Shailza Singh *Editor*

Synthetic Biology

Omics Tools and Their Applications

Springer

Synthetic Biology

Shailza Singh
Editor

Synthetic Biology

Omics Tools and Their Applications

Springer

Editor
Shailza Singh
Department of Pathogenesis and
Cellular Response
National Centre for Cell Science,
Computational and Systems Biology Lab
Pune, Maharashtra, India

ISBN 978-981-10-8692-2 ISBN 978-981-10-8693-9 (eBook)
https://doi.org/10.1007/978-981-10-8693-9

Library of Congress Control Number: 2018957320

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

About the Editor

Shailza Singh is working as Scientist D at National Centre for Cell Science, Pune. She works in the field of Computational and Systems Biology, wherein she is trying to integrate the action of regulatory circuits, cross-talk between pathways and the non-linear kinetics of biochemical processes through mathematical models. The current thrust in her laboratory is to explore the possibility of network-based drug design and how rationalized therapies may benefit from Systems Biology. She has published a book with Springer, titled *Systems Biology Application in Synthetic Biology*. She is the recipient of Rapid Grant for Young Investigators (RGYI) (DBT), DST-Young Scientist and Indian National Science Academy (INSA) (Bilateral Exchange Programme). She is the reviewer of various international and national grants funded from government organizations.

Chapter 1
Integrated Systems and Chemical Biology Approach for Targeted Therapies

Ritika Kabra, Bhavnita Soni, Anurag Kumar, Nutan Chauhan, Prajakta Ingale, and Shailza Singh

Abstract "Omics" tools have revolutionized the systems and synthetic biology concept in past years. Mathematical, computational, and statistical tools have played an enormous role in dealing with the inherent complexity of the biological system whether be metabolic, signaling, or transcriptional network. In this chapter, methods have been discussed for elucidating the mathematical networks underlying different diseases and thereby identifying the key components to predict the response.

Keywords Metabolic network · Signaling network · Systems biology · Omics

1.1 Introduction

Cellular functions are governed by a large complex network of combinatorial interactions. The interacting components being genes, proteins, and metabolites collectively regulate the functioning of various biological systems. Scientists have long focused on the reductionist approach which has successfully identified many important components and their role within various biological systems. Systems biology or integrative biology, on the other hand, uses a holistic approach where the system is visualized as a whole rather than focusing on any specific component. Therefore, systems biology is an interdisciplinary research area which relies on the knowledge gained from biology, mathematics, and computational disciplines (Kitano 2002). The importance of systems biology is now increasing post the Human Genome Project and with the advent of high-throughput technologies. With large-scale datasets generated, the focus is now on integrating the information present such that meaningful interpretations about various systems can be made. The central task of systems biology is therefore to gather information and to integrate and analyze datasets in the form of biological networks. Systems approach to a medicine

R. Kabra · B. Soni · A. Kumar · N. Chauhan · P. Ingale · S. Singh (✉)
National Centre for Cell Science, NCCS Complex, Ganeshkhind, SP Pune University Campus, Pune, India
e-mail: singhs@nccs.res.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_1

strives to unveil the pathogenic mechanisms of diseases and identify biomarkers which would lead to target-based drug designing. With the emergence of large-scale datasets (genome, proteome, transcriptome, metabolome, and personalized datasets), it is imperative that every individual will be surrounded by billions of data points. The convergence of systems biology, digital revolution, and consumer-driven healthcare has made it easier to analyze large-scale datasets thereby transforming medicine to P4 (predictive, preventive, personalized, and participatory) medicine (Hood 2013).

Biological systems are diverse in nature and are broadly classified based on their components, i.e., gene regulatory networks, protein-protein interaction networks, metabolic networks, etc. Biological networks capture, transmit, modulate, and relay the information across various nodes. Due to the large number of biological entities present within a system, systems are mostly nonlinear and complex in nature. It is the interplay of the various components within the network that is important for understanding how networks influence the homeostasis of an organism. Any system is better understood by developing a collective understanding of the system properties, and this can be done by systematically perturbing each and every component and their regulatory reactions. Systems are represented as mathematical models. Models represent a concise summary of our current knowledge of respective systems. They are generally assembled by either the top-down or the bottom-up approaches. While top-down modeling starts with genome-wide experimental datasets in order to discover information pertaining to parts and their interactions within the network, bottom-up approach aims to analyze various subsystems and integrate them such that the properties of higher-order systems can be studied. It is then the spatiotemporal dynamics of the system that are essential for understating the healthy and the diseased state of an individual.

Systems biology is an interdisciplinary means of deciphering the dynamic complexities associated within life. It measures the emergent functions of biological processes that arise from the collective interaction of molecular components. The multitude of technological advancements in the area of "omics" era allows a biologist to gain substantial data on molecular components like gene expression, protein-protein interaction, protein-DNA interaction, lipids, metabolites, etc. (Zhang et al. 2010). These large datasets help infer causal relationship network from statistical correlation. All these interactions forming a network can be integrated into a mathematical model, for understanding complex biological behavior at the systems level.

The two prevalent forms of mathematical modeling in systems biology are data-driven modeling and hypothesis-driven modeling. Data-driven modeling depends on empirical high-throughput measurements (Joyce and Palsson 2006) from microarray, next-generation sequencing and protein mass spectrometry under different experimental conditions like diseased state, knockdown or knockout. This helps in generating comprehensive inventory about the molecular components of the system under study and also delineates their interaction possibilities. Computational and statistical tools integrate the data to infer global or local organization of the system under consideration and elaborate the network structure. The resultant description is

highly complex and requires large number of further experimentation and computation to build an accurate and usable quantitative model (Germain et al. 2011; Kholodenko et al. 2012). Data-driven models have mainly contributed to the discovery of diagnostic, prognostic, and therapeutic markers. Examples of data-driven modeling can be found in the area of transplant rejection vs. tolerance (Roedder et al. 2011), vaccine efficacy (Nakaya et al. 2011, 2012), infections (Ramilo et al. 2007; Ura et al. 2009), and autoimmunity (Liu et al. 2006; Galligan et al. 2009; Pascual et al. 2009).

When presumed knowledge about a system is translated into a mathematical form, it is known as hypothesis-based modeling. In this method, any equation that is associated with the hypothesis will lead to a specific prediction that needs to be verifying with empirical data, i.e., data fitting. This helps in quantitating the cellular components in terms of its concentration and kinetic parameters that drive system behavior, enabling prediction-based design of perturbations to drive the system toward a desired behavior, like designing effective therapy modules (Arazi et al. 2013). For practicality in the predictions, complexity of the model is reduced, i.e., the hypothesis-driven models are simplistic in nature. Such models have provided useful insights in studying immunity and infection dynamics (Antia et al. 2005) driving further investigations.

In either approaches of system modeling, the model is depicted within a mathematical framework which deciphers the interaction between cellular components in the form of a network. A network consists of a set of nodes (genes, proteins, and metabolites) and a set of edges (interactions), like a single chemical reaction or higher-level abstractions and like regulatory interaction involving several chemical reactions. Thus cellular networks can be classified, according to the type of entities and interactions involved, as gene regulatory, metabolic, or protein signaling networks. Such complex systems made up of interacting nodes and edges have two essential features, i.e., they are nonlinear interactions and have multiple feedback loops. Feedback can either be negative or positive. Negative feedback loops inhibit the activity of the pathway from which they arise and stabilize the system. Whereas, positive feedback loops enhance the activity of the pathway; however when constrained by saturation effects, they serve as a mechanism to lock a system in a dynamic state for a long time conferring memory of past conditions (Ingalls 2013).

Once the system has been conceived in a mathematical framework with all the necessary nodes and edges, the dynamism of the system is further explored with simulations governed by the mathematical laws embedded within them. The interactions are written as chemical reactions and physical laws such as mass action, Michaelis-Menten kinetics, Hill-Hinze kinetics, etc. are assigned to them. These set of chemical reactions are then simulated and solved in the form of ordinary differential equations (ODEs) either deterministically or stochastically. The advantage of the ODE model is that they capture intracellular concentrations and short time-scale dynamics, which are critical components of signal transduction. The resultant time series or Boolean data is then put to statistical measures to infer the system dynamism under different perturbation conditions. Once the perturbations introduced gives desired system behavior, the perturbation of choice can be introduced

into the system via a synthetic circuit or device for rewiring purposes (Ingalls 2013; Drubin et al. 2007). These in silico designs should be capable of mimicking the in vitro experiments.

1.1.1 Mathematical Models in Context of Metabolic Network

Systems are described using the mathematical formulations and analyzed in order to generate an experimentally testable hypothesis. Mathematical models consist of a set of equations that describe how the system evolves over a period of time, i.e., the system's dynamic behavior. The resulting model is thus mechanistic in nature where any modifications to the model are assumed to mimic modifications in real system. Mathematical models can be either deterministic or stochastic in nature. The behavior of the deterministic model is exactly reproducible; stochastic models on the other hand exhibit randomness in their behavior. Behavior of the model is generally assessed by simulations, which provide value guide to formulate a testable hypothesis. The end goal of most modeling studies is therefore to generate a comprehensive model which is an iterative process where models are subjected to refinement based on available experimental information. The principle components of the model are the molecular species, where the abundance of each species is assigned as a state variable and a collection of all state variables represents the state of a system. The time course for the collection of the state variables defines the trajectory of the system. In addition to the states, models are also defined by parameters which define the association, dissociation, expression, and degradation of various molecular species within the system. The rate of the reaction is therefore dependent on the abundance of molecular species. Overtime, the concentration of the species changes which also changes the rate of a reaction.

1.1.2 Model Building

The underlying complexity of the system is resolved by solving the model in question where the time-course behavior of the system is studied by simulating the model. Models in the long run arrive at the steady state which is the most persisting state within the system. The behavior of the system is then defined by its trajectory right from the initial state to the steady state. To solve a system and to predict its behavior, the thermodynamics of every chemical reaction in the system needs to be understood and a detailed mathematical description needs to be formed.

1.1.2.1 Kinetic Laws

In kinetic modelling, every chemical reaction is defined with kinetic laws and by their corresponding parameters. The kinetic laws commonly used are:

(a) *Generalized mass action kinetics*: Used for nonenzyme catalyzed reactions, the law of mass action states that the reaction rate is proportional to the product of concentrations of the reactants. The exponent to which each reactant appears in the rate law is then referred to as the kinetic order of the reactant in the reaction. Also the probability of a reaction occurring is proportional to the probability of the reactant colliding with each other. One such example of simple reaction is

$$\mathbf{A}+\mathbf{B}\rightarrow\mathbf{X}$$

Then the rate of the reaction is k_1 [A][B]. The downside of using law of mass action is that most enzymatic reactions are not single step and enzyme saturation is not taken into account.

(b) *Michaelis-Menten kinetics*: Generally used for simple-order enzyme-catalyzed reactions. One of the disadvantages of using law of mass action is that these laws fail to describe the enzyme saturation and such reactions can be modeled by Michaelis-Menten rate laws. The assumptions made are that the reaction proceeds rapidly and the product doesn't bind to the free enzyme. One such simple example is

$\mathbf{A + E} \rightleftharpoons \mathbf{EP} \longrightarrow \mathbf{P + E}$ if k_1 is reaction rate of reaction 1, k_{-1} is the reaction rate of reversible reaction 1, and k_2 is the reaction rate of the reaction 2.

The differential equations applied to the model are

$$\frac{d}{\mathrm{d}t}\, a(t) = -k_1 a\,(t)e(t) + -k_{-1}\mathrm{ep}$$

$$\frac{d}{\mathrm{d}t}\, e(t) = k_{-1}\mathrm{ep}\,(t) - k_1 a(t)e(t) + k_2\mathrm{ep}(t)$$

$$\frac{d}{\mathrm{d}t}\, \mathrm{ep}(t) = -k_{-1}\mathrm{ep}\,(t) + k_1 a(t)e(t) - k_2\mathrm{ep}(t)$$

$$\frac{d}{\mathrm{d}t}\, p(t) = k_2\mathrm{ep}\,(t)$$

The K_m of the reaction is then $\frac{k_{-1}+k_2}{k1}$ and the rate law, i.e., $\frac{d}{\mathrm{d}t}\, p(t) = \frac{V\mathrm{maxa}}{K\mathrm{m}+a}$

(c) *Convenience kinetics*: It is a generalized form of reversible Michaelis-Menten rate law. Unlike Michaelis-Menten equation, convenience kinetics is based on the thermodynamically independent parameters and is good for reactions involving enzyme modifiers (activation, saturation, and inhibition). The assumptions made are that substrate binding is in an arbitrary order and rapid equilibrium assumption is made. Generally used for high-order enzymatic reactions,

convenience kinetics has a drawback that the reactions are thermodynamically constrained. For metabolic systems, the convenience kinetics is generally a good choice if the detailed enzymatic mechanism is unknown (Liebermeister and Klipp 2006).

(d) *Hill-Hinze equation*: Generally used for the transcription reactions. Binding of the transcription factors (activators or repressors) to the regulatory region of the genes is described by the Hill-Hinze kinetics, accounting for the cooperative binding of transcription factors. For an activator X, the Hill equation is $f(X) = \frac{\beta \max . X^n}{K^n + X^n}$ where βmax is the maximal production rate of the promoter-transcription factor complex, n represents the Hill coefficient, and K is the activation coefficient. Likewise the Hill function for a repressor is represented by $f(X) = \frac{\beta \max}{K^n + \left(\frac{X}{K}\right)^n}$ where X represents the repressor.

1.1.2.2 Kinetic Parameters

Reactions are also defined by their kinetic parameters. The kinetic parameters pertaining to the model are to be obtained from experimental observations such as time-course data. Kinetic parameters generally refer to not only kinetic constants such as K_m, K_{cat}, and K_d but also to the input-output relations of various components in the system and the initial concentrations of these components (Neves 2011).

1.1.3 Model Analysis

Mathematical models are generally analyzed based on the type of the model, datasets included in the model, level of detail information included in the model, and feasibility. Modeling a system in silico can be used to summarize knowledge about the system, fill in the missing links, and predict the behavior of the system before actual experiments are designed. In this context, modelling techniques are classified into (a) interaction-based modeling, e.g., graph-based representations of biological networks, (b) constraint-based modeling, and (c) mechanistic modeling. While the mechanistic modeling techniques require information about the details of the kinetic parameters, constraint-based methods don't. Models describe processes at variable states: (a) models at steady state where no temporal variations are seen and (b) dynamic models which have spatiotemporal variations. The dynamic behavior of the system is easily grasped with the help of computer simulations. The most commonly used constraint-based modeling technique is flux balance analysis (FBA). FBA is a constraint-based modeling technique. Unlike mechanistic simulations that depend on accurate kinetic data, FBA is based on the principle of conservation of mass. The technique utilizes a stoichiometric matrix, and based on the specific objective function, optimal reaction flux distribution is obtained. Constraints that the model is subjected to are physiochemical, spatial, or topological

constraints. FBA uses linear optimization to determine the steady-state reaction flux distribution in a metabolic network and aims at maximizing the objective function. FBA is generally used to study the effect of perturbations such as gene deletions and in inhibition studies where in silico predictions can be made about the effect of the perturbation on the proposed objective. The steps involved in carrying out the FBA analysis are (a) defining the system, (b) obtaining the reaction stoichiometry, (c) defining the objective function and constraints the model is subjected to, and (d) model optimization. The dynamic mass balance of the system under study is then defined in the form of a stoichiometric matrix (*Smxn*), with the flux rates being *Vnx*1 to the time derivatives of metabolic concentrations *Xmx*1 as $\frac{dx}{dt} = Sv$ where $v = [v1, v2, v3. \ldots vn1. b1\ b2. \ldots bn]$ *T*, *vi* is internal fluxes, *bi* is the exchange of fluxes, and n is the number of internal metabolites. At the steady state, $\frac{dx}{dt} = 0$ (Kauffman et al. 2003; Orth et al. 2010). The advantages of using FBA are that no kinetic information is required and only information pertaining to the genes and enzymes involved, their transport, and compartmentalization is required. There have been several modifications to the FBA such as the regulatory FBA (rFBA) which has been extremely useful in the gene deletion studies and in experiments with extreme conditions and perturbations (Covert et al. 2001). In addition to FBA, several other constraint-based methods which are used for analysis include MOMA (minimization of metabolic adjustments) (Segre et al. 2002; Shlomi et al. 2005), elementary flux mode analysis, and extreme pathways (Papin et al. 2003).

Constraint-based modeling doesn't require kinetic parameters. However with the advancement in the field of molecular biology and high-throughput technologies, it has now been easy to decipher kinetic parameters associated with the model. Numerical simulation is a process of predicting the behavior of the model based on kinetic parameters and then studying the spatiotemporal behavior of the system. Numerical simulation can be either deterministic, stochastic, or hybrid in nature. In deterministic simulation, chemical reactions within the model are defined in the form of differential equations that are based on law of mass action. The law of mass action states that the rate of the reaction is directly proportional to the concentration of the reactants involved in the particular reaction. Deterministic model assumes that output is certain if a particular input is fixed, thus the behavior of a deterministic model is fixed and predictable. Generally used for population sizes are large, deterministic simulations that are less computationally intensive and are good for fast chemical reactions. Deterministic simulations however are a misfit for reactions with slow rate and with low number of molecules and simulations in which stochastic variations are large. Insights into the behavior of deterministic models can be gained by solving and computing the equilibrium points and verifying the stability by observing the limit cycles and the robustness of the model. Stochastic simulation on the other hand is used to study the random fluctuations of the different molecules within the model. As opposed to the deterministic simulation, the output obtained is uncertain which leads to randomness. Based on the concept of probability, behavior of the model is unpredictable. The advantage of stochastic models is that it accounts for the minor fluctuations of the parameters; however stochastic simulation is good

for models with low number of species and is more computationally intensive. There are several algorithms to carry out stochastic simulation like (a) Monte Carlo method, (b) Gillespie method, (c) tau-leap method, and (d) higher-order tau-leap method. Monte Carlo method uses two random numbers one depicting the time of the event and the other random number indicates the next reaction possible. While the time of the event is decided based on the concentration of the reactants, the reaction rate decides the next reaction within the system. The pattern of the data is then correlated with time. Unlike the deterministic simulation, the number of molecules is taken into account for the generation of random numbers rather than the concentration of the species involved in the reaction. The Gillespie method is a higher-order Monte Carlo method where the reaction rates and the probability of transition of reactions are defined in the form of a chemical master equation (CME). The CME thus governs the time-course evolution of the system starting from one state, jumping to other states during the random walk procedure until the final state is achieved (Gillespie 2001). With the original Gillespie method being slower, the tau-leap method was developed which is approximate but much faster. The method involves setting a time step within which all the reaction rates are calculated, and based on the poison random numbers, the states are updated. The reaction rates are less frequently updated than the original Gillespie method and thus the tau-leap method is much faster (Cao et al. 2006). Multiscale modeling uses the hybrid method of model analysis which uses both the stochastic and the deterministic method of simulation. First proposed by Haseltine and Rawlings, in the hybrid method, the deterministic simulation is performed for fast reactions, while the slow reactions are simulated by the stochastic simulation (Haseltine and Rawlings 2005).

1.1.4 Importance of Systems Biology and Evolutionary Biology in Target Identification

Advances in the field of bioinformatics, genomics, and subsequent emergence of evolutionary genomics in the last decade have led to a more interdisciplinary approach toward drug target discovery. One of the major problems of drug discovery is the quality of the target proteins. It is now possible to better understand a disease by studying the various biological networks associated with the disease. In the diseased state, the envelope of information being transmitted across the network is altered. Disease-specific gene regulatory, metabolic, and protein-protein interaction networks can throw light on the various nodes that are perturbed in the diseased state. Once the target protein is identified, the quality of the target protein can be assessed by understanding the evolutionary patterns associated with the drug targets.

Drug targets are prioritized based on their essentiality, druggability, and uniqueness. While systems biology helps in the identification of essential nodes within the network, evolutionary principles determine the evolutionary rates and patterns of these nodes within the network. Examining the evolutionary rates of various drug

targets is essential as the target proteins with a higher rate of non-synonymous substitutions often have a higher rate of *mutability of* their *amino acids*. Drug interaction with such sites could result in added drug pressure on the pathogen and the emergence of drug resistance. It is therefore imperative to apply all possible in silico strategies to screen and pick the precise drug target for further drug design.

1.1.5 Systems Biology and Synthetic Biology Interface

Research in systems biology has resulted in an offshoot branch known as synthetic biology, which empowers the biologist to design, improve, and rewire biological functions using well-characterized biological parts (promoters, ribosome binding sites, coding sequences for chimeric proteins, RNA switches, etc.) for better system performance (Drubin et al. 2007). Synthetic biology has made progressive advancement with rapid development in molecular biology and bioinformatics techniques.

By applying engineering and computational principles, it has enabled the creation of functional devices, biological networks, and also organisms that do not exist in nature. Complex synthetic gene regulatory circuits like toggle switches (Gardner et al. 2000), oscillators (Elowitz and Leibler 2000; Fung et al. 2005), timers (Ellis et al. 2009), counters (Friedland et al. 2009), clocks (Danino et al. 2010), pattern detectors (Basu et al. 2005), band-pass filters (Tabor et al. 2009), and intercellular communication systems (You et al. 2004; Bulter et al. 2004) have been designed and created in prokaryotes or lower eukaryotes. These early works have opened new avenues for the application of synthetic devices in therapeutics like cancer-targeting therapy (Forbes 2010; Miest and Cattaneo 2014), prevention of cholera infection (Duan and March 2008), and production of precursor of the antimalarial drug artemisinic acid (Ro et al. 2006). Success in engineering bacteria, virus, and yeast for therapeutic purposes has encouraged researchers to engineer mammalian transgene control devices. Some of them have been designed to respond to small molecules (Folcher et al. 2013) and some to respond to light and radio waves (Stanley et al. 2012), and more recently gene homeostatic networks were created to monitor pathological levels of metabolites in the body, such as uric acid, fatty acids, bile acids, and dopamine (Kemmer et al. 2010; Rössger et al. 2013).

Synthetic biology when applied to signaling pathways allows manipulating existing and creating novel pathways. The broad goal of synthetic biology is to engineer novel genetic circuits so as to produce a specific regulatable behavior. Reprogramming the protein-protein interactions and output responses was elegantly demonstrated using the prokaryotic two-component signal transduction systems. In this study the specificity of the osmolarity sensor protein EnvZ from *E. coli* was modified, which normally targets the response regulator protein OmpR. EnvZ was modified by replacing its entire phosphorylation subdomain with either sensor protein RstB from *E. coli* or CC1181 from *Caulobacter crescentus*. The modified EnvZ protein successfully phosphorylated the correct cognate response regulators of either RstB or CC1181 in vitro depending on respective phosphorylation

subdomain. This work was a step forward in engineering of protein-based synthetic pathways (Skerker et al. 2008). The evolutionary conservation of signaling systems provides an attractive platform for developing eukaryotic synthetic signal transduction pathway. Due to the intrinsic dynamism associated with signaling pathways, it is important to specify the exact input-output relationship. Since, in synthetic pathway design, the prime goal is to modify the existing pathways in a cell, it is important that the implementation is reliable and reproducible. Since phosphorylation regulates many cellular processes like gene expression, metabolism, and cellular communication, synthetic signaling can be achieved by targeting the phosphorylation process both in prokaryotes and eukaryotes (Kiel et al. 2010). Like phosphorylation, cell-signaling pathways are influenced by presence of cofactors, multimerization of proteins, enzyme turnover, and the active or passive transport of ions across the membrane, which needs to be considered while designing the synthetic pathway. As in systems biology, in synthetic biology too, synthetic signaling involves design considerations with certain assumptions to be formulated into a mathematical framework (deterministic or stochastic). Estimation and optimization of parameters in the model form a crucial step in obtaining a satisfactory model. Parameter estimation and optimization govern the prediction reliability of the model, while sensitivity and bifurcation analysis direct insights toward an effective design (Zheng and Sriram 2010). A fine-tuned mathematical model should be able to receive unnatural inputs or channel the response via nonnatural route or generate nonnatural outcomes. Last but not the least, orthogonality forms an important component in synthetic signaling circuit designed to help minimize the cross talk with the host (Hansen and Benenson 2016).

Systems and synthetic biology complement each other in dealing with the inherent complexity involved in the biological system and help design synthetic devices for cell-based therapies.

1.2 Mathematical Model in Context of the Signaling Network

1.2.1 What Is Mathematical Modeling?

A mathematical model is a representation of system processes that cannot be observed directly. Biological processes are generally depicted as static drawings, sketches, or three-dimensional (3D) models. Being static in nature, these models do not show time evolution, i.e., dynamism, which is the foremost characteristic of a biological system. To study biological time evolution, mathematical formalism is applied, which has been possible due to enormous data being generated on various high-throughput platforms. Empirical observations are used to describe a biological phenomenon, which are analyzed with mathematical forte for prediction of future behavior that is unmeasured or unseen. These predictions are validated by sets of

experiments which either prove or disapprove them. Disapproval leads to model improvements and data fitting for further empirical data generation. Thus mathematical modeling is an iterative process that helps to predict and validate real-world biological phenomena (Dym 2004).

Any mathematical model building initiative should pose a clear goal that is to be achieved at the end of the modeling process. It should be homogeneous, consistent, conserved, and balanced and must follow the physical laws that define a biological process. Ultimately, the physical laws are used to define the system in the form of ODEs, which when solved or simulated gives the system behavior over time (Materi and Wishart 2007).

Linearity is one of the most important concepts in mathematical modeling. Models of devices or systems are said to be linear when their basic equations—whether algebraic, differential, or integral—are such that the magnitude of their behavior or response produced is directly proportional to the input that drives them. This is important in biology as biological systems are inherently nonlinear, and for making prediction about biology, the approximations should be linearized.

Powerful high-throughput technologies have given an extensive parts list, i.e., proteins, mRNA, DNA, small molecules, etc. of the cell, and have also shed light on a general picture of interactions among them. These make up the metabolic and gene regulatory pathways that play crucial roles in response to external/internal stimuli that ultimately guide the cellular processes. A high degree of cross talk between these pathways has resulted in complexity of the system which makes predicting biological behavior almost impossible. Computer simulations of such physical interactions have proved useful for understanding their topology and dynamics (Materi and Wishart 2007).

A mathematical system can be simulated using either deterministic or stochastic approaches. Deterministic approach does not take into consideration the inherent randomness associated with the system, and therefore the system predictions are consistent, i.e., for an initial set of inputs, the output will be the same. Whereas, stochastic system accounts for the randomness, and therefore for an initial set of inputs, the outputs have to be averaged over using statistical measures. Further with respect to time, these approaches can be of continuous or discrete in nature.

1.2.2 Network Motifs

Most of the complexities that is observed in biological system arise from the complex regulatory circuits that comprise of small set of recurring regulation or wiring patterns, called network motifs (Milo et al. 2002; Shen-Orr et al. 2002). They were first defined in *E. coli*; the same motifs have since been found in other bacteria (Eichenberger et al. 2004; Kalir et al. 2005), yeast (Milo et al. 2002; Lee et al. 2002), plants, and animals (Odom et al. 2004; Boyer et al. 2005; Saddic et al. 2006; Iranfar et al. 2006). The two main types of autoregulatory motifs are the autoregulatory negative and positive feedback loops. Positive feedback loops are often found in

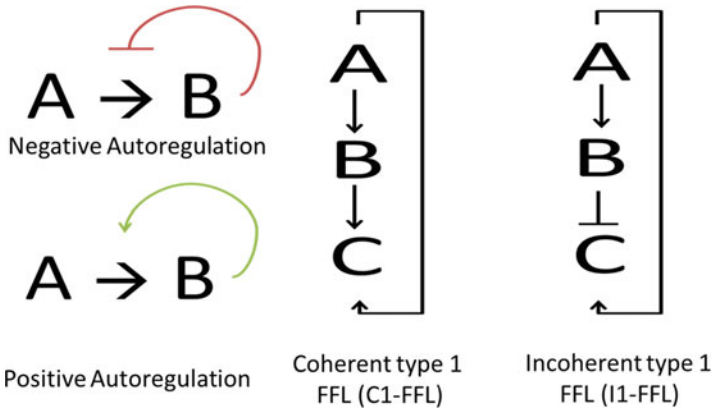

Fig. 1.1 Common network motifs loops in a network

developmental transcription networks in which two transcription factors (TF) regulate each other in a positive manner or a single TF positively regulates itself. In either case, a transient signal can cause the loop to lock irreversibly into a steady-state equilibrium state, providing memory for an input signal. The response times are slowed, and cell-to-cell variation is usually enhanced and shows cooperativity. In negative feedback loop, two repressors repress each other or a single repressor represses itself. It speeds up the response time of gene circuits and reduces cell-to-cell variation in protein levels due to fluctuations in production rate. Other network motifs are feed-forward loops (FFL) which can be coherent or incoherent. FFL appears in hundreds of gene systems from prokaryotes to eukaryotes. In FFL, the regulator X regulates Y and Z, and Z is also regulated by Y. These regulatory interactions can either activation or repression, and therefore FFL can have eight possible structural types. The two most common FFLs are the coherent type 1 FFL (C1-FFL) and the incoherent type 1 FFL (I1-FFL) (Shoval and Alon 2010) (Fig. 1.1).

These feedback loops have a profound effect on the dynamics and outcome of biological systems. Therefore it is important to identify and define these network motifs in the mathematical model, which can give deeper insight into the understanding of the system.

1.2.3 Reconstruction of the Signaling Network

Signaling network reconstruction integrates data from different public resources that describes the biochemical conversions in the form of elementary chemical equations. These chemical transformations are then embedded within a mathematical formulation for quantitative modeling. Doing so helps one to decipher dynamic interactions between the signaling proteins and explain the biological interactions at higher

complexity level. Quantitative modeling of these interactions plays an important role in understanding fundamental intra- and intercellular processes (Papin et al. 2005).

As in any mathematical modeling, the details to be included in the model are only those that could help us understand the dynamics associated with protein concentration changes. In the model the processes of phosphorylation/dephosphorylation and ubiquitination/deubiquitination are embedded within the mathematical formulation and the details are omitted. It is also important to consider the number of reactions to be included in the network and the level of details that is to be accounted for. The reconstruction is done in a way that highlights the importance of highly connected node in the system and a linear pathway depicting the signaling input and output (Papin et al. 2005).

The foremost step in reconstruction of signaling network is the delineation of all the known individual biochemical interaction that comprises the network under consideration. The interactome of the signaling cascade is assembled by extensive literature survey and from signaling databases like the KEGG, INOH Pathway Database (release 4.0), Pathway Interaction Database (PID), and Science Signaling Database.

The interactions between the signaling proteins are entered as elemental chemical reactions along with the kinetics laws that govern them in MATLAB's SimBiology toolbox (The MathWorks, Inc.) which uses the Systems Biology Markup Language (SBML) machine language to represent these interactions and the associated kinetic laws. The kinetic rate laws (mass action for association and dissociation reactions, Henri-Michaelis-Menten for phosphorylation/dephosphorylation/ubiquitilination/deubiquitilination, Hill equation for gene expression) and initial concentrations associated with the reactions are also defined. The reconstructed signaling network model is a hypothesis-driven model; the parameters and variables assigned to the reactions lie well within the biological ranges that have been extensively recorded in literature. Decomposition approach is considered for the reconstruction, i.e., a large network is broken into smaller components. Doing so reduces the problem of parameter estimation associated with each reaction in the network. This approach also significantly improves the computational efficiency by reducing the search space dimensionality (Koh et al. 2006). Reconstructed signaling network is modeled as two modules: one without a feedback loop and the second one with a feedback loop.

The reconstructed network is numerically simulated using ODE15s solver (SimBiology toolbox) which generates the first-order nonlinear ODEs for each node, thus defining the mathematical structure of the model. ODE helps in determining time-dependent changes, i.e., the time series data of the concentrations of the signaling proteins and protein complexes and thus the dynamics associated with it. The model is exported to Copasi (4.8.35) as a SBML file to generate the time series data. The time series data gives us the changes of each component at every time point of the total simulation time.

1.2.4 Parameter Estimation

The equations defining the chemical transformation in the signaling cascade use variables as concentration of species, i.e., signaling proteins involved in the reaction. Also these equations are dependent on the parameters like production and degradation rate. The dynamic behavior of a system is highly dependent on some of the values of the parameters and therefore an accurate and reliable quantification of the parameters is essential. This helps in strengthening the predictability of the reconstructed model.

Often these parameters are unknown, difficult (due to nonlinearity), and expensive to measure. In such situation parameter estimation is done, i.e., the unknown parameters are determined indirectly using computational biology tools (Ma'ayan 2011; Rangamani and Iyengar 2008). Since the reconstructed signaling network is a hypothesis-driven model, no quantitative or repeatedly measured data was needed. The parameters for reconstructed signaling network is manually trained within a range of parameter space limited to biological response, which is further fine-tuned till the model is simulated to obtain a desired dynamic behavior. This is an iterative process and is done till a reproducible graph depicting the biochemical behavior is obtained (Fig. 1.2).

1.2.5 Network Matrices

It includes information about the general and specific properties of nodes (proteins), edges (interactions), and modules within the network using graph theory. It also gives information about the network architecture and helps identify differences between networks. Cytoscape is used to visualize the network topology in the form of a directed acyclic graph (DGA). Properties of nodes and edges in the network allow us to decipher important functions of the proteins and their interaction in the network. The most common network matrices that are used are depicted in (Ma'ayan 2011) Table 1.1.

The shortest path length was computed using adjacency matrix (Fig. 1.3a, b, (Pavlopoulos et al. 2011).

Computational modeling of signaling pathways and regulating them through key players help understand the molecular, genetic, and proteomic function of key protein components in relation to disease. Moreover, the computational tool offers an advantage of analyzing the signaling cascade in terms of its stability and robust performance.

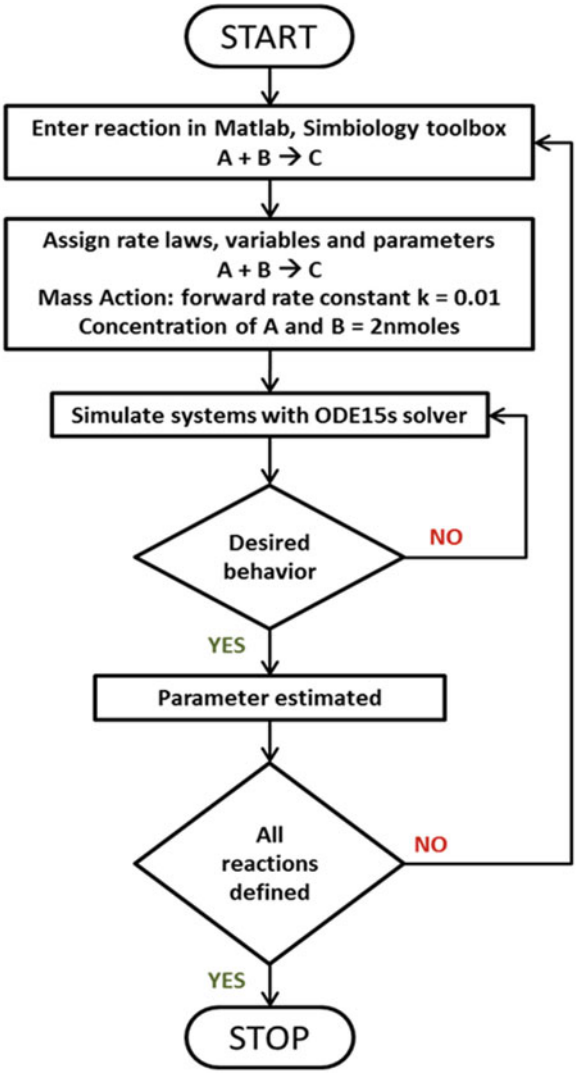

Fig. 1.2 Flowchart for network reconstruction and parameter estimation in MATLAB, SimBiology toolbox

Table 1.1 Network matrices and their description

Network matrices	Description	Mathematical representation
Density	It shows how densely the network is populated with edges. The density is a value between 0 (no edges and isolated nodes) and 1 (a clique). Self-loops and duplicated edges are ignored for calculating density	$D = \frac{E}{N(N-1)}$
Average degree	Number of average edges connected to a node	$k = \frac{2E}{N}$
1. In-degree	(a) In coming edges to a node	
2. Out-degree	(b) Outgoing edges from a node	
Average path length	Average path length is the number of shortest steps taken to move from one component in the network to the other. It is calculated by finding the summation of shortest path between all pairs of nodes divided the total number of pairs. This shows us, on average, the number of steps it takes to get from one member of the network to another	$a = \sum_{s,t,V} \frac{d(s,t)}{N(N-1)}$
		Where $d(s,t)$ is the shortest path between node s and t
Diameter	It is the shortest distance between the two most distant nodes in the network	$d = \max_{v\varepsilon V} e(v)$
Clustering coefficient	It is the degree to which nodes in a network tend to cluster together	$C = \frac{2eN}{kN(kN-1)}$
	Network clustering coefficient is the average clustering coefficient of all the nodes in the network	Where kN is the number of neighbors of N and eN is the number of connected pairs between all neighbors of N
Betweenness centrality	It quantifies the number of times a node acts as a bridge along the shortest path between two other nodes	$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$
		Where σ_{st} is the total number of shortest paths from node s to node t and σ_{st} (v), (v) is the number of those paths that pass through v
Closeness centrality	It is the average length of the shortest path between the node and all other nodes in the graph	$C(x) = \frac{1}{\sum_y d(y,x)}$
		Where $d(y, x)$ is the distance between vertices x and y
Edge betweenness	It is defined as the absolute number of the shortest paths that go through an edge in a network	–

E number of edges, N number of nodes

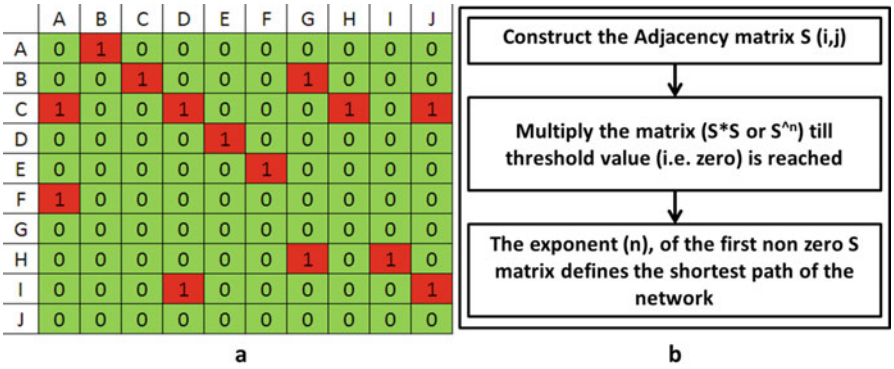

Fig. 1.3 (**a**) Non-weighted adjacency matrix defining the interaction in red (1) and no interaction in green (0) between the components A to J comprising a network. (**b**) Flowchart to calculate the shortest path using the adjacency matrix

References

Antia R, Ganusov VV, Ahmed R (2005) The role of models in understanding CD8+ T-cell memory. Nat Rev Immunol 5:101–111

Arazi A, Pendergraft WF, Ribeiro RM, Perelson AS, Hacohen N (2013) Human systems immunology: hypothesis-based modeling and unbiased data-driven approaches. Semin Immunol 25:193–200

Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R (2005) A synthetic multicellular system for programmed pattern formation. Nature 434:1130–1134

Boyer LA et al (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122:947–956

Bulter T et al (2004) Design of artificial cell – cell communication using gene and metabolic networks. Proc Natl Acad Sci U S A 101:2299–2304

Cao Y, Gillespie DT, Petzold LR (2006) Efficient step size selection for the tau-leaping simulation method. J Chem Phys 124(4):044109

Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. J Theor Biol 213(1):73–88

Danino T, Mondragón-Palomino O, Tsimring L, Hasty J (2010) A synchronized quorum of genetic clocks. Nature 463:326–330

Drubin DA, Way JC, Silver PA (2007) Designing biological systems. Genes Dev 21:42–254

Duan F, March JC (2008) Interrupting Vibrio cholerae infection of human epithelial cells with engineered commensal bacterial signaling. Biotechnol Bioeng 101:128–134

Dym C (2004) Principles of mathematical modeling. Academic, New York

Eichenberger P et al (2004) The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. PLoS Biol 2:e328

Ellis T, Wang X, Collins JJ (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat Biotechnol 27:465–471

Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–338

Folcher M, Xie M, Spinnler A, Fussenegger M (2013) Synthetic mammalian trigger controlled bipartite transcription factors. Nucleic Acids Res 41:e134–e134

Forbes NS (2010) Engineering the perfect (bacterial) cancer therapy. Nat Rev Cancer 10:785–794

Friedland AE et al (2009) Synthetic gene networks that count. Science (80) 324:1199–1202

Fung E et al (2005) A synthetic gene – metabolic oscillator. Nature 435:118–122
Galligan CL et al (2009) Multiparameter phospho-flow analysis of lymphocytes in early rheumatoid arthritis: implications for diagnosis and monitoring drug therapy. PLoS One 4:e6703
Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature 403:339–342
Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser IDC (2011) Systems biology in immunology: a computational modeling perspective. Annu Rev Immunol 29:527–585
Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys 115(4):1716–1733
Hansen J, Benenson Y (2016) Synthetic biology of cell signaling. Nat Comput 15:5–13
Haseltine EL, Rawlings JB (2005) On the origins of approximations for stochastic chemical kinetics. J Chem Phys 123(16):164115
Hood L (2013) Systems biology and p4 medicine: past, present, and future. Rambam Maimonides Med J 4(2):e0012
Ingalls B (2013) Mathematical modelling in systems biology: an introduction. The MIT Press, Cambridge, MA, Internet.[cited p. 117]
Iranfar N, Fuller D, Loomis WF (2006) Transcriptional regulation of post-aggregation genes in Dictyostelium by a feed-forward loop involving GBF and LagC. Dev Biol 290:460–469
Joyce AR, Palsson BØ (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7:198–210
Kalir S, Mangan S, Alon U (2005) A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*. Mol Syst Biol 1:2005
Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14(5):491–496
Kemmer C et al (2010) Self-sufficient control of urate homeostasis in mice by a synthetic circuit. Nat Biotechnol 28:355–360
Kholodenko B, Yaffe MB, Kolch W (2012) Computational approaches for analysing information flow in biological networks. Sci Signal 5:re1 2002961
Kiel C, Yus E, Serrano L (2010) Engineering signal transduction pathways. Cell 140:33–47
Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664
Koh G, Teong HFC, Clément M-V, Hsu D, Thiagarajan PS (2006) A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk. Bioinformatics 22:e271–e280
Lee TI et al (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science (80-) 298:799–804
Liebermeister W, Klipp E (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. Theor Biol Med Model 3(1):1
Liu Z, Maas K, Aune TM (2006) Identification of gene expression signatures in autoimmune disease without the influence of familial resemblance. Hum Mol Genet 15:501–509
Ma'ayan A (2011) Introduction to network analysis in systems biology. Sci Signal 4:tr5
Materi W, Wishart DS (2007) Computational systems biology in drug discovery and development: methods and applications. Drug Discov Today 12:295–303
Miest TS, Cattaneo R (2014) New viruses for cancer therapy: meeting clinical needs. Nat Rev Microbiol 12:23–34
Milo R et al (2002) Network motifs: simple building blocks of complex networks. Science (80-) 298:824–827
Nakaya HI et al (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12:786–795
Nakaya HI, Li S, Pulendran B (2012) Systems vaccinology: learning to compute the behavior of vaccine induced immunity. Wiley Interdiscip Rev Syst Biol Med 4:193–205
Neves SR (2011) Obtaining and estimating kinetic parameters from the literature. Sci Signal 4(191): tr8

Odom DT et al (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science (80-) 303:1378–1381
Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28(3):245–248
Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO (2003) Metabolic pathways in the post-genome era. Trends Biochem Sci 28(5):250–258
Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6:99
Pascual V, Chaussabel D, Banchereau J (2009) A genomic approach to human autoimmune diseases. Annu Rev Immunol 28:535–571
Pavlopoulos GA et al (2011) Using graph theory to analyze biological networks. BioData Min 4:10
Ramilo O et al (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 109:2066–2077
Rangamani P, Iyengar R (2008) Modelling cellular signalling systems. Essays Biochem 45:83–94
Ro D-K et al (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440:940–943
Roedder S, Vitalone M, Khatri P, Sarwal MM (2011) Biomarkers in solid organ transplantation: establishing personalized transplantation medicine. Genome Med 3:37
Rössger K, Charpin-El-Hamri G, Fussenegger M (2013) A closed-loop synthetic gene circuit for the treatment of diet-induced obesity in mice. Nat Commun 4:2825
Saddic LA et al (2006) The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER. Development 133:1673–1682
Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci 99(23):15112–15117
Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64–68
Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc Natl Acad Sci U S A 102(21):7695–7700
Shoval O, Alon U (2010) Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot
Skerker JM et al (2008) Rewiring the specificity of two-component signal transduction systems. Cell 133:1043–1054
Stanley SA et al (2012) Radio-wave heating of iron oxide nanoparticles can regulate plasma glucose in mice. Science (80-) 336:604–608
Tabor JJ et al (2009) A synthetic genetic edge detection program. Cell 137:1272–1281
Ura S et al (2009) Differential microRNA expression between hepatitis B and hepatitis C leading disease progression to hepatocellular carcinoma. Hepatology 49:1098–1112
You L, Cox RS, Weiss R, Arnold FH (2004) Programmed population control by cell- -cell communication and regulated killing. Nature 428:868–871
Zhang K, Beverley SM (2010) Phospholipid and sphingolipid metabolism in Leishmania. Mol Biochem Parasitol 170(2):55–64
Zheng Y, Sriram G (2010) Mathematical modeling: bridging the gap between concept and realization in synthetic biology. Biomed Res Int 2010.

Chapter 2
Application of Bioengineering in Revamping Human Health

Shreya Ghosh, M. S. Kumar, Bhubaneswari Bal, and A. P. Das

Abstract Bioengineering helps in improving the quality of human health and prolongs the lives of the patients. Recent research in this field has witnessed several advances such as pacemakers, artificial organs, and novel technologies such as X-ray machine, magnetic resonance imaging, computed tomography, pulse oximeter, and ventilators. Recent advances in bionics, an integrated part of this field, comprise of prosthetic implants for hips, knees, joints, cochlear implants, mechanical heart assist pumps, contact lenses, breast implants heart valves, etc. Several novel bionic advances such as bionic lung, bioartificial heart, and artificial liver are undergoing clinical trials which are going to revolutionize the biomedical sector in the near future. However, the discoveries in the field of bioengineering rely on the basic research in the fields of physical sciences, biological sciences, and medicine. It may be considered as the spotlight emerged from the convergence of all the above fields to sustain human health. Bioengineering-based biosensors and nanosensors are acquiring higher attention in the aspect of diagnosing infections. They offer higher sensitivity and accuracy being cost-effective and rapid. Various types of biosensors such as electrochemical, magnetoelastic, and cantilever-based and aptamer-based sensors have been successfully employed for diagnosis and treatment. Many nanoparticles were proven to be antibacterial and target specific. Hence, basic research in medicine, chemistry, nanoscience, biology, material science, and biophysics may result in the development of an outstanding technology that would eventually reach the clinic and serve the mankind. An output in the field of bioengineering always inspires further novel research.

Keywords Bioengineering · Nanosensors · Healthcare · Imaging · Implants

S. Ghosh · M. S. Kumar
Bioengineering Laboratory, Centre for Biotechnology, Siksha O Anusandhan (Deemed to Be University), Bhubaneswar, India

B. Bal · A. P. Das (✉)
Department of Chemical and Polymer Engineering, Tripura University (A Central University), Suryamaninagar, Tripura, India
e-mail: alokprasaddas@tripurauniv.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_2

2.1 Introduction

Bioengineering is advancement in technology that relates engineering principles to natural systems and medical expertise. Bioengineering research includes engineered microorganisms to produce food, fertilizers, medicines, chemical, novel diagnostics technology, infection analytic kits, and biomaterial organs (Kumar and Das 2017; Das et al. 2011, 2016; Bal et al. 2017). Biomaterials from health viewpoint are mentioned as "materials with novel characters that build them suitable to come in immediate contact with human cell with no undesirable reactions." Biomaterials are applicable to physically replace any injury or destruction through any pathological processes by stiff or flexible tissue. Addition biomaterials are applicable in various healthcare sectors like disposable health devices, analytical kits, therapeutics, etc. (Das et al. 2014; Bal et al. 2016). Recently biomaterials play a crucial position in the field of biomaterial engineering to advance the medical care of humanity. Bioengineering covers two intimately connected field of attention, i.e., (1) it focuses on the principle of engineering technology to understand the function of living organisms and (2) it applies knowledge to develop novel diagnostic kits and novel biomedical applications. Universally bioengineering primarily focuses on the application of biomaterials to improvise the human health services. With the advancement of bioengineering techniques, there has been a development of human healthcare, i.e., in diagnostic and medical devices. Bioengineered-based biosensors are analytical devices that make use of biomolecules to recognize target molecule in the sample through electrical indication and spectrometric or luminous indicators. In spite of the vast arena of biosensing application, biosensors are broadly separated in two groups: (i) rapid costly devices and (ii) low-priced, easy to work out portable kits (WHO 2012; Das et al. 2015).

Detection of pathogenic strains is essential for disease control in numerous areas such as pharmaceutical products, food, medical implants, ecological sample, and drinkable water. Casualty rates due to pathogenic contamination remain high and have not changed significantly in the last few years. This represents an intensification of examination amount since the last few years time. Although enormous attempts are under examination for making biosensors, comparatively few microbes, and their toxins, can yet be detected by available sensors. Additional development with these expertises has transformed the human health and made the timely healing feasible. Magnetic resonance imaging (MRI) has considerably reduced the mortality rate, thus improving the life span. Further progresses are projected in near future toward the advances of treatment strategies in healthcare sectors.

2.2 Approaches of Biomedical Engineering

Biomedical engineering (BME) advances have led to the overcoming of limitations of conventional diagnostic methods by the introduction of noninvasive tools for diagnosis and disease classification (Fig. 2.1) (Das et al. 2014). The role of these tools in diagnosis and classification of dengue has been described below.

2.2.1 Ultrasound

The vital phase in dengue is due to the increase in capillary permeability leading to the loss in volume of plasma. The degree of plasma leakage is often detected by bodily inspection techniques. However, numerous investigations have now reported the use of ultrasound for the diagnosis of dengue (Srikiatkhachorn et al. 2007; Venkata et al. 2005; Setiawan et al. 1998; Pelupessy et al. 1989). Srikiatkhachorn et al. (2007) have described the delineation of the site and the instance outflow of plasma in DHF by using ultrasound. The results indicated that plasma leakage coincided with defervescence and the general ultrasonographic indication of plasma

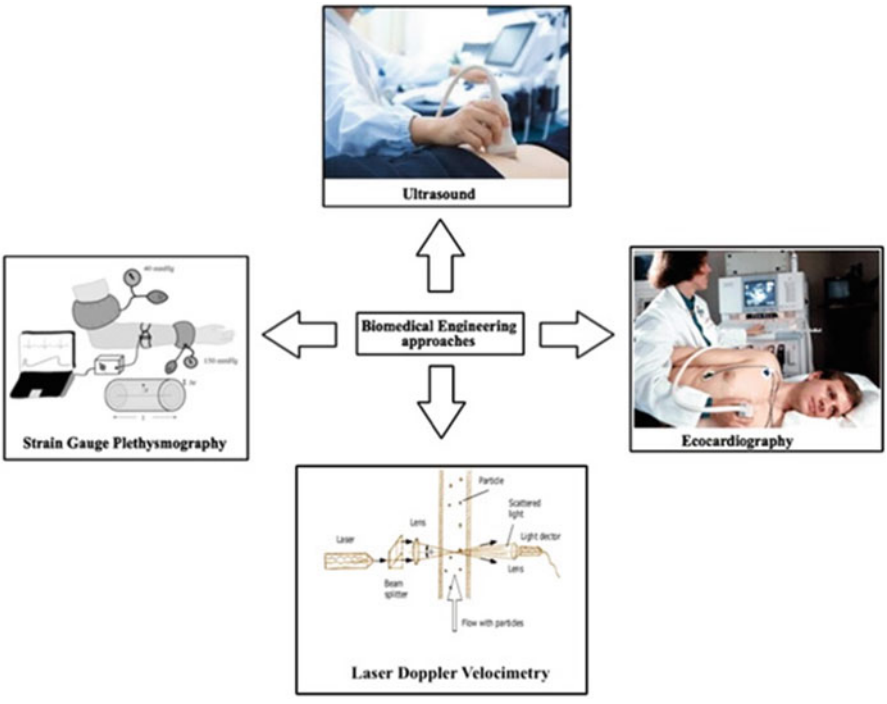

Fig. 2.1 Different approaches of biomedical engineering

leakage was pleural effusion. The study inferred the usefulness of ultrasound imaging for distinguishing plasma leakage in infected patients of dengue. Venkata Sai et al. (2005) also reported an investigation to establish the role of ultrasound in the diagnosis of DHF or DF and also to decide the efficacy of ultrasound into prediction of disease. One hundred and twenty eight suspected dengue patients were studied. Results of 32 patients showed gall bladder wall thickening and pericholecystic fluid. The results of the remaining patients also showed gall bladder wall thickening, ascites, and pleural effusion. This study also demonstrated age role of ultrasound in dengue fever diagnosis. Another study was carried out by Setiawan et al. (1998) that aimed to determine the connection among the clinical proven severity of DHF patients with their sonographic results. The results of this study inferred the importance of ultrasound in early prediction of DHF severity. From the scientific feature, the ultrasound carries an invaluable importance due to its noninvasive, high-quality imaging and comparatively simple execution.

2.2.2 Echocardiography and Electrocardiography (ECG)

Echocardiography and electrocardiography (ECG) have been used in several investigations to study the cardiac functions of dengue infectants (Wali et al. 1998; Yusoff et al. 1993). Pelupessy et al. (1989) used echocardiography to diagnose dengue patients for the presence of pericardial effusion. The investigation revealed that physical examination of DHF patients could not determine any signs of pericardial effusion; however ECG and echocardiogram results clearly demonstrated a low quantity of fluid. This infers that this procedure should be applied to dengue patients that have suffered acute shock. Wali et al. (1998) used echocardiography and ECG for the assessment of the function of the heart s of 17 DHF and DSS infectants. The study concluded that these diagnostic tools were significant in assessing the acute cardiac problem in DHF/DSS. Yusoff et al. (1993) also execute echocardiograms and ECGs on patients for identification of dengue patients. Eighty-seven percent of dengue patients reported abnormal ECGs and/or echocardiograms. The study concluded that ECG and echocardiograms are potent diagnostic tools. They further recommended the early detection of cardiac involvement to identify severe forms of dengue so as to be able to initiate appropriate line of treatment.

2.2.3 Strain Gauge Plethysmography

It has been used for assessing microvascular permeability in patients. Gamble et al. (2000) examined the utilization of age-related variation in microvascular permeability, an indicator of health, and set up the highest price in DSS immature children. Bethell et al. (2001) investigated the underlying pathophysiology of DSS during assessment of the microvascular permeability by means of strain gauge

plethysmography. The study confirmed that the elevated microvascular leakage is found in children.

2.2.4 Laser Doppler Velocimetry

Physiological consideration concerning blood perfusion and movement has been extensively studied by using laser Doppler velocimetry (Kvernebo et al. 1990; Olavi et al. 1991). The changes in microcirculation because of plasma outflow and a boost of microvascular permeability have been evaluated in DHF patients by using this technique (Hassan et al. 2003). Studies indicate the important differentiation among basal laser Doppler flux in usual DHF patients. This outcome hints at the potentiality of this technique to track the microcirculatory changes in DHF patients. However the differentiation of the DHF severity stages has been able to be deduced by this technique.

2.3 Emerging Application of Bioengineering in Healthcare

Bioengineering-inspired materials has been applied in healthcare for an extensive time period. However, noticeable advancement was made and considerable progress has been experimented in medical knowledge and medical equipments since the last 20 years. Reports have publicized that novel modified bioengineered valves with improved performance can be tested for younger patients. A human well-being technology that is modernized by bioengineering conception from the last decade is acting as a pacemaker technology for disease management. Improvements with biosensing machinery currently facilitated the pacemakers to robotically react to different levels of physical activity so as to modify the stimulation pace consequently. Progress in implant medical techniques has grant management to a number of life-menacing sicknesses like neurological disorders. The most significant implication of bioengineering in the area of human health for the last few years is deep brain stimulation for the curing of degenerative illness. Advancement in Healthcare is emerging at a high rate by using the idea of bioengineering and pharmacy. Bioengineering assures a stunning healthcare for numerous patients yearly at minimal medical expenses. Constraint with the conventional methods such as organ transplantation has been taken care by novel bioengineering methods. In recent times bioengineering is developing as possible therapy for craniological diseases. Investigation actions have endeavored to renew the sick heart with therapeutically implantation of artificial tissue. Recently tissue engineering by biomaterials is being applied for infracted myocardium. Advancement in Tissue bioengineering by means of stem cells combined with tissue engineering can be accomplished by utilizing a combination of stem cells with the suitable bioengineered materials for commercial remuneration. Biomaterials for the release of medicines or biomolecules

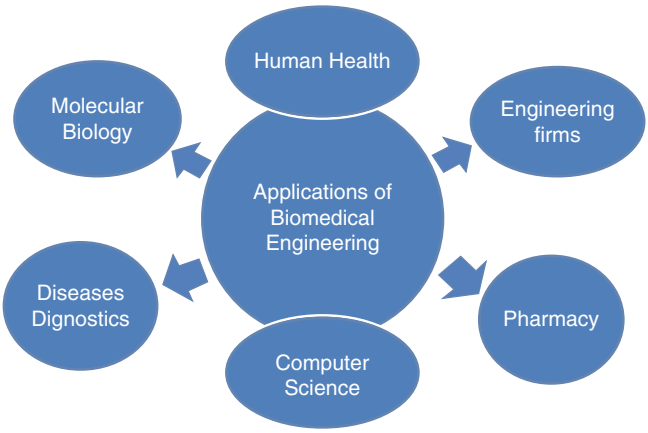

Fig. 2.2 Applications of biomedical engineering

come forward in the 1980s, and this is performing new advances with time. Additional improvements are projected in this area which will get better with commercial and consumer applications (Fig. 2.2).

2.3.1 Role of Gold Nanoparticles

2.3.1.1 Biosensing Applications

Gold nanoparticles (NPs) find their roles in varied biosensing system and bioimaging amplification strategy. Biosensing schemes using gold NPs used for signal amplification have been discussed below.

2.3.1.2 Localized Surface Plasmon Resonance (LSPR) Sensing

Early detection of lethal diseases requires highly reliable and sensitive diagnostic tools. The two key clinical tools for diagnosis are reliant on enzymatic procedures followed by optical diagnostic methods (Jeon et al. 2009). These methods include several procedures and skilled manpower to get accurate results and however depict incomplete multiplexing ability, which is highly essential for rapid and sensitive assays. Therefore, the application of plasmonic nanoparticles in human healthcare can generate a new avenue for the production of cost- effective biosensors having high levels of sensitivity.

2.3.1.3 Propagating Surface Plasmon Resonance (PSPR) Sensing

PSPR is generally energized on constant thin films of metals by the help of prism couplers or grating. The sensitivities of PSPR sensors are extremely high, ranging from pM to fM and they are extremely appropriate for high-throughput assays for studying real-time connections between biomolecules like DNA, antibodies, pathogens, metal ions (i.e., Mg^{2+}, Ca^{2+}), or drugs. There can be a great improvement in the signal intensity of PSPR by incorporating gold NPs, just like the signal amplification approach of LSPR-based biosensing (Zeng et al. 2014).

2.3.1.4 Bioimaging Applications

The optical belongings of gold NPs have made them an attractive substance for bioimaging function. They have tough beam scattering possessions because of LSPR and also pose a robust surface chemistry that makes them suitable for sensitive imaging technologies. Recently developed technologies like existing cell imaging and in vivo imaging procedures are present below.

Live Cell Imaging
Fluorescence-dependent cell imaging is highly popular in life science research. However, the major issue that needs to be addressed now is the partial data and fast photobleaching of fluorophores. The use of plasmon nanoparticles is now treated as an alternative technology for overcoming fluorescence-based techniques and limitations. Dark-field microscope equipped with dark-field condenser generates images due to the white light scattering by plasmonic nanoparticles. Therefore, over a long time span, gold NPs distribution in the cell can be monitored (Wax and Sokolov 2009; Choi and Alivisatos 2010).

In Vivo Imaging
Plasmonic nanoparticles are treated as potent difference factors for imaging purposes like optical imaging, nuclear imaging, and computed tomography (CT). A major problem for obtaining sensitive CT imaging-based diagnoses is the short circulating times (Hainfeld et al. 2006; Lee et al. 2013). To the contrary, extensive circulation times in blood and exact object tissue linked belongings are offered by nano-sized imaging agents. Moreover, greater contrast is offered by gold NPs due to their high atomic number and K-edge value than the conventional iodinated contrast agents. Although gold NPs and iodinated X-ray contrast agents have similar excretion route, the clearance of gold NPs is slower, thereby allowing longer times for imaging. Many investigations have been carried out for determining the surface chemistry and optimal size to obtain competent objective and extensive circulation. Wang et al. reported a study in which they studied in vivo and vitro targeted CT imaging by using gold NPs (Wang et al. 2013). Ahn et al. compared tumor-targeted CT imaging using glycol chitosan-modified AuNPs with NIR fluorescence imaging triggered by a plentiful enzyme in tumor (Sun et al. 2011).

2.3.1.5 Therapeutic Applications

Pharmacodynamics and pharmacokinetics of drug molecules are able to change by using nanocarriers. Moreover, release kinetics can be actively controlled by using gold NPs. The recent technological advances in therapeutic applications of gold NPs have been discussed below.

- *Drug delivery*

Investigations have been carried out on a number of amphiphilic polymers for drug delivery systems due to their safety, stability, and easy ligand-binding chemistry. The inertness and tough surface chemistry of gold NPs have been highly regarded for their use as drug carriers. The optical properties of gold NPs provide the maximum benefit during drug delivery for controlling the drug release kinetics by the application of external light stimuli (Khan et al. 2013). These capabilities of gold NPs also enhance their imaging and therapeutic applications.

- *Photothermal therapy (PTT)*

Gold NPs can absorb visible-NIR light efficiently owing to its structure, and a fast nonradioactive decay process can convert the absorbed energy into local heat energy. The local heat produced damages to abnormal cells significantly, especially cancerous cells because they are most prone to the scratch caused by heat than normal cells (Lal et al. 2008). This demonstrates the efficiency of gold NPs in initiating cell death in tumor cells. Nanostructures are used frequently for efficient PTT because the NIR state of light can cause maximum penetration of light into body tissues (Alkilany et al. 2012a, b; Chen et al. 2005; Li et al. 2015).

2.3.2 Role of Silk Proteins in Healthcare

Silks obtained from silkworms are classified into mulberry and non-mulberry silk. Irrespective of the sources (arachnids or insects), silk fibers are primarily made up of proteins associated with some macromolecules like polysaccharides and lipids (Hardy and Scheibel 2009). Fibroin and sericin are the two most primary proteins that make up the silk fibers, and they consist of 18 different amino acids, alanine, serine, and predominantly glycine. There is an interspecies variation of amino acid sequences of silk proteins which results in a wide range of mechanical properties. For example, the β-sheet regions of *B. mori* fibroin is dominated by a hexapeptide sequence, while repetitive stretches of polyalanine comprise those of *A. pernyi* and *S. cynthia ricini* (Mita et al. 1994; Yukuhiro et al. 1997). Silk proteins have found their roles in several aspects of healthcare and those have been discussed below.

2.3.2.1 Two-Dimensional Coatings and Films

Surface modification strategies can lead to the enhancement of life span and mechanical efficiency of biomedical devices or implants. Silks can be used as coating materials and the anticoagulant properties and compatibility of bio-devices can be influenced. Silk fibroin coatings have been reported to promote the adhesion of human fibroblasts on polyurethane and polycarbonate urethane surfaces (Jin et al. 2004; Arcidiacono et al. 2002) and osteoblasts on multilayered hydroxyapatite-fibroin films (Kino et al. 2007). Silk fibroin blends with polyallylamine have been widely used in biomedical and pharmaceutical industries, and they have increased the process ability and have enhanced water stability. Silk sericin also finds its role in wound dressing due to its possession of anticoagulant activity (Tamada et al. 2004; Akturk et al. 2011).

Silk fibroin coatings may further be coated with nanoparticles for enhancing their biocompatibility (Furuzono et al. 2004; Vachiraroj et al. 2009). While fibroin asserts in vitro antibacterial effects, its combination with silver or titanium dioxide nanoparticles leads to the inhibition of bacterial attachment in vitro, which leads to its successful applications as implants, surgical coatings, wound dressings, and cosmetics (Zhu et al. 2012).

Silk fibroin films have therapeutic implications and can support cell growth in cancer cell lines. For example, doxorubicin (DOX)-loaded silk films had chemo-therapeutic impact of primary tumor growth and metastasis in mice. They had a high primary tumor response and also they were not associated with any toxicity (Seib et al. 2012). Studies using fibroblast cell line demonstrate the ability of silk fibroin in supporting cell attachment, physiological morphology, and growth (Minoura et al. 1995). They can also induce corneal fibroblast elongation with strong expression of corneal ECM (Wu et al. 2012).

2.3.2.2 Implant Applications

Films of silk fibroin are utilized for in vivo wound healing, and the rate of healing was observed to be more rapid in skin wounds of rats and had lower inflammatory response in comparison with the traditional porcine-based wound dressings. Wound dressing using a blend of silk fibroin with chitin leads to the improvement of both keratinocytes and fibroblasts adhesion (Park et al. 2006). Electrospun matrices of collagen and silk can also be used to support human keratinocytes (Yeo et al. 2008). Other implantable applications of silk fibroin include cruciate ligament repair, tendon replacements, knee meniscus grafts, and the repair and regeneration of buccal mucosa (Mandal et al. 2011; Ge et al. 2012; Yang et al. 2007). Fibroin blood vessel constructs are preferred due to their larger diameters and greater mechanical strength (Lovett et al. 2008).

2.3.2.3 Delivery Vehicles

Architectures of silk are ideal vehicles to transport and deliver bioactive molecules across boundaries. Aqueous-based silk fibroin films promote the release of adenosine from the adenosine kinase lacking embryonic stem cell (Szybala et al. 2009) and also cause site-specific drug delivery in neural cells (Benfenati et al. 2010). Glucose oxidase, lipase, and horseradish peroxidase have been kept stable by these films for over a period of 10 months and thereby allow them to be stored at room temperature without losing its enzyme activity (Lu et al. 2009). Silk-hydroxypropyl methyl cellulose (HPMC)-polyethylene glycol (PEG) blended films are effective transmucosal delivery vehicles (Kundu et al. 2008), while silk/chitosan blends are used for the controlled discharge of low molecular weight drugs (Rujiravanit et al. 2003). Architectures of silk fibroin have been used as vehicles for controlled release. The drug loading in nano- or microparticles is carried out by adsorption or covalent coupling (Wenk et al. 2009), and this process regulates the stability of the drug, providing biological potency and release time (Shi and Goh 2012). Delivery of drugs and genes is carried out by composite silk sericin nano- and microparticles of methacrylate as well as chemotherapeutic agent's immobilization. Polyethylene glycol (PEG) composites can also be considered for transmucosal drug delivery. Silk fibroin-derived polypeptides can have a significant role to play in cell delivery for tissue engineering and bone regenerative therapy (Marelli et al. 2012).

2.4 Biomedical Engineering in the Diagnosis of Disease and Management

2.4.1 Dengue

Dengue is a mosquito-borne viral diseases and its incidence has increased extensively in the last five decades (WHO 2009a, b, c). Presently an ultimate detection of dengue infectants can be done in the in vitro condition either by detection of virus or identification of viral antigen (Gubler and Sather 1988). Isolation of virus and their nucleic acid detection is utilized for definitive identification at the early stage of illness; however, serological process is more appropriate when the acute stage of infection has been reached. Various commercial test kits have been developed and they operate through the detection of a combination of either antibodies or antigen (WHO 2009a, b, c). Several dengue patients get well impulsively though others might face a fatal fate due to critical loss of plasma (WHO 1997). Serological tests remain incapable for the diagnosis of the microvascular status of the infectants which is the foremost physiological alteration at the infection stage. Close monitoring of patients for the detection of the arrival of plasma and the prompt administration of intravenous fluid can reduce the fatality of this disease to a great extent (Ng et al. 2007).

2.4.2 Malaria

Malaria is an infective disease affecting both animals and human being and is spread by the infected female *Anopheles*. Fever and headache are the general signs of this disease; however certain cases also report death or coma of the patients. Africa, Asia, and Americas witness the most prevalence of this disease (WHO 2011). Light microscopy analysis of blood sample and the quick diagnostic method are two general procedures for identification of malaria infectants (WHO 2009a, b, c). The major benefit of the light microscopy is its low price, while the other testing kits are preferable due to its high potability (Abba et al. 2011).

2.4.3 Cholera

Vibrio cholera-infected food causes an acute intestinal infection called cholera. The period of incubation of this infection is short and lasts for 3 to 4 days. Severe infection may result in an enterotoxin leading to vomiting and diarrhea. The frequency of cholera disease has been increasing, and the growth of around 44% in the number of cases in the year 2010 and approximately 53% in the number of deaths in comparison with 2009 has been recorded. The mortality rate for the children and pregnant ladies has increased. Intravenous rehydration is the most common treatment protocol (Sack et al. 2001), and there are two types of vaccines which are accessible against this infection.

Individual diagnosis is considered impractical due to limitations like small accuracy, well-developed laboratory, and lengthy time period for diagnosis. The quick identification of *V. cholerae* can be done by dark-field microscopy. Different culturing methods can yield variable results using these assays, and therefore for the improvement of sensitivities and specificities of these assays, fluorescent monoclonal antibody and PCR-based methodology have been used (Alam et al. 2010).

2.4.4 Schistosomiasis

Schistosomiasis is basically a parasitic infection caused by the trematode parasite worms. This disease is categorized into two types: (a) urogenital schistosomiasis and (b) intestinal schistosomiasis. More infection each year is caused due to the quick multiplication and transformation of these schistosoma parasites. Furthermore, the mortality speed of this infection is high (WHO 2012). When parasites penetrate the skin exposed to infested water, schistosomiasis infection occurs. After that, the larva turns into mature schistosomes inside the human body. While immune reactions and tissue damages are caused by the eggs produced by the mature female worms inside the blood vessels, others get excreted through urination.

The foremost competent method for detection of schistosomiasis is the microscopic analysis of eggs in the sample of urine and stool. Kato-Katz procedure can be useful for the detection of *Schistosoma mansoni*. The number of eggs per gram of stool (Cheesbrough 1998) determines the severity of infection rate.

2.4.5 Ebola

The preliminary recognized infection occurred in Africa and a mortality rate of about 88% was reported (WHO 1976). Due to the frequent symptoms of Ebola including fever, vomiting, and weakness, it is often confused with other diseases like malaria (Nabel 2003). ELISA is formally used for its diagnosis, however it is not highly responsive for diverse stages of infection and is also not accessible easily in the areas where this disease is endemic (Leroy et al. 2000). The introduction of optical biosensor and reverse transcriptase polymerase chain reaction (RT-PCR) leads to the increase in the sensitivity and specificity of the disease. Petrosova et al. developed an optical immunosensor for the diagnosis of Ebola virus, and it is based on the technique of photo-immobilization (Petrosova et al. 2007). This sensor was fitted with a tip coated with indium tin oxide (ITO), followed by polypyrrole benzophenone, which is irradiated by means of light of specific wavelength and intensity to cause the excitation of benzophenone radicals which eventually bind to the Ebola virus antigen. The results detect the high efficiency of the immune sensor which is capable of detecting titers as low as around 1:970,000 and 1:2,000,000.00, respectively, for the Sudan and Zaire strains.

2.4.6 Leprosy

Leprosy is infected by *Mycobacterium leprae* and is a chronic disease (Marmor 2002). It is a communal health issue and its transmission makes it a foremost infective disease resulting in disabilities. The first symptom of leprosy is infectious skin, and if they are not treated at proper timing, it can lead to the damage of the eyes, skin, limbs, and nerves. Therefore early diagnosis and treatment is absolutely essential (Hussein et al. 2010). It is not possible to culture *Mycobacterium leprae* in the laboratory, and therefore scientists have been using various innovative techniques for studying host-pathogen interaction. However, no prime process for detection of leprosy in clinical care settings exists (Scollard et al. 2006). Diagnosis problems of leprosy are linked to the dangerous nature of leprosy along with its contradictory clinical, immunological, pathological, and immunological manifestations which cause problems with leprosy diagnosis. An effective diagnostic tool for leprosy is electroneuromyography (ENMG), which involves the study of muscle nerve stimulated with electric current. It has demonstrated 98% effectiveness among

leprosy patients and is especially significant in the identification of neural leprosy (Goulart and Goulart 2008).

2.5 Artificial Heart

The heart is the most vital organ that essentially works as an engine inside our body. The breakdown of the system rendering the pump less effectively is called heart failure, and today heart failure is almost an epidemic plaguing the world. Due to lack of reliable options in case of heart transplants, the latest case of success is that of the AbioCor artificial heart (Shire et al. 2015). The AbioCor heart comprises of an internal and external constituent. The component placed inside the body and is made up of:

- *Thoracic unit*: The thoracic unit is the external component of the artificial heart and has close similarity to the actual physical appearance of the organic heart. It contains inflow and outflow and weighs around 0.9 kg. The thoracic unit is placed in place of the natural heart inside the body and is made up of two hydraulic motors. One of the motors pumps the blood continuously, while the other motors operate the motion of the four heart valves.
- *Implanted TET*: TET (transcutaneous energy transmission) provides electrical energy to the AbioCor system.
- *Implanted controller*: Looks after the AbioCor system and intimates the patient in case of detection of any problem.
- *Implanted battery*: Provides internal system with energy of 30–40 min. It is required to be replaced in every 1 year and need to go through a minor surgical procedure.

The outer constituent of the artificial heart is the outer TET that is positioned on the inner TET over the skin, a console providing power to the external and the internal TET, PCE (patient-carried electronics), PCE battery pack, batteries, and PCE control module.

2.6 Conclusion

Solutions to limited resources required creative engineering and developed technical ability for keeping the expenses low and also improving the time of response. Integration of biomedical technological proficiency with various institutional expertises will not only aid in rapid disease diagnosis but also will strengthen the system of national disaster management. Near future, huge investment in this biomedical engineering sector will lead for the generation of self-sustaining and proficient procedure in human health services at the times of crisis.

References

Abba K, Deeks JJ, Olliaro P, Naing CM, Jackson SM, Takwoingi Y, Donegan S, Garner P (2011) Rapid diagnostic tests for diagnosing uncomplicated *p. Falciparum* malaria in endemic countries. Cochrane Database Syst Rev 7:1–241

Akturk O, Tezcaner A, Bilgili H, Deveci MS, Gecit MR, Keskin D (2011) Evaluation of sericin/collagen membranes as prospective wound dressing biomaterial. J Biosci Bioeng 112:279–288

Alam M, Hasan NA, Sultana M, Nair GB, Sadique A, Faruque A, Endtz HP, Sack R, Huq A, Colwell R (2010) Diagnostic limitations to accurate diagnosis of cholera. J Clin Microbiol 48:3918–3922

Alkilany L, Alberti KP, Mondonge V, Rauzier J, Quilici M-L, Guerin PJ (2012a) Evaluation of a rapid test for the diagnosis of cholera in the absence of a gold standard. PLoS One 7. https://doi.org/10.1371/journal.Pone.0037360

Alkilany M, Thompson LB, Boulos SP, Sisco PN, Murphy CJ (2012b) Adv Drug Deliv Rev 64:190–199

Arcidiacono S, Mello CM, Butler M, Welsh E, Soares JW, Allen A, Ziegler D, Laue T, Chase S (2002) Aqueous processing and fiber spinning of recombinant spider silks. Macromolecules 35:1262–1266

Bal B, Armstrong PB, Das AP (2016) Development of indigenous bio-sensing methodology for rapid and low cost endotoxin detection system. Sens Netw Data Commun S1. 005. https://doi.org/10.4172/2090-4886.S1-005

Bal B, Nayak S, Das AP (2017) Recent advances in molecular techniques for the diagnosis of foodborne diseases. In: Oprea AE, Grumezescu A (eds) Nanotechnology applications in food. Elsevier

Benfenati V, Toffanin S, Capelli R, Camassa LMA, Ferroni S, Kaplan DL, Omenetto FG, Muccini M, Zamboni R (2010) A silk platform that enables electrophysiology and targeted drug delivery in brain astroglial cells. Biomaterials 31:7883–7891

Bethell DB, Gamble J, Loc PP, Dung NM, Chau TTH, Loan HT, Thuy TTN, Tam DTH, Gartside IB, White NJ (2001) Noninvasive measurement of microvascular leakage in patients with dengue hemorrhagic fever. Clin Infect Dis 32:243–253

Cheesbrough M (1998) Parasitological tests, district laboratory practice in tropical countries, part 1. Cambridge University Press, Cambridge

Chen J, Wiley B, Li WY, Campbell D, Saeki F, Cang H, Au L, Lee J, Li X, Xia Y (2005) Adv Mater 17:2255–2261

Choi CL, Alivisatos AP (2010) Annu Rev Phys Chem 61:369–389

Das AP, Sukla LB, Pradhan N, Nayak S (2011) Manganese biomining: a review. Bioresour Technol 102(16):7381–7387

Das AP, Kumar PS, Swain S (2014) Recent advances in biosensor based endotoxin detection. Biosens Bioelectron 51:62–75

Das AP, Bal B, Mahapatra PS (2015) Chromogenic biosensors for pathogen detection. In: Biological and pharmaceutical applications of nanomaterials. CRC Press

Das AP, Ghosh S, Bal B, Nayak S (2016) Advanced Nanosensors for detection food pathogens and toxins. In: Nanobiosensors (nanotechnology in the food industry). Elsevier

Furuzono T, Kishida A, Tanaka J (2004) Nano-scaled hydroxyapatite/polymer composite I. Coating of sintered hydroxyapatite particles on poly (gamma-methacryloxypropyl trimethoxysilane)-grafted silk fibroin fibers through chemical bonding. J Mater Sci Mater Med 15:19–23

Gamble J, Bethell D, Day N, Loc P, Phu N, Gartside I, Farrar J, White N (2000) Age-related changes in microvascular permeability: a significant factor in the susceptibility of children to shock? Clin Sci 98:211–216

Ge Z, Yang Q, Xiang X, Liu KZ (2012) Assessment of silk fibroin for the repair of buccal mucosa in a rat model. Int J Oral Maxillofac Surg 41:673–680

Goulart IMB, Goulart LR (2008) Leprosy: diagnostic and control challenges for a worldwide disease. Arch Dermatol Res 300:269–290

Gubler DJ, Sather GE (1988) Laboratory diagnosis of dengue and dengue hemorrhagic fever. In: Proceedings of international symposium on yellow fever and dengue, Rio de Janeiro, Brazil, 15–19 May, pp 291–322
Hainfeld JF, Slatkin DN, Focella TM, Smilowitz HM (2006) Brit J Radiol 79:248–253
Hardy JG, Scheibel TR (2009) Silk-inspired polymers and proteins. Biochem Soc Trans 37:677–681
Hassan H, Taib MN, Ibrahim F, Abas WABW (2003) Cutaneous microcirculatory flowmetry evaluated by laser Doppler technique in dengue hemorrhage fever patients. In: Proceedings of the Asian conference on sensors, 2003, Kuala Lumpur, Malaysia, 18 July 2003, pp 325–328
Hussein A, Mohammed H, Eltahir A, Sidig A, Gadour M (2010) Frequency of neurological deficits in sudanese lepromatic patients. Sudan. J Med Sci 5:16–24
Jeon J, Lim D-K, Nam J-M (2009) Synthesis of shaped particles and particle arrays by disassembly methods. J Mater Chem 19:2107–2117
Jin HJ, Chen JS, Karageorgiou V, Altman GH, Kaplan DL (2004) Human bone marrow stromal cell responses on electrospun silk fibroin mats. Biomaterials 25:1039–1047
Khan MS, Vishakante GD, Siddaramaiah H (2013) Adv Colloid Interf Sci 199–200:44–58
Kino R, Ikorna T, Yunoki S, Nagai N, Tanaka J, Asakura T, Munekata M (2007) Preparation and characterization of multilayered hydroxyapatite/silk fibroin film. J Biosci Bioeng 103:514–520
Kumar MS, Das AP (2017) Emerging nanotechnology based strategies for diagnosis and therapeutics of urinary tract infections: a review. Adv Colloid Interf Sci 249:53–65
Kundu J, Patra C, Kundu SC (2008) Design, fabrication and characterization of silk fibroin- HPMC-PEG blended films as vehicle for transmucosal delivery. Mat Sci Eng C 28:1376–1380
Kvernebo K, Staxrud L, Salerud E (1990) Assessment of human muscle blood perfusion with single-fiber laser Doppler flowmetry. Microvasc Res 39:376–385
Lal S, Clare SE, Halas NJ (2008) Acc Chem Rev 41:1842–1851
Lee N, Choi SH, Hyeon T (2013) Nano-sized CT contrast agents. Adv Mater 25:2641–2660
Leroy E, Baize S, Lu C, McCormick J, Georges A, Georges-Courbot MC, Lansoud-Soukate J, Fisher-Hoch S (2000) Diagnosis of ebola haemorrhagic fever by RT-PCR in an epidemic setting. J Med Virol 60:463–467
Li R, Cai N, Kawazoe, Chen G (2015) Metallic nanoparticles as synthetic building blocks for cancer diagnostics: from materials design to molecular imaging applications. J Mater Chem B 3:5806–5581
Lovett ML, Cannizzaro CM, Vunjak-Novakovic G, Kaplan DL (2008) Gel spinning of silk tubes for tissue engineering. Biomaterials 29:4650–4657
Lu S, Wang X, Lu Q, Hu X, Uppal N, Omenetto FG, Kaplan DL (2009) Stabilization of enzymes in silk films. Biomacromolecules 10:1032–1042
Mandal BB, Park SH, Gil ES, Kaplan DL (2011) Multilayered silk scaffolds for meniscus tissue engineering. Biomaterials 32:639–651
Marelli B, Ghezzi CE, Alessandrino A, Barralet JE, Freddi G, Nazhat SN (2012) Silk fibroin derived polypeptide-induced biomineralization of collagen. Biomaterials 33:102–108
Marmor MF (2002) The ophthalmic trials of GHA Hansen. Surv Ophthalmol 47:275–287
Minoura N, Aiba SI, Higuchi M, Gotoh Y, Tsukada M, Imai Y (1995) Attachment and growth of fibroblast cells on silk fibroin. Biochem Biophys Res Commun 208:511–516
Mita K, Ichimura S, James TC (1994) Highly repetitive structure and its organization of the silk fibroin gene. J Mol Evol 38:583–592
Nabel GJ (2003) Vaccine for aids and ebola virus infection. Virus Res 92:213–217
Ng CFS, Lum LCS, Ismail NA, Tan LH, Tan CPL (2007) Clinicians' diagnostic practice of dengue infections. J Clin Virol 40:202–206
Olavi A, Kolari P, Esa A (1991) Edema and lower leg perfusion in patients with post-traumatic dysfunction. Acupunct Electr Ther Res 16:7–11
Park K, Kuechle MK, Choe Y, Craik CS, Lawrence OT, Presland RB (2006) Expression and characterization of constitutively active human caspase-14. Biochem Biophys Res Commun 347:941–948

Pelupessy J, Allo E, Jota S (1989) Pericardial effusion in dengue haemorrhagic fever. Paediatr Indones 29:72–75

Petrosova A, Konry T, Cosnier S, Trakht I, Lutwama J, Rwaguma E, Chepurnov A, Mühlberger E, Lobel L, Marks R (2007) Development of a highly sensitive, field operable biosensor for serological studies of ebola virus in central Africa. Sens Actuators B Chem 122:578–586

Rujiravanit R, Kruaykitanon S, Jamieson AM, Tokura S (2003) Preparation of crosslinked chitosan/silk fibroin blend films for drug delivery system. Macromol Biosci 3:604–611

Sack DA, Lyke C, McLaughlin C, Suwanvanichkij V (2001) Antimicrobial resistance in shigellosis, cholera, and campylobacteriosis. World Health Organization, Geneva

Scollard DM, Adams LB, Gillis TP, Krahenbuhl JL, Truman RW, Williams DL (2006) The continuing challenges of leprosy. Clin Microbiol Rev 19:338–381

Seib FP, Maitz MF, Hu XA, Werner C, Kaplan DL (2012) Impact of processing parameters on the haemocompatibility of *Bombyx mori* silk films. Biomaterials 33:1017–1023

Setiawan MW, Samsi TK, Wulur H, Sugianto D, Pool TN (1998) Dengue haemorrhagic fever: ultrasound as an aid to predict the severity of the disease. Pediatr Radiol 28:1–4

Shi PJ, Goh JCH (2012) Self-assembled silk fibroin particles: tunable size and appearance. Powder Technol 215–216:85–90

Shire A, Jawarkar U, Sonone M (2015) A review paper: design of ABIOCOR artificial heart. Int J Innov Sci Eng Technol 2(1)

Srikiatkhachorn A, Krautrachue A, Ratanaprakarn W, Wongtapradit L, Nithipanya N, Kalayanarooj S, Nisalak A, Thomas SJ, Gibbons RV, Mammen MP Jr (2007) Natural history of plasma leakage in dengue hemorrhagic fever: a serial ultrasonographic study. Pediatr Infect Dis J 26:283–290

Sun L, Zhang J, Lu X, Zhang L, Zhang Y (2011) Evaluation to the antioxidant activity of total flavonoids extract from persimmon (Diospyros kaki L.) leaves. Food Chem Toxicol 49(10):2689–2696

Szybala C, Pritchard EM, Lusardi TA, Li TF, Wilz A, Kaplan DL, Boison D (2009) Antiepileptic effects of silk-polymer based adenosine release in kindled rats. Exp Neurol 219:126–135

Tamada Y, Sano M, Niwa K, Imai T, Yoshino G (2004) Sulfation of silk sericin and anticoagulant activity of sulfated sericin. J Biomat Sci Polym E 15:971–980

Vachiraroj N, Ratanavaraporn J, Damrongsakkul S, Pichyangkura R, Banaprasert T, Kanokpanont S (2009) A comparison of Thai silk fibroin-based and chitosan-based materials on in vitro biocompatibility for bone substitutes. Int J Biol Macromol 45:470–477

Venkata Sai PM, Dev B, Krishnan R (2005) Role of ultrasound in dengue fever. Br J Radiol 78:416–418

Wali J, Biswas A, Chandra S, Malhotra A, Aggarwal P, Handa R, Wig N, Bahl V (1998) Cardiac involvement in dengue haemorrhagic fever. Int J Cardiol 64:31–36

Wang H, Zheng L, Peng C, Shen M, Shi X, Zhang G (2013) Biomaterials 34:470–480

Wax A, Sokolov K (2009) Laser Photon Rev 3:146–170 158

Wenk E, Meinel AJ, Wildy S, Merkle HP, Meinel L (2009) Microporous silk fibroin scaffolds embedding PLGA microparticles for controlled growth factor delivery in tissue engineering. Biomaterials 30:2571–2581

World Health Organization (1976) International study team for percentage of cases in health care workers and comments for Nzara, and Maridi area, Sudan, 1976. WHO, Geneva

World Health Organization (1997) Dengue haemorrhagic fever diagnosis, treatment, prevention and control, 2nd edn. WHO, Geneva

World Health Organization (2009a) Malaria case management: operations manual. WHO, Geneva

World Health Organization (2009b) Dengue haemorrhagic fever: diagnosis, treatment, prevention and control, New edn. WHO, Geneva

World Health Organization (2009c) Special Programme for Research and Training in Tropical Diseases (Tdr): dengue guidelines for diagnosis, treatment, prevention and control. WHO, Geneva

World Health Organization (2011) World malaria report. WHO, Geneva

World Health Organization (2012) Lymphatic Filariasis, fact sheets. WHO, Geneva
Wu J, Du YQ, Watkins SC, Funderburgh JL, Wagner WR (2012) The engineering of organized human corneal tissue through the spatial guidance of corneal stromal stem cells. Biomaterials 33:1343–1352
Yang Y, Ding F, Wu J, Hu W, Liu W, Liu J, Gu X (2007) Development and evaluation of silk fibroin-based nerve grafts used for peripheral nerve regeneration. Biomaterials 28:5526–5535
Yeo IS, Oh JE, Jeong L, Lee TS, Lee SJ, Park WH, Min BM (2008) Collagen-based biomimetic nanofibrous scaffolds: preparation and characterization of collagen/silk fibroin bicomponent nanofibrous structures. Biomacromolecules 9:1106–1116
Yukuhiro K, Kanda T, Tamura T (1997) Preferential codon usage and two types of repetitive motifs in the fibroin gene of the Chinese oak silkworm, Antheraea pernyi. Insect Mol Biol 6:89–95
Yusoff K, Roslawati J, Sinniah M, Khalid B (1993) Electrocardiographic and echocardiographic changes during the acute phase of dengue infection in adults. J Hong Kong Coll Cardiol 1:93–96
Zeng S, Baillargeat D, Ho H-P, Yong K-T (2014) Nanomaterials enhanced surface plasmon resonance for biological and chemical sensing applications. Chem Soc Rev 43:3426–3452
Zhu YR, Chen YY, Xu GH, Ye XJ, He DN, Zhong J (2012) Micropattern of nano- hydroxyapatite/silk fibroin composite onto Ti alloy surface via template-assisted electrostatic spray deposition. Mat Sci Eng C 32:390–394

Chapter 3
Integrative Omics for Interactomes

Debangana Chakravorty, Krishnendu Banerjee, and Sudipto Saha

Abstract Single-layer omics provide limited insight, whereas integrated multi-omics layers allow understanding of their combined influence on the complex biological process. The integrative omics approach has been initially applied to cancer research and later used in understanding host-pathogen interactions and pluripotency regulatory networks in stem cells. Here, different multi-omics layers along with databases and tools specific for multiple data integration, visualization, and integrated network modeling are described. In summary, this chapter focuses on integrative analysis of different multi-omics layers and modeling of interactomes to identify robust biomarkers and biological processes associated with diseases.

Keywords Multi-omics · Protein-protein interactions · TCGA · CPTAC · Integrative analysis

3.1 Introduction

The initial multi-omics data was generated by The Cancer Genome Atlas (TCGA) project on different tumors and cancer cell lines. It provided a comprehensive genomics profiles including genetic mutations, gene expression, microRNA, copy number, and methylation data of 32 types of human tumors. This genomics dataset was possible due to the availability of next-generation sequencing (NGS) technology that provided the complete genome-wide coverage with low cost. After that, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) used the same TCGA tumor samples and generated tandem mass spectrometry (MS/MS)-based proteomics data. All these multi-omics data from TCGA and CPTAC projects were analyzed and stored in LinkedOmics database (Vasaikar et al. 2018). Detailed proteogenomics analysis was performed in TCGA breast cancer samples, where functional consequences of somatic mutations were reported (Mertins et al. 2016). Large-scale protein-protein interactions of human and other model organisms were

D. Chakravorty · K. Banerjee · S. Saha (✉)
Bioinformatics Centre, Bose Institute, Kolkata, India

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_3

generated using affinity purification followed by mass spectrometry and yeast two-hybrid-based techniques (Ewing et al. 2007; Rual et al. 2005; Krogan et al. 2006; Uetz et al. 2000). Multi-omics data was not only restricted to cancer, but there were other applications of multi-omics integrative studies such as understanding host-pathogen interaction (Jean Beltran et al. 2017), host signaling regulation by the gut microbiota (Manes et al. 2017), and pluripotency regulatory network in embryonic stem cells (Stumpf et al. 2016).

There are several bioinformatics tools available for integrating, visualizing, and modeling multi-omics data and networks. Bayesian support vector machine and clustering methods have been used to integrate the data of mixed types (Yifeng et al. 2016). Cytoscape is an open-source software that can be used for visualizing the integrated networks (Cline et al. 2007). Network-based approaches used graph theory to integrate multiple homogeneous networks (e.g., protein-protein interaction), where node represents gene or protein and edge represents interaction. There can be two different types of interaction in heterogeneous networks (e.g., protein-protein, protein-DNA and DNA-metabolite interactions), one is the intraspecies interaction (protein-protein) and the other is the interspecies interaction (protein-DNA). The latter interaction is mainly involved in cross talk among multiple layers of the interactome. In summary, multi-omics approaches along with bioinformatics tools allow the integration of data generated from different omic levels and aid in understanding the complex and wired biological networks.

This chapter will first highlight different multi-omics layers and four different types of integrative analysis of multi-omics datasets, including (1) integrative analysis of genomics, epigenomics, and transcriptomics data; (2) integrative analysis of transcriptomics, proteomics, and protein interaction networks; (3) integrative analysis of transcriptomics and metabolomics; and (4) integrative analysis of multi-omics data. Next, the databases and tools used for multi-omics studies will be presented. And finally, the future perspectives and challenges of integrative omics studies will be discussed.

3.2 Multi-omics Layers

A single layer of "omics" including genomics, epigenomics, transcriptomics, proteomics, and metabolomics provides specific insight of DNA, RNA, protein, and metabolite level into the biological process of a cell. Genomics, involving the sequencing and analysis of genomes, uses high-throughput DNA sequencing such as next-generation sequencing (NGS), whole-genome sequencing (WGS), whole-exome sequencing, real-time PCR (RT-PCR), and single nucleotide polymorphism (SNP) along with bioinformatics to assemble and analyze the function and structure of the entire genomes (Concepts of genetics 2012; Culver and Labow 2002). Epigenomics, on the other hand, involves the study of reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence. The study of epigenetics on a global level has been made possible only

recently through the adaptation of genomic high-throughput assays such as chromatin immunoprecipitation followed by microarray (ChIP-chip), chromatin immunoprecipitation followed by sequencing (ChIP-seq), methylated DNA immunoprecipitation (Me-DIP) (Friedman and Rando 2015), and ATAC-seq (Buenrostro et al. 2013). Transcriptomics refers to the study of the information content of an organism present in DNA, which includes mRNA and noncoding RNAs such as tRNA, rRNA, microRNA, and long ncRNA. The various RNA pools differ dramatically in abundance relative to each other and can change across experimental conditions (Yang et al. 2011). The standard protocol for transcriptome analysis involves RNA extraction, reverse transcription, cDNA amplification using quantitative reverse transcription-PCR (qRT-PCR), and hybridization using microarrays followed by library construction and sequencing (RNA-Seq). Proteomics refers to the large-scale analysis of the whole set of proteins which has significantly benefited from the Human Genome Project, accumulation of both DNA and protein sequence databases, improvements in mass spectrometry, and the development of computer algorithms for database searching (Graves and Haystead 2002). Metabolomics aims to measure the low molecular weight compounds called metabolites. The metabolome composition reflects the current status of the organism and is considered to be a chemical reflection of a molecular phenotype (Bujak et al. 2015). Numerous analytical platforms are commonly used in both targeted and untargeted metabolomic studies such as nuclear magnetic resonance (NMR) and mass spectrometry (MS), coupled with different separation techniques (Lindon and Nicholson 2008).

Multi-omics approaches integrate data from different omics levels to understand their combined influence on the biological process. For example, pluripotent stem cells show a high degree of regulation between multiple species of molecules. Studies have shown that the pluripotent state in mouse and human cells is regulated at multiple levels, including transcriptional (Boyer et al. 2005), epigenetic (Lee et al. 2006), signaling (Chen et al. 2008), and metabolic (Moussaieff et al. 2015) layers. Studies by Stumpf et al. shows that in the presence of external stimuli (Ying et al. 2008), the pluripotent state is maintained by a set of TFs, Oct4, Sox2, and Nanog along with secondary factors such as Klf4, Myc, and Lin28 (MacArthur et al. 2012). These core TFs interact with a range of auxiliary TFs via PPIs (Wang et al. 2006) and collectively control transcription of a large number of genes. Transcriptional control is exerted either directly, by binding to gene promoters (Boyer et al. 2005), or indirectly, by mediating the effects of epigenetic remodeling complexes (Orkin and Hochedlinger 2011). To add to this is a network of microRNAs (Wang et al. 2007) which ensures that appropriate protein levels are robustly maintained. Collectively, these reports indicate that pluripotency is regulated by cross talk among multi-omics layers to form interactome (Fig. 3.1) and involves layers of combinatorial regulatory control, including complex feedback relationships between the transcriptional, epigenetic, and signaling strata. Thus, the cross talk between multi-omics layers cannot be determined by single omics reduction approach.

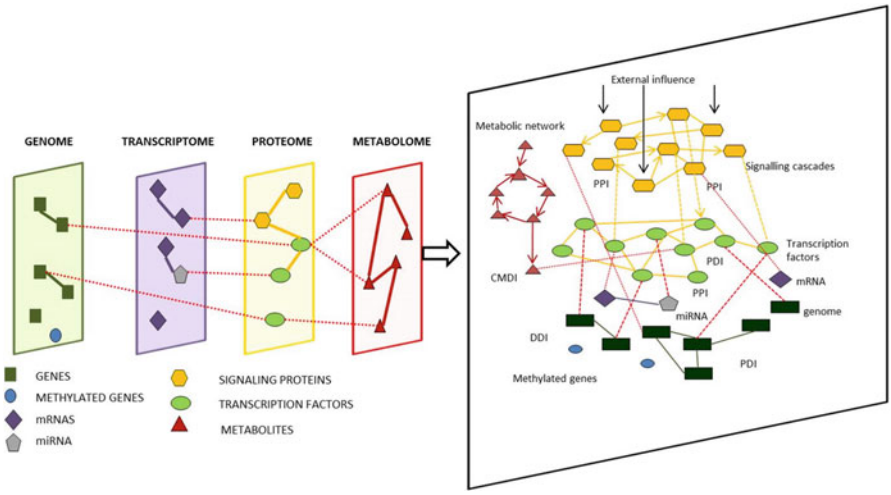

Fig. 3.1 Cross talk of the multi-omics layer to form interactome. Single omics approach like genomics, epigenomics, transcriptomics, proteomics, and metabolomics integrated by combining the interconnections of all layers within species and across species. The molecular species involved at each level is illustrated with nodes of different shapes and colors, and a key is provided below. The edges in dotted red lines show intermolecular species interaction, while the edges solid lines and color matched with nodes show intraspecies interaction, whereas solid arrows indicate external influence. (Partially adapted from Stumpf et al. Proteomics 2016)

3.2.1 Integrative Analysis of Genomics, Epigenomics, and Transcriptomics Data

The TCGA data provides comprehensive genomics profiles including genetic mutations, gene expression, microRNA sequencing, and copy number alterations of over 30 human tumors. Thus, the TCGA data is well studied for integrating multi-omics datasets. The effect of copy number alterations (CNA) on mRNA levels was studied in breast cancer samples, and it was seen that 64% of all genes studied have a positive correlation between CNA and mRNA levels (Mertins et al. 2016). In another study of integrative analysis in liver cancer, it was observed that cancer gene expression could be correlated with DNA copy number (CNVcor) and with DNA methylation (METcor) (Woo et al. 2017). Expression profiles of these CNVcor and METcor genes were able to predict subgroups in hepatocellular cancer. There are few bioinformatics tools available for integrating genomics, epigenomics, and transcriptomics datasets like DINGO, BioWardrobe, and mixOmics (Ha et al. 2015; Kartashov and Barski 2015; Rohart et al. 2017). These tools allow building differential networks and identifying common hub genes found in expression datasets of multiple layers.

3.2.2 *Integrative Analysis of Transcriptomics, Proteomics, and Protein Interaction Networks*

Integrating transcriptomics and proteomics data with protein interaction networks have been used for discovery of biomarkers and novel biological processes. In the field of biomarker discovery, the overlapping genes and proteins observed in multiple layers are common targets or a part of a feedback loop and so possibly better targets for therapeutics (Chakravorty et al. 2017). In a study by Mertins et al. from Broad Institute, results show a correlation between protein expression and gene expression across breast cancer samples taken from TCGA data (Mertins et al. 2016). These results demonstrate the utility of integrated transcriptome and proteome analysis for confirmation of regulatory mechanisms and identification of candidate regulators.

There is a higher coverage of transcriptome data as compared with mass spectrometry-based proteomics approach. Thus, gene expression datasets are merged with protein-protein interaction (PPI) network for the identification of novel biological process and active subnetworks as shown in Fig. 3.2. NetworkAnalyst and jActiveModules allow to merge gene expression and PPI networks. This approach has been studied for a better understanding of cancer and host-pathogen interactions (Jean Beltran et al. 2017; Saha and Ewing 2011).

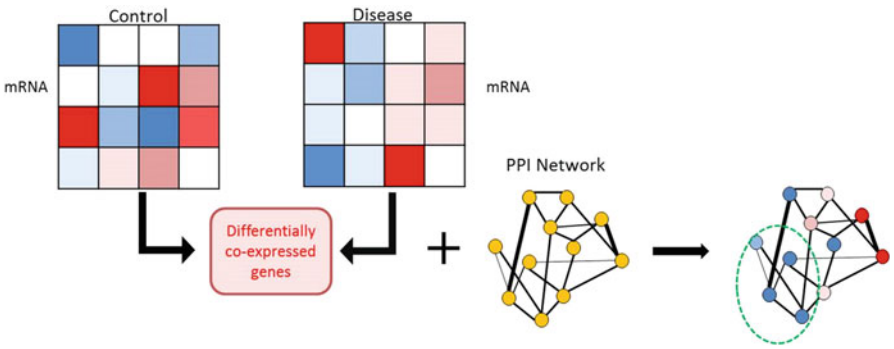

Fig. 3.2 Integrating transcriptomics and proteomics data to generate an integrative omics network. The gene expression profile from microarray data of disease versus control is combined with protein-protein interaction network to generate an integrative network. The red and blue color gradients represent overexpression and under-expression of differentially regulated genes. Yellow node indicates the proteins, and black lines indicate the edges of the PPI network, and thicknesses of the edges indicate the confidence of interaction. In the integrative network, the red to blue color gradients indicate the gene expression profile of the proteins involved in PPI. Green dotted circle highlights the subnetwork

3.2.3 Integrative Analysis of Transcriptomics and Metabolomics

Metabolomics is an important functional layer in studying multi-omics datasets, since it links genotype to phenotype. Integrative approaches for metabolomics and transcriptomics have been well established in the plant system (Urbanczyk-Wochniak et al. 2003). Datasets from metabolomics and transcriptomics studies are integrated using the correlation-based method, multivariate-based method that uses partial least square (PLS) regression and principal component analysis (PCA), and finally pathway-based method (Cavill et al. 2016). Integrated Molecular Pathway-Level Analysis (IMPaLA) is a web-based freely available tool frequently used for integration of two types of datasets (Kamburov et al. 2011). Other tools like Metscape 2 and Paintomics also perform similar kind of integrative analysis.

3.2.4 Integrative Analysis of Multi-omics Data

With the availability of TCGA and LinkedOmics resources, analyzing multi-omics dataset is possible. Various bioinformatics tools like Lemon-Tree and Omics Integrator allow network-based interpretation of multi-omics datasets (Bonnet et al. 2015). These are open-source, platform-independent and allow integrating multiple types of high-throughput datasets for creating networks.

3.3 Databases and Tools Used for Multi-omics Data

3.3.1 Database

Several databases contain multi-omics data as shown in Table 3.1. The first multi-omics database is The Cancer Genome Atlas (TCGA) that provides an interactive data system for researchers to search, download, upload, and analyze various cancer genomic datasets (Wang et al. 2016). The Library of Integrated Network-Based Cellular Signatures (LINCS) program provides an extensive reference library of cell-based perturbation-response signatures (Koleti et al. 2018). The LinkedOmics database includes information about mass spectrometry-based global proteomics data on TCGA tumor samples along with clinical data (Vasaikar et al. 2018). Multi-Omics Profiling Expression Database (MOPED) contains processed data for gene, protein, and pathway expression of human and model organism (Montague et al. 2015). Very few organ-specific diseases like heart and kidney diseases have multi-omics databases available (Alexandar et al. 2015; Fernandes and Husi 2017). Taken together, most of the integrative resources compiled various types of multi-omics datasets of tumors and cancer cell lines.

Table 3.1 Databases of multi-omics studies

Database name	Brief description	References
CardioGenBase	A literature-based multi-omics database for major cardiovascular diseases	Alexandar et al. (2015)
CKDdb	An integrative multi-omics expression database of chronic kidney disease (CKD)	Fernandes and Husi (2017)
LINCS	Library of Integrated Network-Based Cellular Signatures	Koleti et al. (2018)
LinkedOmics	Integrates mass spectrometry (MS)-based global proteomics data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) on selected TCGA tumor samples	Vasaikar et al. (2018)
MOPED	A freely accessible multi-omics expression database	Montague et al. (2015)
TCGA	An archive of genomic sequence, expression, methylation, and copy number variation data over 30 different types of cancer	Wang et al. (2016)

3.3.2 Tools

The availability of multi-omics cancer data from the same samples from TCGA allows developing various tools specific for multiple data integration, visualization, and integrated network modeling. The list of software dedicated for multi-omics data along with brief description is shown in Table 3.2. Tools like DINGO, BioWardrobe, and mixOmics are used for integrated analysis of mRNA/miRNA expression, DNA copy number, and methylation (Ha et al. 2015; Kartashov and Barski 2015; Rohart et al. 2017). Tools like jActiveModules (Cytoscape plugins) and NetworkAnalyst are used for integrating gene expression and PPI networks. Similarly, there are tools like Metscape 2 and Paintomics for integration of mRNA expression and metabolites data (Karnovsky et al. 2012; Garcia-Alcalde et al. 2011) and tools like ZikaVR and Immunet for Zika virus and immunological disease research, respectively (Gorenshteyn et al. 2015; Gupta et al. 2016). Omics Integrator software integrates several types of omics data and constructs a heterogeneous network of phosphorylated proteins, metabolites, and mRNA expression (Tuncbag et al. 2016). Lemon-Tree software uses large-scale multi-omics datasets and predicts network modules and pathways (Bonnet et al. 2015). In summary, there are several integrative analysis tools for multi-omics datasets and inferring network modules and pathways for understanding complex biological processes.

Table 3.2 Software for integrating multi-omics studies

Software	Integrating omics	References
BioWardrobe	mRNA expression and DNA methylation	Kartashov and Barski (2015)
DINGO	mRNA expression, DNA copy number, methylation, and microRNA expression	Ha et al. (2015)
Immunet	mRNA expression, protein-protein interaction, miRNA motif profiles, and transcription factor motif profiles	Gorenshteyn et al. (2015)
IMPaLA	Web server for integrating transcriptomics and metabolic datasets	Kamburov et al. (2011)
jActiveModules	Genetic or protein-protein interaction network and mRNA expression	Cline et al. (2007)
Lemon-Tree	Integrative multi-omics module network inference	Bonnet et al. (2015)
Metscape 2	Metabolites and gene expression data	Karnovsky et al. (2012)
mixOmics	mRNA expression, miRNA expression, DNA methylation, and protein	Rohart et al. (2017)
NetworkAnalyst	Gene expression data and protein-protein interaction	Xia et al. (2015)
Omics Integrator	Integrates a variety of omics data and identifies putative molecular pathways	Tuncbag et al. (2016)
Paintomics	Metabolites and mRNA expression	Garcia-Alcalde et al. (2011)
ZikaVR	Gene, protein, miRNA, siRNA, and shRNA	Gupta et al. (2016)

3.4 Future Prospective and Challenges

The primary requirement of the integrative multi-omics is that all the omics studies have to be performed in the same sample. So, there are few challenges in integrating multi-omics datasets. First, for integrating protein-protein interactions data, it was observed that most of the data available was from HEK 293 cell line in case of AP-MS studies. There is a considerable gap in generating PPIs of all the proteins from other human cell lines and tissues including healthy and diseased states. Second, for integrating metabolomics and transcriptomics data, it was seen that the metabolites are mainly isolated from blood or urine, while transcriptomics data can be derived from all tissue samples related to the disease. As there is a need for experimental sample source parity, there is also the need for establishing data processing standards and data normalization procedures across different omics layers. So far, most of the multi-omics studies are mainly focused on tumor and cancer cell lines. Besides cancer, there are various diseases like respiratory and cardiac diseases, which need urgent attention for understanding biological mechanisms of these diseases using integrative analysis of multi-omics data.

References

Alexandar V, Nayar PG, Murugesan R, Mary B, Darshana P et al (2015) CardioGenBase: a literature based multi-omics database for major cardiovascular diseases. PLoS One 10: e0143188

Bonnet E, Calzone L, Michoel T (2015) Integrative multi-omics module network inference with Lemon-Tree. PLoS Comput Biol 11:e1003983

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS et al (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122:947–956

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10:1213–1218

Bujak R, Struck-Lewicka W, Markuszewski MJ, Kaliszan R (2015) Metabolomics for laboratory diagnostics. J Pharm Biomed Anal 113:108–120

Cavill R, Jennen D, Kleinjans J, Briede JJ (2016) Transcriptomic and metabolomic data integration. Brief Bioinform 17:891–901

Chakravorty D, Jana T, Das Mandal S, Seth A, Bhattacharya A et al (2017) MYCbase: a database of functional sites and biochemical properties of Myc in both normal and cancer cells. BMC Bioinformatics 18:224

Chen X, Xu H, Yuan P, Fang F, Huss M et al (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133:1106–1117

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N et al (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2:2366–2382

Concepts of genetics (2012) (10th ed.). San Francisco: Pearson Education. ISBN:978-0-321-72412-0

Culver KW, Labow MA (2002) Genomics. In: Robinson R (ed) Genetics. Macmillan Science Library. Macmillan Reference USA, New York ISBN:978-0-02-865606-9

Ewing RM, Chu P, Elisma F, Li H, Taylor P et al (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 3:89

Fernandes M, Husi H (2017) Establishment of a integrative multi-omics expression database CKDdb in the context of chronic kidney disease (CKD). Sci Rep 7:40367

Friedman N, Rando OJ (2015) Epigenomics and the structure of the living genome. Genome Res 25:1482–1490

Garcia-Alcalde F, Garcia-Lopez F, Dopazo J, Conesa A (2011) Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics 27:137–139

Gorenshteyn D, Zaslavsky E, Fribourg M, Park CY, Wong AK et al (2015) Interactive big data resource to elucidate human immune pathways and diseases. Immunity 43:605–614

Graves PR, Haystead TA (2002) Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev 66:39–63

Gupta AK, Kaur K, Rajput A, Dhanda SK, Sehgal M et al (2016) ZikaVR: an integrated Zika virus resource for genomics, proteomics, Phylogenetic and Therapeutic Analysis. Sci Rep 6:32713

Ha MJ, Baladandayuthapani V, Do KA (2015) DINGO: differential network analysis in genomics. Bioinformatics 31:3413–3420

Jean Beltran PM, Federspiel JD, Sheng X, Cristea IM (2017) Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases. Mol Syst Biol 13:922

Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics 27:2917–2918

Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G et al (2012) Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 28:373–380

Kartashov AV, Barski A (2015) BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data. Genome Biol 16:158

Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ et al (2018) Data portal for the library of integrated network-based cellular signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Res 46:D558–D566

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:637–643

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS et al (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125:301–313

Lindon JC, Nicholson JK (2008) Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trac-Trend Anal Chem 27:194–204

MacArthur BD, Sevilla A, Lenz M, Muller FJ, Schuldt BM et al (2012) Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. Nat Cell Biol 14:1139–1147

Manes NP, Shulzhenko N, Nuccio AG, Azeem S, Morgun A et al (2017) Multi-omics comparative analysis reveals multiple layers of host signaling pathway regulation by the gut microbiota. mSystems 2:e00107

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR et al (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534:55–62

Montague E, Janko I, Stanberry L, Lee E, Choiniere J et al (2015) Beyond protein expression, MOPED goes multi-omics. Nucleic Acids Res 43:D1145–D1151

Moussaieff A, Rouleau M, Kitsberg D, Cohen M, Levy G et al (2015) Glycolysis-mediated changes in acetyl-CoA and histone acetylation control the early differentiation of embryonic stem cells. Cell Metab 21:392–402

Orkin SH, Hochedlinger K (2011) Chromatin connections to pluripotency and cellular reprogramming. Cell 145:835–850

Rohart F, Gautier B, Singh A, Le Cao KA (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13:e1005752

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437:1173–1178

Saha S, Ewing R (2011) Systemic discovery of condition specific Wnt signalling subnetworks. In: IEEE international conference on bioinformatics and biomedicine workshops

Stumpf PS, Ewing R, MacArthur BD (2016) Single-cell pluripotency regulatory networks. Proteomics 16:2303–2312

Tuncbag N, Gosline SJ, Kedaigle A, Soltis AR, Gitter A et al (2016) Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. PLoS Comput Biol 12:e1004879

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623–627

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U et al (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. EMBO Rep 4:989–993

Vasaikar SV, Straub P, Wang J, Zhang B (2018) LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res 4:D956–D963

Wang J, Rao S, Chu J, Shen X, Levasseur DN et al (2006) A protein interaction network for pluripotency of embryonic stem cells. Nature 444:364–368

Wang Y, Medvid R, Melton C, Jaenisch R, Blelloch R (2007) DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet 39:380–385

Wang Z, Jensen MA, Zenklusen JC (2016) A practical guide to the Cancer genome atlas (TCGA). Methods Mol Biol 1418:111–141

Woo HG, Choi JH, Yoon S, Jee BA, Cho EJ et al (2017) Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. Nat Commun 8:839

Xia J, Gill EE, Hancock RE (2015) NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc 10:823–844

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL (2011) Genomewide characterization of non-polyadenylated RNAs. Genome Biol 12:R16

Yifeng L, Fang-Xiang W, Alioune N (2016) A review on machine learning principles for multi-view biological data integration. Brief Bioinform 1–16

Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B et al (2008) The ground state of embryonic stem cell self-renewal. Nature 453:519–523

Chapter 4
Studying Parasite Gene Function and Interaction Through Ribozymes and Riboswitches Design Mechanism

Harish Shukla and Timir Tripathi

Abstract Riboswitches are short mRNA sequences that can change their structural conformation to regulate the expression of adjacent genes. They regulate various biological processes via utilizing their secondary structure and consist of two distinct regions: (i) an evolutionarily conserved ligand-binding aptamer region and (ii) a variable expression platform regulating the gene expression. Many ligands bind to riboswitches, and its accurate and selective recognition requires a specific architecture that completely matches a given molecule. In general, the ligand-binding site can be found within the adjacent junction or regions; however, certain ligands may interact with the distant riboswitch regions. Gene expressions can be regulated by switching between two alternative RNA conformations; one of these conformations is favored in the presence of bound metabolite, while the other is favored in its absence. Riboswitches can regulate genes via metabolic pathways that are involved in the biosynthesis of vitamins, amino acids, and purines. In the present chapter, we aim to explain the structure, functions, and biological significance of these molecules.

Keywords Ribozymes · Riboswitches · Gene expression · Structure · Function

4.1 Introduction

Catalytic RNAs or ribozymes are present mostly as RNA–protein complexes. After its discovery in the early 1980s, ribozymes could be used to logically explain evolutionary molecular biology, i.e., whether nucleic acids or proteins came first. Ribozymes also supported the "RNA world hypothesis," which assumes that the prebiotic self-replicating molecules were composed exclusively of RNA (Scott 2007). The advent of ribozymes discards the notion that all enzymes are proteins.

H. Shukla · T. Tripathi (✉)
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India
e-mail: ttripathi@nehu.ac.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_4

In 1990, Cech and Altman shared the Nobel Prize for their demonstration that RNA could act as an enzyme. Ribozymes are small RNA structures that can catalytically cleave covalent bonds in target DNA. The most apparent example of ribozyme is believed to be the ribosomes (Nissen et al. 2000; Steitz and Moore 2003). They can inhibit gene expression in a sequence-specific manner and can have the therapeutic potential to eliminate mRNA in cancer and viral diseases.

A riboswitch is also a functional RNA: a ligand-dependent and cis-acting gene regulator in mRNA. Riboswitches are typically comprised of two domains, an aptamer domain that recognizes the ligand and an expression platform that regulates transcription termination, translation initiation, self-cleavage, or splicing of its own mRNA. Riboswitches are single- or double-stranded nucleic acids that can bind proteins or other small molecules. They are a type of noncoding RNA that up- or downregulates gene expression by switching from one structural conformation to another upon ligand binding. They bind specifically to a wide variety of molecules and macromolecules, varying from organic dyes and antibiotics to proteins and polysaccharides.

The various classes of riboswitches discovered so far are differentiated by the ligand, which on binding induces a conformational switch. Every class of riboswitch is characterized by the aptamer domain and the expression platform. The sequence and structure of the aptamer domain are highly conserved in same-class riboswitches. In contrast to ribozymes, binding is based on shape recognition and not on sequence. As therapeutics, aptamers may have a wide range of protein targets, including transcription factors, extracellular proteins, and cell surface molecules. The size of riboswitches varies from 34 nucleotides (pre-queuosine riboswitch) to up to 200 nucleotides for lysine riboswitch. Riboswitches are abundantly present and are involved in the regulation of nearly 4% of genes (Lunse et al. 2011) in eukaryotes (Cheah et al. 2007) and most likely in archaea also (Weinberg et al. 2010).

4.2 Different Tools Used to Predict Ribozymes and Riboswitch

The prediction of riboswitches can be performed by different servers using the hidden Markov model and few other methods. Thus, computational modeling for ribozymes and riboswitch characterization are the most common methods used. Computational prediction of putative riboswitches can provide direction to the biologists to study riboswitch-mediated gene expression.

4.2.1 Riboswitch Scanner

The Riboswitch Scanner (http://service.iiserkol.ac.in/%E2%88%BCriboscan/) (Mukherjee and Sengupta 2016) is a web server developed by IISER-Kolkata, India. It uses the HMM-based automated pipelines for prediction of riboswitches. The pHMM are appropriate for modeling sequence profiles and searching databases for remotely homologous sequences (Eddy 1998). They use HMMER3 (Eddy 2011) program for construction of HMM profiles of each class of riboswitches. The *hmmbuild* program of HMMER3 packages is used for the construction of an HMM model for each class. The speed of the server is very fast, and it takes an average of 30 s for one million base pairs to generate the results. The secondary structure of detected riboswitches is generated by using RNAfold program of ViennaRNA package (Lorenz et al. 2011). It is an efficient server with high sensitivity and specificity and accepts the nucleotide sequences in FASTA format. The server predicts putative riboswitches from the whole genome with their locations in the genomic sequences for 24 classes of riboswitches.

4.2.2 RibEx Server

RibEx (riboswitch explorer) (Abreu-Goodger and Merino 2005) is a web server used to determine the known sequences of riboswitches and other highly conserved bacterial regulatory elements. The server can be accessed using the following address: http://www.ibt.unam.mx/biocomputo/ribex.html. The server program is written in *Perl* programming language. The server has two modules for result prediction and predicts the result in four steps. First, it splits the given sequence and detects the riboswitch-like elements. Second, the open reading frames are done for bacterial genome. In the third step, it predicts the secondary structure and free energy of attenuator. Lastly, it gives the output using GD graphics library.

4.2.3 Riboswitch Finder

Riboswitch finder (Bengert and Dandekar 2004) is a web server (http://www.biozentrum.uni-wuerzburg.de/bioinformatik/Riboswitch/) used to examine the specific sequence, secondary structure, and folding energy of RNA. In this web server, the users can submit the sequence in batch mode, i.e., simultaneous input of multiple sequences, and each sequence must be separated by FASTA format. It uses the bona fide riboswitches as a test set (Mandal et al. 2003). In the final version, the program uses sequence, secondary structure, and folding routines to define and identify riboswitches. The riboswitch finder can be installed in local machine also.

4.2.4 RiboSW Server

The RiboSW (Chang et al. 2009) is a web server accessible through the following web address: http://ribosw.mbc.nctu.edu.tw/. The server efficiently and conveniently identifies riboswitches within the mRNA sequences. The server programming is written in C^{++} language. Based on the secondary structure and functional regions, 12 kinds of riboswitches have been classified in the server—purine, lysine, FMN, TPP, glycine, yybP-ykoY, ykkC-yxkD, SAM alpha, PreQ1, SAM, cobalamin, and glmS—through a literature survey and the Rfam database (Griffiths-Jones et al. 2005). On the basis of this information, the servers generate the models of these characteristics. These generated models are then used for searching putative riboswitches in a sequence. In other words, the server predicts the putative riboswitches from the input sequence.

4.2.5 Riboswitch Detector

The Denison riboswitch detector (DRD) (Havill et al. 2014) is a web server for predicting riboswitches from DNA sequence and can be accessed via the following URL: http://drd.denison.edu. The DRD can achieve 88–99% sensitivity and >99.99% specificity on 13 riboswitch families. The program for DRD is written in C++ language. The sequence in this server can be submitted in FASTA format. It has prior information about the riboswitches; thus, it first breaks the input sequence and then assigns that sequence to a few hundred nucleotide segments. After processing, it gives result by the PHP scripting language with much information about riboswitches.

4.2.6 Sfold Web Server

The Sfold (Ding et al. 2004) is a web server for the prediction of ribozymes and can be accessed by the following URL: http://sfold.wadsworth.org/. It is maintained by the Bioinformatics Center, Wadsworth Center, New York State Department of Health, USA. This software uses new statistical paradigm to predict the RNA-targeting nucleic acids such as antisense oligonucleotides, small interfering RNAs (siRNAs), and trans-cleaving ribozymes for gene knockdown studies. The siRNAs, antisense oligonucleotides, and ribozymes are predicted by Sirna, Soligo, and Sribo application modules. It has an easy graphical user interface server where users can submit the sequence and predict the ribozymes.

4.3 Classification

The living cell requires efficient and accurate biochemical catalysts. In the present world, most of the enzymes are proteins. However, in the primordial world, it is possible that the nucleic acids (ribozymes) carried out most of the enzymatic reactions, which support the credibility of an RNA world and suggest that nucleic acids were the original biocatalysts (Talini et al. 2009). In present times, ribozymes also play a central role in all aspects of cell life as it is involved in gene regulation, expression, and protein synthesis (Nissen et al. 2000; Teixeira et al. 2004; Winkler et al. 2004). These findings propel the belief that an RNA world is a real thing and that the ribozymes can contribute to earlier life forms for their bioprocesses. These properties of RNA were exploited by the scientists to develop ribozymes with desired activities. SELEX (systematic evolution of ligands by exponential amplification) in vitro selection is used to develop ribozymes with desired activities. Based on the occurrence, the ribozymes are classified into two categories: natural ribozymes and artificial ribozymes.

4.3.1 Natural Ribozymes

The main function of naturally occurring ribozymes is to carry out phosphoryl transfer, catalyzing the cleavage or ligation of the RNA phosphodiester backbone with the only exception of the ribosome, whose peptidyl transferase activity keeps it distinct from other ribozymes. On the basis of catalytic mechanism, the natural ribozymes are further categorized into two different groups: (1) self-cleaving ribozymes include the hammerhead (Blount and Uhlenbeck 2005), hairpin (Fedor 2000), hepatitis delta virus (HDV) (Shih and Been 2002), Varkud satellite (VS) (Lilley 2004), and glmS ribozymes (Winkler et al. 2004); and (2) self-splicing ribozymes include group I and II introns, group I-like cleavage ribozyme (GIR1 branching ribozyme) (Pyle 2005), and ribonuclease (RNase) P.

4.3.1.1 Hairpin Ribozyme

The hairpin ribozyme is a metallo-ribozyme whose catalysis involves the participation of divalent metal ions. Divalent metal ions are mainly required for the proper folding of the ribozyme rather than the direct involvement in the catalysis. This hypothesis is confirmed by replacing the magnesium ions with cobalt(III) amines, which is isosteric with magnesium ions. However, unlike the water ligands of magnesium ions, the amino ligand is not easily dissociated. Hence, the metal ion does not chemically participate in the catalysis, and its ligands cannot be easily replaced by RNA functional groups to form direct cation–RNA (Young et al. 1997; Nesbitt et al. 1997; Hampel and Cowan 1997).

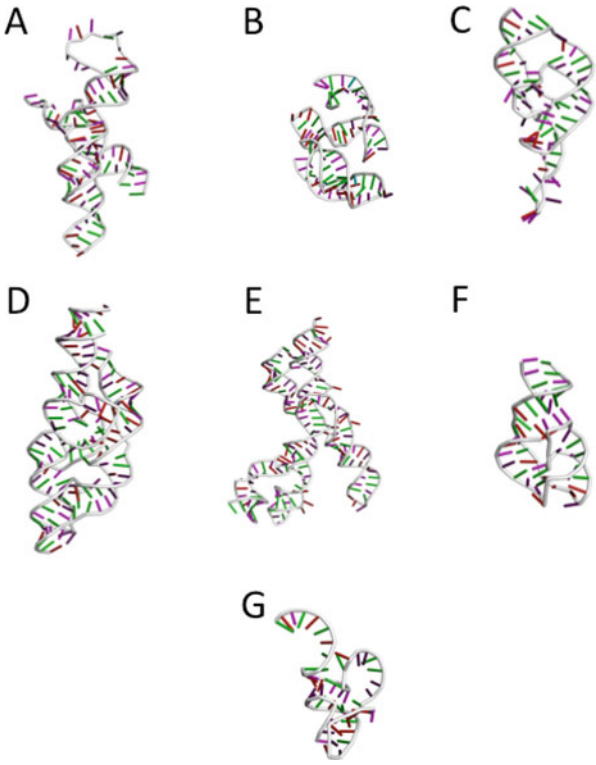

Fig. 4.1 Structure of the self-cleaving enzymes. (**a**) Hairpin ribozyme (PDB ID 1m5k), (**b**) hammerhead ribozyme (PDB ID 1hmh), (**c**) hepatitis delta virus ribozyme (PDB ID 1drj), (**d**) glmS ribozyme (PDB ID 2gcs), (**e**) Varkud satellite ribozyme (PDB ID 5v3i), (**f**) twister ribozyme (PDB ID 4oji), and (**g**) pistol ribozyme (PDB ID 5ktj)

Structurally, hairpin ribozyme contains two independently folding domains, termed as A and B, each with an internal loop flanked by two helices (H1 and H2 in domain A, H3 and H4 in domain B). The reactive phosphodiester is located within loop A (Feldstein et al. 1989; Hampel and Tritz 1989; Haseloff and Gerlach 1989) (Fig. 4.1a). The secondary structure of hairpin ribozyme is confirmed by using functional and physical studies including truncation and primer extension experiments (Feldstein et al. 1989, 1990; Haseloff and Gerlach 1989), site-directed mutagenesis (Feldstein et al. 1990; Anderson et al. 1994; Barroso-delJesus et al. 1999; Chowrira and Burke 1991; Hampel et al. 1990; Joseph et al. 1993; Perez-Ruiz et al. 1999; Sekiguchi et al. 1991), sequence comparisons of naturally occurring hairpin ribozyme variants (DeYoung et al. 1995; Haseloff and Gerlach 1988; Lian et al. 1999; Rubino et al. 1990), identification of functional variants through in vitro selection (Joseph et al. 1993; Berzal-Herranz et al. 1992; Berzal-Herranz et al. 1993; Joseph and Burke 1993; Siwkowski et al. 1998), and structure mapping through chemical protection (Angelucci et al. 2010; Butcher and Burke 1994; Hampel 1998). Mutational studies show that base-paired helices tolerate most nucleotide substitutions that maintain complementarity apart from a requirement for G11 (Feldstein et al. 1990; Anderson et al. 1994; Hampel et al. 1990; Joseph et al. 1993). Importance of loop residues is confirmed by mutational studies as the

modification in these residues compromises the catalysis, evidence that loop nucleotides participate in tertiary interactions or have direct catalytic roles (Anderson et al. 1994; Chowrira and Burke 1991; Hampel et al. 1990; Joseph et al. 1993; Perez-Ruiz et al. 1999; Berzal-Herranz et al. 1992, 1993; Joseph and Burke 1993; Siwkowski et al. 1998; Chowrira et al. 1991; Grasby et al. 1995; Ryder and Strobel 1999; Schmidt et al. 1996; Shippy et al. 1998; Siwkowski et al. 1997; Young et al. 1999; zu Putlitz et al. 1999).

4.3.1.2 Hammerhead (HHR) Ribozyme

Hammerhead ribozyme (HHR) is a member of small nucleolytic ribozymes that catalyze a self-transesterification reaction in a highly sequence-specific manner. It is the smallest known RNA catalytic motif with endoribonucleolytic activity (Buzayan et al. 1986; Uhlenbeck 1987). It has the secondary structure that has a hammerhead-like fold, and the motif adopts a "Y"-shaped fold. Here, helix III coaxially stacks with helix II, and helix I is parallel to the coaxial stack interacting with helix II through tertiary interactions required for efficient self-cleavage in vivo (Chi et al. 2008; De la Pena et al. 2003; Khvorova et al. 2003; Martick and Scott 2006) (Fig. 4.1b). The cleavage mechanism of HHRs has been highly characterized (Martick and Scott 2006; Ruffner et al. 1990). Three major kinds of HHRs have been described in the genomes of vertebrates: retroposon-like HHRs in lower vertebrates, a "discontinuous" HHR found in some mammals, and intronic HHRs conserved in amniotes. All the three groups share sequence and structural homology between the motifs of HHR (de la Pena and Garcia-Robles 2010a; Ferbeyre et al. 1998; Martick et al. 2008), suggesting either a phylogenetic or convergence relationship among them.

The first group of HHRs (tandem repeats of the satellite DNA of newts and salamanders) (Epstein and Gall 1987) and type I HHRs of the lamprey *Petromyzon marinus* and the *Xenopus tropicalis* are similar (de la Pena and Garcia-Robles 2010b; Perreault et al. 2011; Seehafer et al. 2010). These HHRs efficiently self-cleave through dimeric motifs using an elongated and more stable helix III (Forster et al. 1988; Prody et al. 1986). Type I HHRs have been widely detected among many metazoan genomes, suggesting that these repeats are related to SINE-like retrotransposons acting (Ferbeyre et al. 1998; Epstein and Gall 1987; de la Pena and Garcia-Robles 2010b).

The so-called "discontinuous" HHR (Martick et al. 2008) is an unusual type III ribozyme found in the 3′ untranslated region (UTR) of some mammalian *Clec-2* genes (Scott et al. 2009). Its specific biological role needs to be clarified, although its different functions in posttranscriptional gene regulation, like the control of mRNA decay or alternative polyadenylation sites, are defined.

The last vertebrate HHR family includes a group of highly conserved motifs occurring within the introns of specific genes of amniotes (de la Pena and Garcia-Robles 2010a). These intronic ribozymes are type I HHRs and are similar to those

found in amphibians. However, amniota HHRs show an elongated helix III with at least four Watson–Crick base pairs instead of only one.

4.3.1.3 Hepatitis Delta Virus Ribozyme (HDV Ribozyme)

The HDV ribozyme was initially discovered in the human pathogen HDV, where it was found to be closely related to genomic and antigenomic forms (Kuo et al. 1988; Sharmeen et al. 1988). HDV ribozyme is associated with HDV, which is an RNA satellite virus of hepatitis B virus (HBV) having circular, single-stranded RNA genome. This association of HDV with HBV increases the efficacy of disease than the HBV infection alone (Lai 1995). HDV ribozyme is the only known example of catalytic RNA associated with an animal virus and is active in vitro in the absence of any protein. The replication of HDV genome involves the double-rolling-circle mechanism that is carried out using the host RNA polymerase (Lai 1995; Been and Wickham 1997). The entrench ribozyme activity is necessary for post-processing of replicon generated after the replication to produce the RNA of unit length (Kuo et al. 1988; Sharmeen et al. 1988; Wu et al. 1989). Catalytically, the HDV ribozyme is the fastest known naturally occurring self-cleaving RNA that does not require a specific metal ion for action and has higher tolerance toward various denaturants (Duhamel et al. 1996; Rosenstein and Been 1990; Suh et al. 1993; Tanner et al. 1994; Thill et al. 1993) than the other known self-cleaving ribozymes. Recently, HDV-like ribozymes have been discovered throughout all kingdoms of life, including the human genome (Salehi-Ashtiani et al. 2006; Webb et al. 2009).

The crystal structure of HDV ribozyme shows the presence of a compact structure with five helical (P1 to P4 and P1.1) segments connected as a nested, double pseudoknot (Fig. 4.1c). These five helical segments form two parallel stacks: P1, P1.1, and P4 form a nearly coaxial stack, whereas P2 and P3 form a second coaxial stack (Ferre-D'Amare et al. 1998). Watson–Crick base pairs adopt the compact intricate double-pseudoknot fold to stabilize the HDV ribozyme. Mutation, which deletes the pseudoknots, reduces the ribozyme activity by three to five orders of magnitude (Tanner et al. 1994; Perrotta and Been 1996).

4.3.1.4 glmS Ribozyme

A glmS ribozyme is also an important member of self-cleaving enzyme, and interestingly, it acts both as a ribozyme as it catalyzes chemical reaction similar to other ribozymes of the same class and a riboswitch since it regulates gene expression in response to metabolite (Winkler et al. 2004; Ferre-D'Amare and Scott 2010; Ferre-D'Amare 2010). From the biochemical perspective, the most remarkable feature of the glmS ribozyme-riboswitch is its GlcN6P-induced catalytic activation. It is the first known example of a natural ribozyme that requires binding of an exogenous small molecule for the activity (Winkler et al. 2004; Barrick et al. 2004). The glmS ribozyme is a part of the 50 untranslated regions (50-UTR) of the mRNA

that encodes glucosamine-6-phosphate (GlcN6P) synthetase, and binding to GlcN6P activates its self-cleavage activity. When present at physiologic concentration, this small molecule binds to the glmS ribozyme domain and downregulates the gene expression by cleaving its own mRNA. GlcN6P is essential for efficient cleavage activity, though other primary amines with a vicinal hydroxyl group can also activate the ribozyme in vitro, including serinol and Tris (McCarthy et al. 2005). The primary amine is essential for the reaction; substitution of the amine with a hydroxyl results in complete loss of activity. Glucose-6 phosphate (Glc6P) can still bind to the ribozyme but functions as a competitive inhibitor. glmS ribozyme is widely distributed in prokaryotes (Gelfand et al. 1999; Miranda-Rios et al. 2001; Rodionov et al. 2002; Stormo and Ji 2001) and eukaryotes (Sudarsan et al. 2003). The presence of glmS ribozyme indicates that at earlier times as in RNA world, ribozymes could act as both catalysis as ribozyme and also regulate the gene expression in response to binding of small metabolites as riboswitch. During evolution, some of the ribozymes shed its riboswitch function and solely functioned as ribozyme and vice versa. glmS ribozyme adopts its catalytically active structure in the absence of GlcN6P (Hampel and Tinsley 2006; Klein and Ferre-D'Amare 2006). In contrast to all other known riboswitches (Baird and Ferre-D'Amare 2010), GlcN6P does not induce a conformational changes in the glmS riboswitch, but instead, it acts as a coenzyme and provides a catalytically essential amine group to the active site (Cochrane et al. 2007, 2009), thereby acting as an acid in the self-cleavage of glmS (Xin and Hamelberg 2010).

Most of the glmS ribozymes include a consensus sequence and are structurally divided into active core domain and supportive subdomains. These core domains are well conserved, and this strict conservation is required for the dual necessity bind GlcN6P selectively, and this selective binding positions the ligand to promote RNA transesterification efficiently.

Crystal structure of representative glmS ribozyme of *Thermoanaerobacter tengcongensis* (Klein and Ferre-D'Amare 2006) and from *Bacillus anthracis* (Cochrane et al. 2007) comprises several roughly coaxial RNA helices and three pseudoknots (Fig. 4.1d). Specifically, the P2.1 helix lies between the P4 helix and the P1, P2.2, P2, P3, and P3.1 helices. For every glmS ribozyme, the P4 helix serves as a support for the activity core and binding pocket by utilizing its GNRA tetraloop as a docking component to form a tertiary interaction with the P1 helix (Klein and Ferre-D'Amare 2006; Cochrane et al. 2007). When the P4 helix is removed, catalysis is retained (though with a substantially reduced rate constant) when high concentrations of cations are present to aid in stable structural formation (Roth et al. 2006). P2.1 and P2.2 segments possess the catalytic core (Winkler et al. 2004; Klein and Ferre-D'Amare 2006; Cochrane et al. 2007; McCown et al. 2011). The nucleotides within these regions responsible for direct binding to GlcN6P are A(-1), G1, C2, A50, U51, and G65 (Klein and Ferre-D'Amare 2006; Cochrane et al. 2007).

4.3.1.5 Varkud Satellite Ribozyme

The VS ribozyme was first discovered in the filamentous fungi *Neurospora* more than 25 years ago as part of the VS RNA, a mitochondrial RNA (Saville and Collins 1990, 1991). It is the largest member of the small self-cleaving ribozyme having seven helical segments (helices 1–7) organized in a two-prominent three-way helical junction and probably plays a key role in the folding of the ribozyme. Out of the seven helices, helix 1 acts as a substrate helix that consists of helix 1a, an internal cleavage loop, helix 1b, and the terminal loop. It can be separated from the rest of the RNA and undergoes reaction in trans with helices 2–6 (Lilley 2004; Guo and Collins 1995). Catalytic helices (helices 2–6) interact with the substrate helix through tertiary interactions that include a kissing loop interaction between the closing loops of helices 1 and 5 and interactions between the internal cleavage loop of the substrate and helices 2 and 6 (Fig. 4.1e). The mutation that disrupts these interactions reduces the rate of cleavage reaction (Bouchard and Legault 2014; Rastogi et al. 1996). Crystal structure of the VS ribozyme shows that it attains a domain-swapped dimer, where the substrate helix from one protomer docks into the catalytic cleft created by helices 2 and 6 of the other protomer (Suslov et al. 2015). The structural data suggest that the kissing loop formed by the interaction of helices 1 and 5 with substrate helix stabilizes a shift in secondary structure, facilitating the reorganization of the internal cleavage loop into its catalytically active conformation (Flinders and Dieckmann 2001; Michiels et al. 2000).

4.3.1.6 Twister Ribozyme

Twister ribozyme is one of the newest classes of small ribozymes (Roth et al. 2013; Ren et al. 2014; Liu et al. 2014; Eiler et al. 2014) present in several eukaryotes as well as bacteria. As in many other ribozymes, the general acid−base catalysis plays a significant role in the self-cleavage mechanism of twister ribozyme (Bevilacqua 2003; Das and Piccirilli 2005; Ganguly et al. 2014; Kath-Schorr et al. 2012; Nakano et al. 2000; Veeraraghavan et al. 2011; Wilson and Lilley 2010; Wilson et al. 2007; Zhang et al. 2014). Its size and structural complexity are similar to those of riboswitches (Roth et al. 2013). The metal ions, although important for folding, do not play a direct role in catalysis of the twister ribozyme (Roth et al. 2013).

The structure of the twister ribozyme is different from other known ribozyme structures (Martick and Scott 2006; Klein and Ferre-D'Amare 2006; Cochrane et al. 2007; Ke et al. 2004; Rupert and Ferre-D'Amare 2001). The structures show two pseudoknots, T1 and T2 (Fig. 4.1f), in which helices P1, T1, P2, and T2 are coaxially aligned, although there is a large helical twist angle between helices P1 and T1. Four Mg^{2+} ions are bound within the structure of the ribozyme fold. Two have complete inner coordination shells of water molecules, whereas two have exchanged some inner-sphere water ligands for phosphate non-bridging oxygen atoms at the self-cleavage site (Ren et al. 2014; Liu et al. 2014; Eiler et al. 2014). In addition, four

non-pseudoknot stems are predicted in the secondary structures (Roth et al. 2013) along with ten highly conserved nucleotides.

4.3.1.7 Pistol Ribozyme

Pistol RNAs often appear near bacteriophage-related genes and are mostly found in the *Firmicutes* phylum. There are ten highly conserved nucleotides and several modestly conserved nucleotides within the sequence of pistol ribozymes. The conserved nucleotides play key roles in nucleolytic activity, and the variable nucleotides are thought to be associated with a slight diversity in their architectures and biochemical properties. Recently, two structures of the pistol ribozyme have been determined (PDB ID: 5KTJ and 5K7C). The structures revealed similar folds with a limited secondary structure compared to other self-cleaving ribozymes (Fig. 4.1g) (Nguyen et al. 2017; Ren et al. 2016). They are characterized by three stems, P1, P2, and P3, as well as hairpin and internal loops. A six-base-pair pseudoknot helix is formed by two complementary regions located on the P1 loop and the junction connecting P2 and P3; the pseudoknot duplex is spatially situated between stems P1 and P3. Overall, the three-layered stacking of the ribozyme involves stem P1, pseudoknot stem, and a segment of stem P3 showing nontypical base pairing, and stem P2 is positioned opposite the pseudoknot stem and contributes to an overall compact fold. The main difference between the two pistol RNAs is the kind of nucleotide 32 (G or A) at the active site. The nucleotide at this position is highly conserved as a purine (Weinberg et al. 2015). The cleavage site of the pistol ribozyme resides in the G-U dinucleotide junction that links P2 and P3, which are modestly conserved, but the length of the junction is highly conserved (Ren et al. 2016).

4.3.1.8 Hatchet Ribozyme

Hatchet RNAs are self-cleaving ribozymes and are recently discovered (Lee and Lee 2017; Li et al. 2015). No 3D structure has yet been determined for hatchet ribozymes, but their sequence and secondary structure model have been predicted (Weinberg et al. 2015; Li et al. 2015). The secondary structure of hatchet ribozyme is characteristic of four major stems, P1–P4, and an additional P5 hairpin that is functionally unessential. Hatchet mutants that disrupt base pairings of stems P2 or P4 show a loss of ribozyme activity, whereas compensatory mutants that restore base pairing enable ribozyme activity (Li et al. 2015). Highly conserved nucleotides are located in a short sequence that connects P1 and P2 and in two internal bulges between P2 and P3. These nucleotides play important roles in comprising the RNA cleavage site (Li et al. 2015).

4.3.1.9 Self-Splicing Ribozymes

The group of self-splicing ribozymes includes self-splicing introns (groups I and II) (Pyle 2005). The mechanisms of self-splicing introns and the GIR1 ribozyme are quite similar as they catalyze their own splicing from the nascent mRNA and also covalently join the flanking exonic sequences (Peebles et al. 1986; Pyle 2010; Schmelzer and Schweyen 1986).

Group I intron is first identified in the large rRNA subunit of *Tetrahymena thermophila* (Cech et al. 1981), which prompted the idea that RNA catalysts continue to guide part of protein synthesis and RNA splicing in eukaryotic cells. Structurally, approximately all group I introns have the common secondary structure, the splicing pathway (Breaker et al. 2006; Strobel and Cochrane 2007), and the conserved core (Michel et al. 1982) that provides the additional support for the ancient and common origin of these molecules. Splicing by group I intron acts via two-metal ion mechanism (Stahley and Strobel 2005), suggesting that the active site of the introns is mechanistically equivalent to a large number of protein-based phosphoryl transferases, including all known DNA and RNA polymerases. These analogies between the active site of group I intron and polymerases could be an example of convergent evolution, and it is possible that this was a mechanism used by the ancient RNA polymerases that continue to be used by other RNA splicing systems.

Group II introns are large ribozymes that catalyze the RNA splicing and retrotransposition. Splicing by group II introns plays a major role in the metabolism of plants, fungi, and bacteria. Group II introns are a prevalent and structurally defined class of ribozymes that catalyze certain biochemically important reactions (Pyle 2010; Lambowitz and Belfort 2015; Toro et al. 2007). Through their reactions, these highly reactive and structurally complex RNA molecules have contributed to the evolution of almost every form of life on the earth (Lambowitz and Belfort 2015; Zimmerly and Semper 2015). Group II introns perhaps left the most noticeable inscription of the RNA world on modern genomes. It is largely believed that the eukaryotic spliceosomal introns, as well as snRNAs, evolved from this group (Michel and Ferat 1995). Moreover, the ability of group II introns to act as autonomous mobile elements suggests that they are the ancestors of modern non-LTR retrotransposons (Lambowitz and Zimmerly 2004; Robart and Zimmerly 2005). Additionally, the mechanistic and structural similarities between self-splicing group II introns and spliceosomal RNAs strongly suggest that the spliceosomal RNAs are the catalytic components of spliceosomes (Valadkhan 2007).

4.3.2 Artificial Ribozymes

The synthesis of artificial ribozymes in the laboratory depends on the dual functioning of RNAs as a catalyst and as an information carrier. This synthesis includes the

mutation of natural ribozymes that is initiated by reverse transcription using the enzyme reverse transcriptase. This results in the generation of various cDNA. This method can be used to synthesize relatively short RNA molecules, such as a 165-base-long RNA and a 189-base-long RNA that can duplicate others. These ribozymes can polymerize RNA primers. A short RNA molecule that mimics one of the two rRNAs of ribozymes has been synthesized in vitro and catalyzes the synthesis of peptide bond.

Tang and Breaker used in vitro selection of RNAs from random RNA sequence to isolate self-cleaving RNAs (Tang and Breaker 2000) that have a good enzymatic activity. Some of the synthetic ribozymes have unique structures, while others resemble the naturally occurring HHR. Researchers have engineered a tethered ribosome that works almost as well as the authentic cellular component that produces all the proteins and enzymes within the cell. It has been named as Ribo-T (Orelle et al. 2015). An RNA enzyme system has been created that can self-replicate in around 1 h (Lincoln and Joyce 2009). This system uses molecular competition or in vitro evolution to establish an RNA enzyme pair from a candidate RNA mixture. Another study has shown that an artificial riboswitch, a ligand-dependent self-cleaving ribozyme (aptazyme), can knockdown expression of an adenovirus and a measles virus structural gene, inhibiting viral genome replication and infectivity, respectively (Ketzer et al. 2014).

The techniques used to create artificial ribozymes involve Darwinian directed evolution. This approach is based on the ability of RNA to act both as a catalytic molecule and an information carrier. This dual nature of RNA provides an enormous advantage to a researcher, who could easily produce large amounts of RNA enzymes using polymerases. The ribozymes are mutated by reverse transcription into various cDNA and amplified using PCR. The selection parameters in these experiments often differ. One method for selecting a ligase ribozyme involves using biotin tags that are covalently linked to a substrate. If a molecule possesses the required ligase activity, a streptavidin matrix can be used to recover the active molecules.

4.4 Trans-acting

Riboswitches control gene expression usually through acting in cis, controlling the expression of their downstream genes through a metabolite-induced alteration of their secondary structure (Lu et al. 2008; Nudler 2006; Roth and Breaker 2009). Since the discovery of riboswitches, several cis elements have been described that trigger changes in gene expression by binding to riboswitches, i.e., nucleobases, amino acids, cofactors, and amino sugar glucosamine-6-phosphate (Nudler 2006). These types of regulatory mechanisms that are devoid of the involvement of proteins are frequently present in thiamine pyrophosphate-dependent switches, SAM-dependent switches, etc. There is a common mechanism of action of riboswitches that involves ligand binding to an aptamer domain, which subsequently generates structural changes in the 5′-UTR of an mRNA at which bacterial

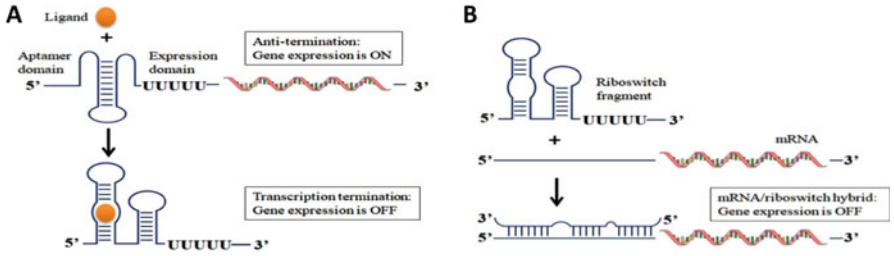

Fig. 4.2 A trans-acting riboswitch. (**a**) Via the formation of an antiterminator stem, the cis-acting riboswitches allow the transcription of full-length mRNA. Ligand binding to an aptamer domain results in structural changes in the terminator, premature termination of transcription, and inhibition of gene expression. (**b**) In *L. monocytogenes*, the terminated fragments of riboswitch bind to other mRNAs to regulate the expression in trans. (Adapted from Hartig 2010)

riboswitches are present. This ligand binding leads to the structural rearrangements of the so-called expression platform and causes changes in transcription termination or initiation of translation and, hence, ultimately regulates the output of gene expression (Fig. 4.2a).

However, trans-acting small regulatory RNAs (sRNAs) have also been described in bacteria in trans-acting (Vogel and Wagner 2007) via base pairing to mRNAs, which contain complementary sequences (Fig. 4.2b). An example of trans-acting riboswitches is observed in a pathogen *Listeria monocytogenes* (Loh et al. 2009), which results in listeriosis via consumption of infected foods. In this pathogen, the riboswitch present in the mRNA codes for cysteine, and the methionine metabolism uptake depends on SAM. SreA and SreB are two riboswitches that act by SAM-dependent termination of transcription that influences the expression of virulence factor PrfA and several other genes. These two riboswitches act in trans through an antisense mechanism. This SAM riboswitch fragment originates from SAM-dependent transcription termination of SreA that hybridizes to the target PrfA mRNA and causes the downregulation of gene expression. Here, it is interesting to note that the SAM binding does not inhibit PrfA. After the hybridization of SreA riboswitch RNA with the PrfA mRNA, the exact mechanism in repression of PrfA needs to be clarified because the ribosome binding site located downstream of the hybridization site remains unaffected by this interaction.

4.5 Inducible Gene Knockout

One of the main goals of synthetic biology is to reprogram organisms to autonomously perform complex tasks. By integrating RNA aptamers within the control regions of mRNAs, their expressions can be controlled by the addition of an appropriate ligand. These gene regulatory aptamers or riboswitches provide sophisticated means to control mRNAs without the need for protein cofactors. In nature,

bacteria use metabolite-responsive riboswitches to regulate several metabolic pathways at the RNA level (Breaker 2011; Serganov and Patel 2009). The riboswitches can control the gene expression in response to a variety of ligands using a number of different mechanisms. These mechanisms can operate either transcriptionally or posttranscriptionally, can repress or activate gene expression, can perform Boolean logic, and can respond in "digital" or "analog" fashions.

The first synthetic riboswitches were created by Werstuck and Green (Werstuck and Green 1998). They found that the insertion of RNA aptamers that bound the dye H33342 within the 50-UTR of a reporter mRNA enabled translation to be selectively blocked upon addition of the dye to living mammalian cells. Since then many synthetic riboswitches have been developed to control gene expression in response to small molecules and to build synthetic genetic circuits and logic gates (Beisel and Smolke 2009; Desai and Gallivan 2004; Dohno et al. 2013; Muranaka and Yokobayashi 2010; Sudarsan et al. 2006). Artificial riboswitch-enabled bacteria move along a concentration gradient of a small molecule (Topp and Gallivan 2007). Topp and Gallivan showed that *E. coli* can be reprogrammed to detect, follow, and accurately localize to a completely new chemical signal by using a synthetic riboswitch to recognize a ligand. These reprogrammed cells not only retain the gradient-sensing behavior of chemotactic *E. coli* but also have the unique ability to localize to a specific chemical signal.

4.6 Tools for Attenuation of Gene Expression

Gene attenuation is an important method to regulate gene regulation and is done either by structural changes in RNA or by transcriptional repressors. The regulation of gene expression by modulation of RNA structure is termed as transcription attenuation, which was discovered in the tryptophan operon of *E. coli* in 1981 (Yanofsky 1981). This attenuation was accomplished by alternative folding in the leader region upstream of the coding sequence and leads to structural RNA changes and premature termination of transcription.

Artificial riboswitches have also been engineered for the manipulation of gene expression; for example, a theophylline-sensing synthetic RNA switch causes reduced access to an adjacent SD sequence on theophylline addition (Wieland and Hartig 2008). Riboswitches as the regulator of the gene expression were discovered in 2002 (Mironov et al. 2002; Winkler et al. 2002). One of the riboswitches is a transcription attenuation mechanism mediated by binding of FMN (flavin mononucleotide) and FAD (flavin adenine dinucleotide) to the leader region of nascent riboflavin mRNA in *B. subtilis* (Mironov et al. 2002). This attenuation is accomplished by the binding of FMN or FAD to a specific box in the leader region of the mRNA (rfn box) that leads to the change in the folding pathway, preventing the formation of an antiterminator that enables the formation of a transcriptional terminator. In the absence of FMN or FAD, the RNA leader region folds into an antiterminator, resulting in transcriptional read-through (Fig. 4.3a). This mechanism

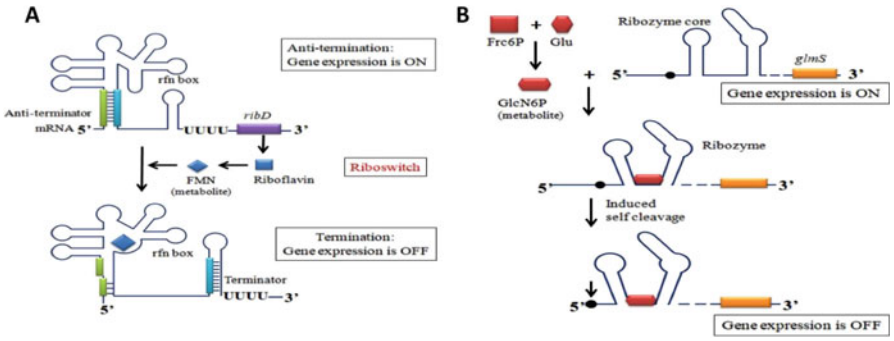

Fig. 4.3 The working mechanism of riboswitch and ribozyme. (**a**) The riboflavin riboswitch of *Bacillus subtilis*. In the absence of FMN, the structure of the leader region of the ribD mRNA facilitates the formation of a transcriptional antiterminator. The binding of FMN to the RFN (rifampin) box results in the change and disruption of the folding behavior of antiterminator and facilitates the terminator formation, leading to premature termination of transcription and abolishment of the riboflavin biosynthesis. (**b**) The glmS ribozyme switch of *B. subtilis*. The glmS enzyme aids in the conversion of precursors of glucosamine-6-phosphate (GlcN6P) into GlcN6P, which is a translated product of glmS mRNA. The GlcN6P binds to untranslated glmS leader region to activate the ribozyme causing the cleavage of the mRNA immediately upstream of the ribozyme core and leads into the translationally inactive mRNA which ultimately shut off GlmS enzyme synthesis. (Adapted from Brantl 2004)

is mainly found in low GCC gram-positive bacteria. Another mechanism is present in the gram-negative bacteria, which act by reducing translation of the *E. coli* thiamine mRNA through sequestration of the SD (Shine–Dalgarno) sequence (Brantl and Wagner 2002). This metabolite binding promotes specific folding pathways of both types of riboswitches. The above-described attenuation mechanism is found in a variety of genes encoding metabolic enzymes (Mandal and Breaker 2004; Nudler and Mironov 2004) such as the riboflavin and thiamine operons in *B. subtilis* (Mironov et al. 2002) and *E. coli* (Winkler et al. 2002), the vitamin B12 operon in *E. coli* (Nahvi et al. 2002), genes involved in methionine and lysine metabolism in *B. subtilis* (Epshtein et al. 2003; Grundy et al. 2003), and several genes involved in guanine biosynthesis in low GCC gram-positive bacteria (Mandal et al. 2003).

Transcription attenuation is also accomplished by a novel riboswitch that can act as ribozyme. This unexpected riboswitch is present in *B. subtilis*. The glutaminefructose-6-phosphate amidotransferase is encoded by glmS gene and generates glucosamine-6-phosphate (GlcN6P) by using fructose-6-phosphate (Fru6P) and glutamine (Fig. 4.3b). In contrast to classical attenuation systems, the binding of GlcN6P was found to induce a conformational change in the leader region of glmS mRNA and activate a possibly true allosteric ribozyme, which then cleaves the leader region upstream of the ribozyme element, resulting in reduced expression of the glmS gene (Winkler et al. 2004). Transcription attenuation can also be induced by artificial ribozymes that can be generated by adding RNA aptamers containing

ribozyme sequences that induce ribozyme-mediated cleavage of the RNA upon binding of small molecules (Silverman 2003).

References

Abreu-Goodger C, Merino E (2005) RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res 33:W690–W692

Anderson P, Monforte J, Tritz R, Nesbitt S, Hearst J, Hampel A (1994) Mutagenesis of the hairpin ribozyme. Nucleic Acids Res 22:1096–1100

Angelucci F, Dimastrogiovanni D, Boumis G, Brunori M, Miele AE, Saccoccia F, Bellelli A (2010) Mapping the catalytic cycle of *Schistosoma mansoni* thioredoxin glutathione reductase by X-ray crystallography. J Biol Chem 285:32557–32567

Baird NJ, Ferre-D'Amare AR (2010) Idiosyncratically tuned switching behavior of riboswitch aptamer domains revealed by comparative small-angle X-ray scattering analysis. RNA 16:598–609

Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc Natl Acad Sci U S A 101:6421–6426

Barroso-delJesus A, Tabler M, Berzal-Herranz A (1999) Comparative kinetic analysis of structural variants of the hairpin ribozyme reveals further potential to optimize its catalytic performance. Antisense Nucleic Acid Drug Dev 9:433–440

Been MD, Wickham GS (1997) Self-cleaving ribozymes of hepatitis delta virus RNA. Eur J Biochem 247:741–753

Beisel CL, Smolke CD (2009) Design principles for riboswitch function. PLoS Comput Biol 5: e1000363

Bengert P, Dandekar T (2004) Riboswitch finder – a tool for identification of riboswitch RNAs. Nucleic Acids Res 32:W154–W159

Berzal-Herranz A, Joseph S, Burke JM (1992) In vitro selection of active hairpin ribozymes by sequential RNA-catalyzed cleavage and ligation reactions. Genes Dev 6:129–134

Berzal-Herranz A, Joseph S, Chowrira BM, Butcher SE, Burke JM (1993) Essential nucleotide sequences and secondary structure elements of the hairpin ribozyme. EMBO J 12:2567–2573

Bevilacqua PC (2003) Mechanistic considerations for general acid-base catalysis by RNA: revisiting the mechanism of the hairpin ribozyme. Biochemistry 42:2259–2265

Blount KF, Uhlenbeck OC (2005) The structure-function dilemma of the hammerhead ribozyme. Annu Rev Biophys Biomol Struct 34:415–440

Bouchard P, Legault P (2014) A remarkably stable kissing-loop interaction defines substrate recognition by the Neurospora Varkud Satellite ribozyme. RNA 20:1451–1464

Brantl S (2004) Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes. Trends Microbiol 12:473–475

Brantl S, Wagner EG (2002) An antisense RNA-mediated transcriptional attenuation mechanism functions in *Escherichia coli*. J Bacteriol 184:2740–2747

Breaker RR (2011) Prospects for riboswitch discovery and analysis. Mol Cell 43:867–879

Breaker RR, Gesteland RF, Cech TR, Atkins JF (2006) The RNA world. Cold Spring Harbor Laboratory Press, New York

Butcher SE, Burke JM (1994) Structure-mapping of the hairpin ribozyme. Magnesium-dependent folding and evidence for tertiary interactions within the ribozyme-substrate complex. J Mol Biol 244:52–63

Buzayan JM, Gerlach WL, Bruening G (1986) Non-enzymatic cleavage and ligation of RNAs complementary to a plant virus satellite RNA. Nature 323:349–353

Cech TR, Zaug AJ, Grabowski PJ (1981) In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell 27:487–496

Chang TH, Huang HD, Wu LC, Yeh CT, Liu BJ, Horng JT (2009) Computational identification of riboswitches based on RNA conserved functional sequences and conformations. RNA 15:1426–1430

Cheah MT, Wachter A, Sudarsan N, Breaker RR (2007) Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. Nature 447:497–500

Chi YI, Martick M, Lares M, Kim R, Scott WG, Kim SH (2008) Capturing hammerhead ribozyme structures in action by modulating general base catalysis. PLoS Biol 6:e234

Chowrira BM, Burke JM (1991) Binding and cleavage of nucleic acids by the "hairpin" ribozyme. Biochemistry 30:8518–8522

Chowrira BM, Berzal-Herranz A, Burke JM (1991) Novel guanosine requirement for catalysis by the hairpin ribozyme. Nature 354:320–322

Cochrane JC, Lipchock SV, Strobel SA (2007) Structural investigation of the GlmS ribozyme bound to its catalytic cofactor. Chem Biol 14:97–105

Cochrane JC, Lipchock SV, Smith KD, Strobel SA (2009) Structural and chemical basis for glucosamine 6-phosphate binding and activation of the glmS ribozyme. Biochemistry 48:3239–3246

Das SR, Piccirilli JA (2005) General acid catalysis by the hepatitis delta virus ribozyme. Nat Chem Biol 1:45–52

de la Pena M, Garcia-Robles I (2010a) Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. RNA 16:1943–1950

de la Pena M, Garcia-Robles I (2010b) Intronic hammerhead ribozymes are ultraconserved in the human genome. EMBO Rep 11:711–716

De la Pena M, Gago S, Flores R (2003) Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. EMBO J 22:5561–5570

Desai SK, Gallivan JP (2004) Genetic screens and selections for small molecules based on a synthetic riboswitch that activates protein translation. J Am Chem Soc 126:13247–13254

DeYoung M, Siwkowski AM, Lian Y, Hampel A (1995) Catalytic properties of hairpin ribozymes derived from Chicory yellow mottle virus and arabis mosaic virus satellite RNAs. Biochemistry 34:15785–15791

Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res 32:W135–W141

Dohno C, Kohyama I, Kimura M, Hagihara M, Nakatani K (2013) A synthetic riboswitch that operates using a rationally designed ligand-RNA pair. Angew Chem Int Ed Engl 52:9976–9979

Duhamel J, Liu DM, Evilia C, Fleysh N, Dinter-Gottlieb G, Lu P (1996) Secondary structure content of the HDV ribozyme in 95% formamide. Nucleic Acids Res 24:3911–3917

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195

Eiler D, Wang J, Steitz TA (2014) Structural basis for the fast self-cleavage reaction catalyzed by the twister ribozyme. Proc Natl Acad Sci U S A 111:13028–13033

Epshtein V, Mironov AS, Nudler E (2003) The riboswitch-mediated control of sulfur metabolism in bacteria. Proc Natl Acad Sci U S A 100:5052–5056

Epstein LM, Gall JG (1987) Self-cleaving transcripts of satellite DNA from the newt. Cell 48:535–543

Fedor MJ (2000) Structure and function of the hairpin ribozyme. J Mol Biol 297:269–291

Feldstein PA, Buzayan JM, Bruening G (1989) Two sequences participating in the autolytic processing of satellite tobacco ringspot virus complementary RNA. Gene 82:53–61

Feldstein PA, Buzayan JM, van Tol H, Gough GR, Gilham PT, Bruening G (1990) Specific association between an endoribonucleolytic sequence from a satellite RNA and a substrate analogue containing a 2′-5′phosphodiester. Proc Natl Acad Sci 87:2623–2627

Ferbeyre G, Smith JM, Cedergren R (1998) Schistosome satellite DNA encodes active hammerhead ribozymes. Mol Cell Biol 18:3880–3888
Ferre-D'Amare AR (2010) The glmS ribozyme: use of a small molecule coenzyme by a gene-regulatory RNA. Q Rev Biophys 43:423–447
Ferre-D'Amare AR, Scott WG (2010) Small self-cleaving ribozymes. Cold Spring Harb Perspect Biol 2:a003574
Ferre-D'Amare AR, Zhou K, Doudna JA (1998) Crystal structure of a hepatitis delta virus ribozyme. Nature 395:567–574
Flinders J, Dieckmann T (2001) A pH controlled conformational switch in the cleavage site of the VS ribozyme substrate RNA. J Mol Biol 308:665–679
Forster AC, Davies C, Sheldon CC, Jeffries AC, Symons RH (1988) Self-cleaving viroid and newt RNAs may only be active as dimers. Nature 334:265–267
Ganguly A, Thaplyal P, Rosta E, Bevilacqua PC, Hammes-Schiffer S (2014) Quantum mechanical/molecular mechanical free energy simulations of the self-cleavage reaction in the hepatitis delta virus ribozyme. J Am Chem Soc 136:1483–1496
Gelfand MS, Mironov AA, Jomantas J, Kozlov YI, Perumov DA (1999) A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. Trends Genet 15:439–442
Grasby JA, Mersmann K, Singh M, Gait MJ (1995) Purine functional groups in essential residues of the hairpin ribozyme required for catalytic cleavage of RNA. Biochemistry 34:4068–4076
Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33:D121–D124
Grundy FJ, Lehman SC, Henkin TM (2003) The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. Proc Natl Acad Sci U S A 100:12057–12062
Guo HC, Collins RA (1995) Efficient trans-cleavage of a stem-loop RNA substrate by a ribozyme derived from neurospora VS RNA. EMBO J 14:368–376
Hampel A (1998) The hairpin ribozyme: discovery, two-dimensional model, and development for gene therapy. Prog Nucleic Acid Res Mol Biol 58:1–39
Hampel A, Cowan JA (1997) A unique mechanism for RNA catalysis: the role of metal cofactors in hairpin ribozyme cleavage. Chem Biol 4:513–517
Hampel KJ, Tinsley MM (2006) Evidence for preorganization of the glmS ribozyme ligand binding pocket. Biochemistry 45:7861–7871
Hampel A, Tritz R (1989) RNA catalytic properties of the minimum (-)sTRSV sequence. Biochemistry 28:4929–4933
Hampel A, Tritz R, Hicks M, Cruz P (1990) 'Hairpin' catalytic RNA model: evidence for helices and sequence requirement for substrate RNA. Nucleic Acids Res 18:299–304
Hartig JS (2010) Turning riboswitches loose. Chembiochem: Eur J Chem Biol 11:640–641
Haseloff J, Gerlach WL (1988) Simple RNA enzymes with new and highly specific endoribonuclease activities. Nature 334:585–591
Haseloff J, Gerlach WL (1989) Sequences required for self-catalysed cleavage of the satellite RNA of tobacco ringspot virus. Gene 82:43–52
Havill JT, Bhatiya C, Johnson SM, Sheets JD, Thompson JS (2014) A new approach for detecting riboswitches in DNA sequences. Bioinformatics 30:3012–3019
Joseph S, Burke JM (1993) Optimization of an anti-HIV hairpin ribozyme by in vitro selection. J Biol Chem 268:24515–24518
Joseph S, Berzal-Herranz A, Chowrira BM, Butcher SE, Burke JM (1993) Substrate selection rules for the hairpin ribozyme determined by in vitro selection, mutation, and analysis of mismatched substrates. Genes Dev 7:130–138
Kath-Schorr S, Wilson TJ, Li NS, Lu J, Piccirilli JA, Lilley DM (2012) General acid-base catalysis mediated by nucleobases in the hairpin ribozyme. J Am Chem Soc 134:16717–16724
Ke A, Zhou K, Ding F, Cate JH, Doudna JA (2004) A conformational switch controls hepatitis delta virus ribozyme catalysis. Nature 429:201–205

Ketzer P, Kaufmann JK, Engelhardt S, Bossow S, von Kalle C, Hartig JS, Ungerechts G, Nettelbeck DM (2014) Artificial riboswitches for gene expression and replication control of DNA and RNA viruses. Proc Natl Acad Sci U S A 111:E554–E562

Khvorova A, Lescoute A, Westhof E, Jayasena SD (2003) Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. Nat Struct Biol 10:708–712

Klein DJ, Ferre-D'Amare AR (2006) Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. Science 313:1752–1756

Kuo MY, Sharmeen L, Dinter-Gottlieb G, Taylor J (1988) Characterization of self-cleaving RNA sequences on the genome and antigenome of human hepatitis delta virus. J Virol 62:4439–4444

Lai MM (1995) The molecular biology of hepatitis delta virus. Annu Rev Biochem 64:259–286

Lambowitz AM, Belfort M (2015) Mobile bacterial group II introns at the crux of eukaryotic evolution. Microbiol Spectr 3:MDNA3-0050-2014

Lambowitz AM, Zimmerly S (2004) Mobile group II introns. Annu Rev Genet 38:1–35

Lee KY, Lee BJ (2017) Structural and biochemical properties of novel self-cleaving ribozymes. Molecules 22:E678

Li S, Lunse CE, Harris KA, Breaker RR (2015) Biochemical analysis of hatchet self-cleaving ribozymes. RNA 21:1845–1851

Lian Y, De Young MB, Siwkowski A, Hampel A, Rappaport J (1999) The sCYMV1 hairpin ribozyme: targeting rules and cleavage of heterologous RNA. Gene Ther 6:1114–1119

Lilley DM (2004) The Varkud satellite ribozyme. RNA 10:151–158

Lincoln TA, Joyce GF (2009) Self-sustained replication of an RNA enzyme. Science 323:1229–1232

Liu Y, Wilson TJ, McPhee SA, Lilley DM (2014) Crystal structure and mechanistic investigation of the twister ribozyme. Nat Chem Biol 10:739–744

Loh E, Dussurget O, Gripenland J, Vaitkevicius K, Tiensuu T, Mandin P, Repoila F, Buchrieser C, Cossart P, Johansson J (2009) A trans-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. Cell 139:770–779

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) Vienna RNA package 2.0. Algorithm Mol Biol 6:26

Lu C, Smith AM, Fuchs RT, Ding F, Rajashankar K, Henkin TM, Ke A (2008) Crystal structures of the SAM-III/S(MK) riboswitch reveal the SAM-dependent translation inhibition mechanism. Nat Struct Mol Biol 15:1076–1083

Lunse CE, Schmidt MS, Wittmann V, Mayer G (2011) Carba-sugars activate the glmS-riboswitch of *Staphylococcus aureus*. ACS Chem Biol 6:675–678

Mandal M, Breaker RR (2004) Gene regulation by riboswitches. Nat Rev Mol Cell Biol 5:451–463

Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR (2003) Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. Cell 113:577–586

Martick M, Scott WG (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. Cell 126:309–320

Martick M, Horan LH, Noller HF, Scott WG (2008) A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. Nature 454:899–902

McCarthy TJ, Plog MA, Floy SA, Jansen JA, Soukup JK, Soukup GA (2005) Ligand requirements for glmS ribozyme self-cleavage. Chem Biol 12:1221–1226

McCown PJ, Roth A, Breaker RR (2011) An expanded collection and refined consensus model of glmS ribozymes. RNA 17:728–736

Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64:435–461

Michel F, Jacquier A, Dujon B (1982) Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. Biochimie 64:867–881

Michiels PJ, Schouten CH, Hilbers CW, Heus HA (2000) Structure of the ribozyme substrate hairpin of Neurospora VS RNA: a close look at the cleavage site. RNA 6:1821–1832

Miranda-Rios J, Navarro M, Soberon M (2001) A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. Proc Natl Acad Sci U S A 98:9736–9741

Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell 111:747–756

Mukherjee S, Sengupta S (2016) Riboswitch scanner: an efficient pHMM-based web-server to detect riboswitches in genomic sequences. Bioinformatics 32:776–778

Muranaka N, Yokobayashi Y (2010) A synthetic riboswitch with chemical band-pass response. Chem Commun (Camb) 46:6825–6827

Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR (2002) Genetic control by a metabolite binding mRNA. Chem Biol 9:1043

Nakano S, Chadalavada DM, Bevilacqua PC (2000) General acid-base catalysis in the mechanism of a hepatitis delta virus ribozyme. Science 287:1493–1497

Nesbitt S, Hegg LA, Fedor MJ (1997) An unusual pH-independent and metal-ion-independent mechanism for hairpin ribozyme catalysis. Chem Biol 4:619–630

Nguyen LA, Wang J, Steitz TA (2017) Crystal structure of pistol, a class of self-cleaving ribozyme. Proc Natl Acad Sci U S A 114:1021–1026

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. Science 289:920–930

Nudler E (2006) Flipping riboswitches. Cell 126:19–22

Nudler E, Mironov AS (2004) The riboswitch control of bacterial metabolism. Trends Biochem Sci 29:11–17

Orelle C, Carlson ED, Szal T, Florin T, Jewett MC, Mankin AS (2015) Protein synthesis by ribosomes with tethered subunits. Nature 524:119–124

Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL (1986) A self-splicing RNA excises an intron lariat. Cell 44:213–223

Perez-Ruiz M, Barroso-DelJesus A, Berzal-Herranz A (1999) Specificity of the hairpin ribozyme. Sequence requirements surrounding the cleavage site. J Biol Chem 274:29376–29380

Perreault J, Weinberg Z, Roth A, Popescu O, Chartrand P, Ferbeyre G, Breaker RR (2011) Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. PLoS Comput Biol 7:e1002031

Perrotta AT, Been MD (1996) Core sequences and a cleavage site wobble pair required for HDV antigenomic ribozyme self-cleavage. Nucleic Acids Res 24:1314–1321

Prody GA, Bakos JT, Buzayan JM, Schneider IR, Bruening G (1986) Autolytic processing of dimeric plant virus satellite RNA. Science 231:1577–1580

Pyle AM (2005) Capping by branching: a new ribozyme makes tiny lariats. Science 309:1530–1531

Pyle AM (2010) The tertiary structure of group II introns: implications for biological function and evolution. Crit Rev Biochem Mol Biol 45:215–232

Rastogi T, Beattie TL, Olive JE, Collins RA (1996) A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. EMBO J 15:2820–2825

Ren A, Kosutic M, Rajashankar KR, Frener M, Santner T, Westhof E, Micura R, Patel DJ (2014) In-line alignment and Mg(2)(+) coordination at the cleavage site of the env22 twister ribozyme. Nat Commun 5:5534

Ren A, Vusurovic N, Gebetsberger J, Gao P, Juen M, Kreutz C, Micura R, Patel DJ (2016) Pistol ribozyme adopts a pseudoknot fold facilitating site-specific in-line cleavage. Nat Chem Biol 12:702–708

Robart AR, Zimmerly S (2005) Group II intron retroelements: function and diversity. Cytogenet Genome Res 110:589–597

Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2002) Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. J Biol Chem 277:48949–48959

Rosenstein SP, Been MD (1990) Self-cleavage of hepatitis delta virus genomic strand RNA is enhanced under partially denaturing conditions. Biochemistry 29:8011–8016
Roth A, Breaker RR (2009) The structural and functional diversity of metabolite-binding riboswitches. Annu Rev Biochem 78:305–334
Roth A, Nahvi A, Lee M, Jona I, Breaker RR (2006) Characteristics of the glmS ribozyme suggest only structural roles for divalent metal ions. RNA 12:607–619
Roth A, Weinberg Z, Chen AG, Kim PB, Ames TD, Breaker RR (2013) A widespread self-cleaving ribozyme class is revealed by bioinformatics. Nat Chem Biol 10:56–60
Rubino L, Tousignant ME, Steger G, Kaper JM (1990) Nucleotide sequence and structural analysis of two satellite RNAs associated with chicory yellow mottle virus. J Gen Virol 71 (Pt 9):1897–1903
Ruffner DE, Stormo GD, Uhlenbeck OC (1990) Sequence requirements of the hammerhead RNA self-cleavage reaction. Biochemistry 29:10695–10702
Rupert PB, Ferre-D'Amare AR (2001) Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. Nature 410:780–786
Ryder SP, Strobel SA (1999) Nucleotide analog interference mapping of the hairpin ribozyme: implications for secondary and tertiary structure formation. J Mol Biol 291:295–311
Salehi-Ashtiani K, Luptak A, Litovchick A, Szostak JW (2006) A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. Science 313:1788–1792
Saville BJ, Collins RA (1990) A site-specific self-cleavage reaction performed by a novel RNA in Neurospora mitochondria. Cell 61:685–696
Saville BJ, Collins RA (1991) RNA-mediated ligation of self-cleavage products of a Neurospora mitochondrial plasmid transcript. Proc Natl Acad Sci U S A 88:8826–8830
Schmelzer C, Schweyen RJ (1986) Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation. Cell 46:557–565
Schmidt S, Beigelman L, Karpeisky A, Usman N, Sorensen US, Gait MJ (1996) Base and sugar requirements for RNA cleavage of essential nucleoside residues in internal loop B of the hairpin ribozyme: implications for secondary structure. Nucleic Acids Res 24:573–581
Scott WG (2007) Ribozymes. Curr Opin Struct Biol 17:280–286
Scott WG, Martick M, Chi YI (2009) Structure and function of regulatory RNA elements: ribozymes that regulate gene expression. Biochim Biophys Acta 1789:634–641
Seehafer C, Kalweit A, Steger G, Graf S, Hammann C (2010) From alpaca to zebrafish: hammerhead ribozymes wherever you look. RNA 17:21–26
Sekiguchi A, Komatsu Y, Koizumi M, Ohtsuka E (1991) Mutagenesis and self-ligation of the self-cleavage domain of the satellite RNA minus strand of tobacco ringspot virus and its binding to polyamines. Nucleic Acids Res 19:6833–6838
Serganov A, Patel DJ (2009) Amino acid recognition and gene regulation by riboswitches. Biochim Biophys Acta 1789:592–611
Sharmeen L, Kuo MY, Dinter-Gottlieb G, Taylor J (1988) Antigenomic RNA of human hepatitis delta virus can undergo self-cleavage. J Virol 62:2674–2679
Shih IH, Been MD (2002) Catalytic strategies of the hepatitis delta virus ribozymes. Annu Rev Biochem 71:887–917
Shippy R, Siwkowski A, Hampel A (1998) Mutational analysis of loops 1 and 5 of the hairpin ribozyme. Biochemistry 37:564–570
Silverman SK (2003) Rube Goldberg goes (ribo)nuclear? Molecular switches and sensors made from RNA. RNA 9:377–383
Siwkowski A, Shippy R, Hampel A (1997) Analysis of hairpin ribozyme base mutations in loops 2 and 4 and their effects on cis-cleavage in vitro. Biochemistry 36:3930–3940
Siwkowski A, Humphrey M, De-Young MB, Hampel A (1998) Screening for important base identities in the hairpin ribozyme by in vitro selection for cleavage. BioTechniques 24:278–284
Stahley MR, Strobel SA (2005) Structural evidence for a two-metal-ion mechanism of group I intron splicing. Science 309:1587–1590

Steitz TA, Moore PB (2003) RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem Sci 28:411–418
Stormo GD, Ji Y (2001) Do mRNAs act as direct sensors of small molecules to control their expression? Proc Natl Acad Sci U S A 98:9465–9467
Strobel SA, Cochrane JC (2007) RNA catalysis: ribozymes, ribosomes, and riboswitches. Curr Opin Chem Biol 11:636–643
Sudarsan N, Barrick JE, Breaker RR (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA 9:644–647
Sudarsan N, Hammond MC, Block KF, Welz R, Barrick JE, Roth A, Breaker RR (2006) Tandem riboswitch architectures exhibit complex gene control functions. Science 314:300–304
Suh YA, Kumar PK, Taira K, Nishikawa S (1993) Self-cleavage activity of the genomic HDV ribozyme in the presence of various divalent metal ions. Nucleic Acids Res 21:3277–3280
Suslov NB, DasGupta S, Huang H, Fuller JR, Lilley DM, Rice PA, Piccirilli JA (2015) Crystal structure of the Varkud satellite ribozyme. Nat Chem Biol 11:840–846
Talini G, Gallori E, Maurel MC (2009) Natural and unnatural ribozymes: back to the primordial RNA world. Res Microbiol 160:457–465
Tang J, Breaker RR (2000) Structural diversity of self-cleaving ribozymes. Proc Natl Acad Sci U S A 97:5784–5789
Tanner NK, Schaff S, Thill G, Petit-Koskas E, Crain-Denoyelle AM, Westhof E (1994) A three-dimensional model of hepatitis delta virus ribozyme based on biochemical and mutational analyses. Curr Biol 4:488–498
Teixeira A, Tahiri-Alaoui A, West S, Thomas B, Ramadass A, Martianov I, Dye M, James W, Proudfoot NJ, Akoulitchev A (2004) Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. Nature 432:526–530
Thill G, Vasseur M, Tanner NK (1993) Structural and sequence elements required for the self-cleaving activity of the hepatitis delta virus ribozyme. Biochemistry 32:4254–4262
Topp S, Gallivan JP (2007) Guiding bacteria with small molecules and RNA. J Am Chem Soc 129:6807–6811
Toro N, Jimenez-Zurdo JI, Garcia-Rodriguez FM (2007) Bacterial group II introns: not just splicing. FEMS Microbiol Rev 31:342–358
Uhlenbeck OC (1987) A small catalytic oligoribonucleotide. Nature 328:596–600
Valadkhan S (2007) The spliceosome: a ribozyme at heart? Biol Chem 388:693–697
Veeraraghavan N, Ganguly A, Golden BL, Bevilacqua PC, Hammes-Schiffer S (2011) Mechanistic strategies in the HDV ribozyme: chelated and diffuse metal ion interactions and active site protonation. J Phys Chem B 115:8346–8357
Vogel J, Wagner EG (2007) Target identification of small noncoding RNAs in bacteria. Curr Opin Microbiol 10:262–270
Webb CH, Riccitelli NJ, Ruminski DJ, Luptak A (2009) Widespread occurrence of self-cleaving ribozymes. Science 326:953
Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol 11:R31
Weinberg Z, Kim PB, Chen TH, Li S, Harris KA, Lunse CE, Breaker RR (2015) New classes of self-cleaving ribozymes revealed by comparative genomics analysis. Nat Chem Biol 11:606–610
Werstuck G, Green MR (1998) Controlling gene expression in living cells through small molecule-RNA interactions. Science 282:296–298
Wieland M, Hartig JS (2008) Artificial riboswitches: synthetic mRNA-based regulators of gene expression. Chembiochem: Eur J Chem Biol 9:1873–1878
Wilson TJ, Lilley DM (2010) Do the hairpin and VS ribozymes share a common catalytic mechanism based on general acid-base catalysis? A critical assessment of available experimental data. RNA 17:213–221

Wilson TJ, McLeod AC, Lilley DM (2007) A guanine nucleobase important for catalysis by the VS ribozyme. EMBO J 26:2489–2500

Winkler W, Nahvi A, Breaker RR (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature 419:952–956

Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR (2004) Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428:281–286

Wu HN, Lin YJ, Lin FP, Makino S, Chang MF, Lai MM (1989) Human hepatitis delta virus RNA subfragments contain an autocleavage activity. Proc Natl Acad Sci U S A 86:1831–1835

Xin Y, Hamelberg D (2010) Deciphering the role of glucosamine-6-phosphate in the riboswitch action of glmS ribozyme. RNA 16:2455–2463

Yanofsky C (1981) Attenuation in the control of expression of bacterial operons. Nature 289:751–758

Young KJ, Gill F, Grasby JA (1997) Metal ions play a passive role in the hairpin ribozyme catalysed reaction. Nucleic Acids Res 25:3760–3766

Young KJ, Vyle JS, Pickering TJ, Cohen MA, Holmes SC, Merkel O, Grasby JA (1999) The role of essential pyrimidines in the hairpin ribozyme-catalysed reaction. J Mol Biol 288:853–866

Zhang S, Ganguly A, Goyal P, Bingaman JL, Bevilacqua PC, Hammes-Schiffer S (2014) Role of the active site guanine in the glmS ribozyme self-cleavage mechanism: quantum mechanical/molecular mechanical free energy simulations. J Am Chem Soc 137:784–798

Zimmerly S, Semper C (2015) Evolution of group II introns. Mob DNA 6:7

zu Putlitz J, Yu Q, Burke JM, Wands JR (1999) Combinatorial screening and intracellular antiviral activity of hairpin ribozymes directed against hepatitis B virus. J Virol 73:5381–5387

Chapter 5
Genome Microbiology for Synthetic Applications

Taj Mohammad and Md. Imtaiyaz Hassan

Abstract Synthetic biology integrates the knowledge of engineering, mathematics and physics into the biological systems to investigate and get deeper insight into the natural cellular phenomena and thus is implicated for a variety of applications. A key step in synthetic biology is logical combination of simple circuits into higher-order systems which work like a genetic switchboard which is subsequently implicated in building novel biological entities on an ever more complex level for novel application. In recent years, the field has emerged extensively and constructed many complex circuits which find its use in several fields ranging from simple laboratory experiments to clinic. Most important application of synthetic biology is the development of novel and efficient therapies for the treatment of a large number of life-threatening infectious diseases, development of vaccine, cell therapy, regenerative medicine and microbiome engineering. This chapter is aimed at providing a brief description on different applications of synthetic biology.

Keyword Synthetic biology · Genetic switchboard · Drug design and discovery, genome editing · Metabolic engineering · Cell-free biology

5.1 Introduction

Synthetic biology is an emerging field with the key aim of engineering and developing biomolecular systems with novel functionalities (Ruder et al. 2011). This potential research area has numerous applications in pharmaceutical, chemical, agricultural and energy industries. It allows a new perception, where biology, computer, chemistry and engineering have mutually revisited the older problems in biology (Weber and Fussenegger 2012). The field has designed and built increasingly complex circuits inspired by electrical engineering, and contributing new solutions to biomedical challenges such as emerging antibiotic resistance in bacteria,

T. Mohammad · M. I. Hassan (✉)
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: mihassan@jmi.ac.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_5

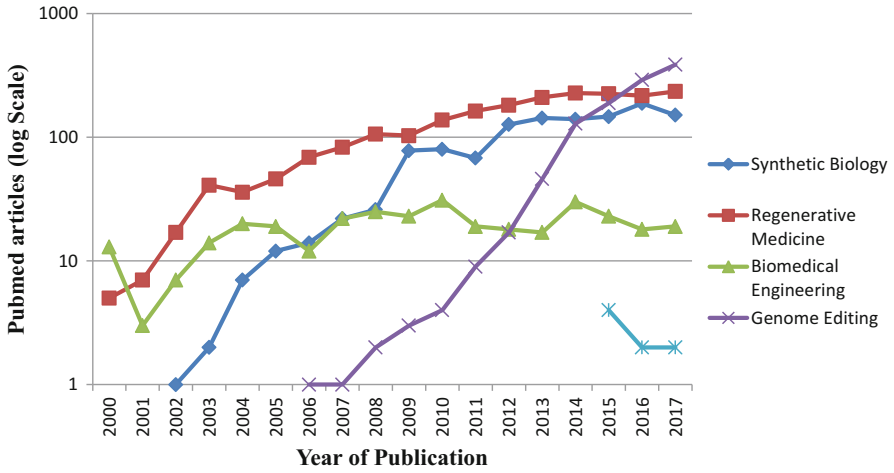

Fig. 5.1 Rise of synthetic biology and related articles in biological sciences over the period of January 2000 to December 2017. (Graph was obtained by performing a PubMed search in December 2017)

emerging infectious diseases and several other complex diseases including cancer as well as in vaccine development, microbiome engineering, cell therapy, regenerative medicine, etc. (Cheng and Lu 2012).

Synthetic biology potentially has enabled scientists to modify and engineer the cell components and even the whole living organism very precisely and efficiently to gain medical improvements in therapeutics (Arkin 2008). This field already covers a number of medical applications, for instance, diagnostics, designing and synthesis of cells and tissue with customized features, and desired biological products such as drugs, enzymes, bacteria and synthetic organs (Benner and Sismour 2005). These medical applications will probably expand rapidly over the next couple of years because of the fast and efficient easy-to-use genome editing tools now accessible and developing with newer advancements (Church et al. 2014). Most exciting example is clustered regularly interspaced short palindromic repeats (CRISPR)-associated system (Cas), used to edit genome of a living organism (Mali et al. 2013). This system has swiftly transformed preceding gene editing tools like zinc finger nucleases. It paved the way of synthetic biology toward non-specialists and has tremendous implications in many organisms being easy to use (Maeder et al. 2013).

A literature analysis of PubMed data reveals a significant rise of synthetic biology and related article in biological science over the period of January 2000 to December 2017 (Fig. 5.1). In this chapter, we discuss various approaches, applications and recent advances in the synthetic biology with special attention in the development of human therapeutics. We have categorized this chapter into different sections and discussed in detail.

5.1.1 Application of Synthetic Biology in Biomedical Engineering and Human Health

5.1.1.1 Diagnostics

Synthetic biology plays a vital role in modern clinical science. In the field of medical diagnosis, a paper-based diagnostic tool has been developed for diagnostics purpose (Martinez et al. 2009). This tool can detect pathogens in the saliva or blood which can also be engineered to detect antibiotic resistance against bacteria and infectious viruses. Engineering bacteriophages is another promising approach where output indicator such as bioluminescence is produced when a particular bacterial strain is detected (Lu et al. 2013). The mechanism of engineered bacteriophage is illustrated in Fig. 5.2. Bacteria can also be engineered for real-time biosensing which is an attractive approach of *in vivo* biosensing to monitor changes in the cellular environment of an organism (D'Souza 2001). Tailored microorganisms can be used as a biosensor for whole-cell biosensing using genetic engineering for early detection of pathogenic infections. Furthermore, genetic engineering combined with synthetic biology allows scientists to engineer bacteria and phages selectively for a particular substance. For instance, bacteria have been designed and successfully synthesized to detect arsenic in water. RNA-based biosensor is another beautiful example which is used to detect RNA sequences and metabolites specific to a particular disease (Kellenberger et al. 2013). Such diagnostic tools are needed to track public health problems for human welfare.

5.1.1.2 Treatment of Infectious Diseases

Development of synthetic constructs for the treatment of bacterial infections is currently used to control disease and to further improve the efficacy of existing antibiotics. Engineered bacteriophage is one of the best examples of biomedical

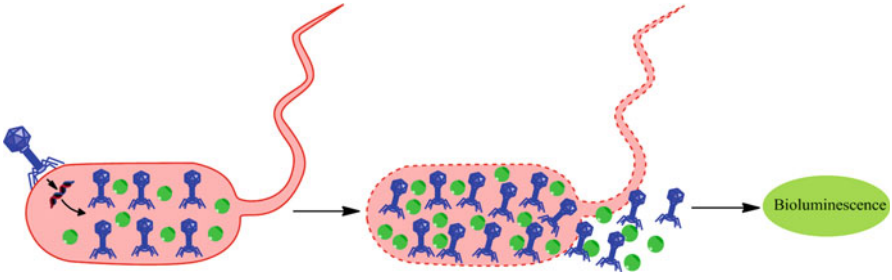

Fig. 5.2 *In vitro* phage-based diagnostics by engineered phage particles which recognize target bacteria. Once the phage has bound, the engineered phage genome is injected into the targeted cells where the reporter gene is expressed (i.e., luciferase, green circles) and phage replicates, and the resultant bioluminescence is detected

engineering which can attack resistant bacteria by disrupting its antibiotic defense mechanism (Lu et al. 2013). Enzymatic bacteriophages have been engineered to degrade bacterial biofilms to prevent the process of pathogenesis (Lu and Collins 2007). Modified lytic T7 phage can rapidly replicate during infection which can express dispersin B (DspB), an enzyme that can degrade biofilm matrix which exposes unprotected bacterial cells to the released phage, resulting in removal of all the bacteria in treated biofilms. The designed synthetic constructs has the engineered bacteriophage which can enhance the killing efficacy of existing antibiotics by disrupting bacterial system networks that regulate antibiotic defense mechanism. To overcome the issue of multidrug resistance, engineered bacteriophage are extensively exploited (Viertel et al. 2014). Bacteriophages are being engineered to overexpress LexA3, a repressor of SOS to increase the efficacy of antibiotics by inhibiting antibiotic resistance. M13 phage can be engineered to inhibit *Chlamydia trachomatis* infection, a sexually transmitted disease. Another interesting approach of developing antibacterial drugs is genome editing tools; ZFNs (Urnov et al. 2010), TALENs (Joung and Sander 2013) and CRISPR/Cas system (Hwang et al. 2013) to make a site-specific cut at DNA strand(s) and remove pathogenic bacterial strains. These tools can be used to target endogenous genes at species-specific sites to develop species-specific bactericidal agents and used to develop viral/bacterial resistance in human (Gaj et al. 2013).

5.1.1.3 Cancer Therapy

Engineered bacteria capable of invading and killing cancer cells has a promising approach for the treatment of cancer (Forbes 2010). Several bacterial strains such as *Salmonella* and *Clostridium* are able to kill cancer cells which can be further engineered to be even more potent anticancer bacterial strains. In many recent attempts, these bacterial strains have been engineered to target cancerous cells without affecting the normal cells. In such system the bacterial cells are programmed to sense molecules associated with injury or cancer and to activate the gene that stimulates repair or kills cancerous cells. Oncolytic virotherapy is another developing approach that focuses on engineering tumor-specific oncolytic viruses to kill the cancerous cells (Russell et al. 2012). In such engineered viruses, specificity can be induced for tumors at the stage of virus entry by modifying receptor binding proteins. The arming strategy may be used to engineer virus to express a protein that sensitizes both infected tumor cells and surrounding uninfected tumor cells. In the future, such programmable synthetic bacteria or virus could target and invade specific cancer-related pathways under specific *in vivo* conditions with tumor specificity (Ruder et al. 2011).

5.1.1.4 Genome Editing

Gene therapy and disease modeling are not possible without synthetic biology because it provides a wide range of genome editing tools for gene therapy and disease modeling. CRISPR/Cas9 nuclease system, TALEN and ZFN are emerging tools which have been important in genetic engineering by knocking out and knocking in the gene of interest for screening, disease modeling and drug discovery for several genetic and complex diseases (Gaj et al. 2013).

5.1.1.5 Treatment of Metabolic Disorders

Genetic enzyme deficiency or epigenetic mutations can cause various complex metabolic diseases. These disorders disrupt normal metabolism and affect several critical biochemical reactions. These days' metabolic disorders are being controlled either by dietary restriction or supplementation only; but, still no effective treatment is available. To address this problem, engineered bacterial circuits are implicated to restore normal metabolic functions (Khalil and Collins 2010). A basic representation of using programmed synthetic circuit approach for therapeutic purpose is illustrated in Fig. 5.3. Diabetes is one of the major public health problems resulting from dysfunctional insulin secretion. And for the effective treatment of this malady, engineered bacteria are being used to secrete insulin in response to the glucose concentration in the blood. Another interesting synthetic biology approach is the use of pharmaceutically controlled mammalian gene circuit to address such metabolic

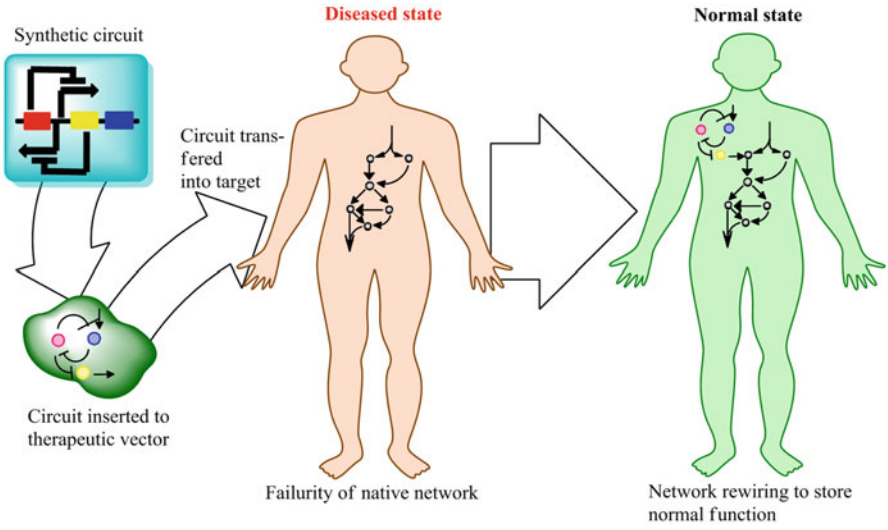

Fig. 5.3 A therapeutic approach of synthetic biology: an engineered genetic circuit is inserted to a biological network to restore normal/desired function

disorders (Ye et al. 2013). Electromagnetic gene regulation using optogenetic devices is another interesting approach. The optogenetic devices control production of therapeutic protein in diseased animals is being developed (Ye et al. 2011). However, a major limitation of implementing optogenetic devices is that the delivery of light signal can be affected by the dense animal tissues. But, instead of visible light, radio frequencies can be used for the solution which passes easily through the human tissues.

5.1.1.6 Vaccine Development

Vaccine development is one of the most challenging issues in medical science because of various rapidly evolving pathogens and diseases. Conventional vaccine development is based on live attenuated, dead or inactivated microorganisms and recombinant proteins or genome. This strategy of developing new vaccines is limited by several disadvantages such as risks in handling and use of attenuated pathogens and complexity in modifying vaccine target specificity. Synthetic biology-based development of vaccines is a new and advanced way to avoid these issues. This approach provides a safer, economic and fast synthesis of vaccine candidates. It was already shown that influenza vaccine could be finished in less than seven days using synthetic biology approach involving bioinformatics, computational biology, gene synthesis and a mammalian cell line production system. Whole-genome synthesis accomplished with computer-aided design is being used to reprogram influenza and polio viruses for safe and effective vaccination.

RNA-based vaccine is another application where synthetic RNA is transferred into the cells to produce vaccine for rapid and efficient immunity (Ulmer et al. 2012). Here, safety concern is not associated with RNA vaccines, because RNA does not mutate the genome. Although, RNA vaccines are less stable than DNA vaccines but, it can be produced at large scale in a short period of time. Development of vaccines to stimulate immunity against cancers is a beautiful example of RNA-based vaccines (Ying et al. 1999). Some of these therapeutic vaccines are being approved for several types of cancers such as prostate cancer, and many are currently in clinical trials.

Vaccines are usually delivered by injection which is a discomfort and risky procedure. Engineered plant tissue-based vaccine is a better option to address this issue which is having a low risk of contamination and does not require expensive purification and processing. There is some limitation in the development of new vaccines because of several drawbacks such as risks related with the use of attenuated pathogens along with difficulties altering vaccine target specificity. These issues can be addressed by combining synthetic circuits with recent genomic engineering advances. Synthetic circuit can be prepared to change homologous codons in the infected cells' genome automatically, and then cell-based systems would be possible to fight with the viral infection. Some examples of synthetic biology-based vaccines are RNA-, nanoparticle-, liposome- and peptide-based vaccines against microbial pathogens (McDaniel and Weiss 2005).

5.1.1.7 Cell Therapy and Regenerative Medicine

Cell therapy is a major application of biomedical engineering and synthetic biology in which the prescribed patient's own cells are introduced into the body to treat disease(s) (Caplan 2007). But, challenges remain there due to the inability to control cell behavior and phenotype after implantation. Here, uploading synthetic circuits into cells before implantation can control the systems which could be an effective solution for this problem. Unfortunately, most of the synthetic circuits are limited to microbes only and now being extended to mammalian cells which can open the door to new therapies. This approach of controlling mammalian gene expression will be very useful in various medical applications such as cell therapy and to check whether a disease phenotype is the result of changes in gene expression or not (Jaenisch and Bird 2003).

Designing engineered cells according to a patient's physiology is a new but very rapidly growing method in the field of regenerative medicine where the cell therapy involves tissues creation from the stem cells of the same patient. In the future, biomedical engineering can create some synthetic constructs in the future that can target and restore injured, diseased or aged tissue (Liu et al. 2008). One of the most important applications of synthetic biology in biomedical engineering is synthetic and hybrid (natural-synthetic) polymers which are widely used for various medical applications such as PVC in blood bags, surgical meshes, dialysis membrane, bone cement, intraocular lens, dentistry, etc. These biomedical polymers are extensively used in coatings and as nano-carriers for drug delivery system (Haag 2004).

5.1.1.8 Personalized Medicine

Design and discovery of personalized therapy is a major contribution of synthetic biology in the field of personalized medicine (Gonzalez-Angulo et al. 2010). It provides personalized sequencing of cancer cells which can be used to predict individual tumor progression and drug response. This development in synthetic biology such as personalized sequencing, biosensors, synthetic molecular tools and novel therapeutic approaches are producing a number of opportunities for future biomedical applications (Chan and Ginsburg 2011).

5.1.1.9 Designing Life Tools and Living Therapies

Synthetic biology can be used in biomedical engineering to design the living systems which can make bio-active chemical compounds such as aspirin. In modern biomedical science, synthetic biology is being used for cell-based treatments to cure several complex and rare diseases (Brenner et al. 2008). Scientists are working in this field to synthesize some unique enzymes and proteins for faster and efficient manufacturing of drugs. These engineered enzymes can be put directly into the cells,

so that each cell can perform similar to a drug-making factory. These enzymes can catalyze several biochemical pathways to formulate drugs in cells instead of reactors. This approach may be used in human cell, where a person's cells can be engineered to make drug required to cure a disease.

Another exciting finding is the generation of engineered yeast to produce cheap synthetic artemisinic acid (an antimalarial drug precursor) for the treatment of malaria (Ro et al. 2006). For this purpose, a genomic pathway from the *Artemisia annua* can be assembled and then be spliced into the genome of microbes. Hopefully, production of high-quality and economical artemisinin may be exploited to save millions of lives. In living therapeutics approach, a microbe is genetically engineered to identify a pathogen and subsequently to destroy it. As we know that bacteria can have both good and bad sides, for example *Salmonella* is being genetically altered to deliver vaccines which can work like a carrier as well as a bioreactor in the human body.

The first entirely synthetic life, termed "JCVI-syn1.0" based on existing bacterium *Mycoplasma capricolum* has been designed, assembled and synthesized (Gibson et al. 2010). This is a self-replicating single-cell bacterium whose entire genome is chemically synthesized from 1.08 million nucleotide base pairs. This new synthetic bacterium is named as *Mycoplasma mycoides* which are controlled by purely synthetic chromosome and having expected phenotypes and constant self-replication.

5.1.1.10 Tailoring Tissues and Organs

Application of synthetic biology is not only limited to gene and protein modifications but also extends to 3D printing which helps in constructing biological materials, tissues and organs which provides a newer deeper insight into medical science. Today, cell-like compartment separated by lipid bilayers are being successfully synthesized by printing tens of thousands of droplets in picoliters volume (Elani et al. 2014). 3D printing technology controls the size and shape of droplets in tailoring tissues and organs which are very similar to biological ones, and hopefully can be used in regenerative medicine, drug screening, damage repair, organ transplantation and advanced clinical applications in the future. These printed cells can construct functional tissues for medicinal use, biological research and clinical purpose. This may further be engineered to specify tissue properties. The membrane of these cells behaves like a biological membrane which can allow electrical communication in neurons.

5.1.1.11 Improving Proteins

Proteins have enzymatic and structural roles in all cells which can be engineered to perform useful operations both *in vitro* and *in vivo*. Protein engineering incorporates modifications and designing the protein to improve its activities for particular cell or

cell function (Wong and Schwaneberg 2003). Engineered proteins have remarkable potential in medical science where mutated or dysfunctional proteins can be replaced for therapeutic purpose. Synthesis of novel or uncommon amino acids such as selenocysteine and pyrrolysine for innovative medical treatment based on cell-based therapies, protein drugs and vaccines is also reflecting the advancement of protein engineering.

5.1.1.12 Immunological Disorders

Synthetic biology is very helpful in the investigation of several human disorders as it provides an outline to generate disease models and can discover new drug targets for drug development (Way et al. 2014). For example, agammaglobulinemia, an immunodeficiency disorder in which abnormal B cells are formed was investigated by redesigning the natural complexes in an orthogonal environment. Another great example is synthetic arrangement of a human peptidome on T7 phage surface which can discover autoantigens (self-antigens) for the accurate diagnosis of autoimmune diseases.

5.1.1.13 Drug Discovery and Production

Novel compounds with enhanced medicinal properties are needed for the drug development. Novel synthetic biology approaches in biomedical engineering for drug discovery and development are in demand because of their fast and accurate action (Li and Vederas 2009). For example, ethionamide, a novel antituberculosis compound has been discovered by the use of a synthetic mammalian gene circuit. Synthetic biology also provides a new way to discover novel anticancer agents that can differentiate between malignant and normal cells to target cancer. Plant-derived natural compounds are very useful and valuable in the area of medical therapeutic against various infectious diseases. However, these drugs are being produced in their native hosts which are usually expensive with negative impact on environment. Engineered microorganisms and plants are relatively better solution for large-scale production of drugs. Taxadiene, precursor to Taxol, an anticancer drug is a recent example of large-scale production by pathway optimization by dividing it into two modules in *E. coli* that was used for expression (Levskaya et al. 2005).

Synthetic gene network is an exciting approach for drug discovery (Weber and Fussenegger 2009). While synthetic gene networks are used in diagnosis, disease modeling and treatments, it can also be used for screening in drug discovery to develop novel drug compounds. Gene networks allow screening of pharmaceutically active compounds for the development of new drugs against various diseases such as cancer and tuberculosis. Synthetic biology provides several advantages over conventional methods for drug discovery. In this approach, engineered cells facilitate the redesigning of cells to screen drug molecules which may reduce the drug-

discovery time and cost. Furthermore, engineered cells can produce desirable molecules with high effectiveness and less toxicity.

5.2 Conclusion and Future Prospects

Synthetic biology is an emerging discipline covering vast areas of science, especially in biomedical engineering with an ultimate goal to improve human health. It deals with design, modifications and creation of new biological parts, devices and biological systems, as well as redesign of existing natural systems for useful therapeutic purposes. In this chapter, we have discussed a number of biomedical engineering approaches and applications of synthetic biology in diagnosis, disease modeling and medical therapeutics. We further covered current examples in drug discovery and production and progress of synthetic biology strategies. Synthetic biology tools extensively exploited to understand the mechanisms of diseases and identification and validation of potential drug targets consequently provide newer avenue to design and discover novel biopharmaceuticals. On the other hand, synthetic biology approaches are exploited to design cost-effective microbial production of complex natural products.

The aim of synthetic biology is to bring biologist and engineers together to design and construct novel biomolecular components, networks and pathways. Interestingly, these constructs may further be implicated to rewire and reprogram living organisms which may have the potential to change our lives in the near future by producing cheaper drugs, green fuel and targeted therapies against life-threatening diseases. In conclusion, synthetic biology tools are exploited to elucidate the dynamics of simple processes. Designed and biologically engineered devices may contribute to the understanding of disease mechanisms, development of newer clinical diagnostic tools and cost-effective production of therapeutic molecules for the treatment of cancer, immune diseases, metabolic disorders (diabetes and gout) and other infectious diseases.

Acknowledgments The authors sincerely acknowledge the Indian Council of Medical Research, Council for Scientific Industrial Research, University Grants Commission, and Department of Science and Technology (India) for financial support.

Conflict of Interest The authors declare no conflict of interest.

References

Arkin A (2008) Setting the standard in synthetic biology. Nat Biotechnol 26:771–774
Benner SA, Sismour AM (2005) Synthetic biology. Nat Rev Genet 6:533–543
Brenner K, You L, Arnold FH (2008) Engineering microbial consortia: a new frontier in synthetic biology. Trends Biotechnol 26:483–489

Caplan AI (2007) Adult mesenchymal stem cells for tissue engineering versus regenerative medicine. J Cell Physiol 213:341–347
Chan IS, Ginsburg GS (2011) Personalized medicine: progress and promise. Annu Rev Genomics Hum Genet 12:217–244
Cheng AA, Lu TK (2012) Synthetic biology: an emerging engineering discipline. Annu Rev Biomed Eng 14:155–178
Church GM, Elowitz MB, Smolke CD, Voigt CA, Weiss R (2014) Realizing the potential of synthetic biology. Nat Rev Mol Cell Biol 15:289–294
D'souza S (2001) Microbial biosensors. Biosens Bioelectron 16:337–353
Elani Y, Law RV, Ces O (2014) Vesicle-based artificial cells as chemical microreactors with spatially segregated reaction pathways. Nat Commun 5:5305
Forbes NS (2010) Engineering the perfect (bacterial) cancer therapy. Nat Rev Cancer 10:785–794
Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31:397–405
Gibson DG et al (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329:52–56
Gonzalez-Angulo AM, Hennessy BT, Mills GB (2010) Future of personalized medicine in oncology: a systems biology approach. J Clin Oncol 28:2777–2783
Haag R (2004) Supramolecular drug-delivery systems based on polymeric core–shell architectures. Angew Chem Int Ed 43:278–282
Hwang WY et al (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol 31:227–229
Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33:245–254
Joung JK, Sander JD (2013) TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol 14:49–55
Kellenberger CA, Wilson SC, Sales-Lee J, Hammond MC (2013) RNA-based fluorescent biosensors for live cell imaging of second messengers cyclic di-GMP and cyclic AMP-GMP. J Am Chem Soc 135:4906–4909
Khalil AS, Collins JJ (2010) Synthetic biology: applications come of age. Nat Rev Genet 11:367–379
Levskaya A et al (2005) Synthetic biology: engineering *Escherichia coli* to see light. Nature 438:441–442
Li JW-H, Vederas JC (2009) Drug discovery and natural products: end of an era or an endless frontier? Science 325:161–165
Liu Z, Jiao Y, Wang Y, Zhou C, Zhang Z (2008) Polysaccharides-based nanoparticles as drug delivery systems. Adv Drug Deliv Rev 60:1650–1662
Lu TK, Collins JJ (2007) Dispersing biofilms with engineered enzymatic bacteriophage. Proc Natl Acad Sci 104:11197–11202
Lu TK, Bowers J, Koeris MS (2013) Advancing bacteriophage-based microbial diagnostics with synthetic biology. Trends Biotechnol 31:325–327
Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK (2013) CRISPR RNA-guided activation of endogenous human genes. Nat Methods 10:977–979
Mali P et al (2013) RNA-guided human genome engineering via Cas9. Science 339:823–826
Martinez AW, Phillips ST, Whitesides GM, Carrilho E (2009) Diagnostics for the developing world: microfluidic paper-based analytical devices. ACS Publications
McDaniel R, Weiss R (2005) Advances in synthetic biology: on the path from prototypes to applications. Curr Opin Biotechnol 16:476–483
Ro D-K et al (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440:940–943
Ruder WC, Lu T, Collins JJ (2011) Synthetic biology moving into the clinic. Science 333:1248–1252
Russell SJ, Peng K-W, Bell JC (2012) Oncolytic virotherapy. Nat Biotechnol 30:658–670

Ulmer JB, Mason PW, Geall A, Mandl CW (2012) RNA-based vaccines. Vaccine 30:4414–4418
Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. Nat Rev Genet 11:636–646
Viertel TM, Ritter K, Horz H-P (2014) Viruses versus bacteria—novel approaches to phage therapy as a tool against multidrug-resistant pathogens. J Antimicrob Chemother 69:2326–2336
Way JC, Collins JJ, Keasling JD, Silver PA (2014) Integrating biological redesign: where synthetic biology came from and where it needs to go. Cell 157:151–161
Weber W, Fussenegger M (2009) The impact of synthetic biology on drug discovery. Drug Discov Today 14:956–963
Weber W, Fussenegger M (2012) Emerging biomedical applications of synthetic biology. Nat Rev Genet 13:21–35
Wong TS, Schwaneberg U (2003) Protein engineering in bioelectrocatalysis. Curr Opin Biotechnol 14:590–596
Ye H, Daoud-El Baba M, Peng R-W, Fussenegger M (2011) A synthetic optogenetic transcription device enhances blood-glucose homeostasis in mice. Science 332:1565–1568
Ye H, Charpin-El Hamri G, Zwicky K, Christen M, Folcher M, Fussenegger M (2013) Pharmaceutically controlled designer circuit for the treatment of the metabolic syndrome. Proc Natl Acad Sci 110:141–146
Ying H et al (1999) Cancer therapy using a self-replicating RNA vaccine. Nat Med 5:823–827

Chapter 6
Medicinal Application of Synthetic Biology

Umesh Panwar, Poonam Singh, and Sanjeev Kumar Singh

Abstract Today, the world's most difficult task is to improve the human health in the form of potential and functional efficiency. Therefore, synthetic biology is an emerging approach of medicinal science which is now on the edge of creating a novel life with potential and effective functions at genetic level. It represents the responsibility of engineering into the biological science to redesign the new biological system that doesn't exist in the nature. Thus, medicinal application of synthetic biology provides a fruitful hope to improve the treatment of existing diseases and enhance the human health. Herein, the presented chapter sketches the basic concept of medicinal application role into the synthetic biology towards the improvement and development of healthy life with appropriate quality, stability and efficiency.

Keywords Human health · Synthetic biology · Biomedical engineering · Genome editing · Quorum sensing · Microbiome engineering · Medicinal applications

6.1 Introduction

Synthetic biology is the basic concept of systematic engineering process along with biological science to design and build the novel biological system to understand the mechanism of biological pathways, genetic networks and genomic functions of an organism (Cachat and Davies 2011; Davies 2016). Here, we portray the practical medicinal applications of synthetic biology in biomedical engineering, genome editing, development of quorum sensing mechanism and microbiome engineering

U. Panwar · S. K. Singh (✉)
Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India

P. Singh
Corrosion & Materials Protection Division, C.S.I.R – Central Electrochemical Research Institute (CECRI), Karaikudi, Tamil Nadu, India

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_6

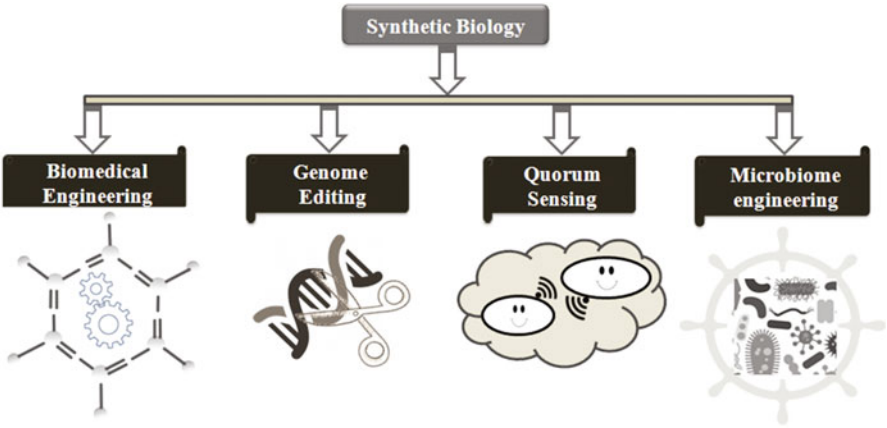

Fig. 6.1 General concept of medicinal application of synthetic biology

to make synthetic biology platforms more promising, effective and immense. The general concept of medicinal applications of synthetic biology has been provided in Fig. 6.1.

6.2 Biomedical Engineering and Human Health

Biomedical engineering is an innovative field of research in engineering and biology that applies the advanced principles of technology to improve the lifestyle of human health. It is one of the innovative field in the research and development towards the design and management of all kinds of medical resources for the treatment of public health in a better way. It is the state-of-the-art for developing the advanced treatment against human diseases and helping people worldwide lowering the expenditure of healthcare. The present scenario of research knowledge from artificial intelligence of synthetic biology helps the medical folks in making an accurate and fast response towards the biological system (Colas Fustero and Guillen Arredondo 2010; Domach 2004). A creative mind with problem-solving skills always gives a fruitful result in science, well known as a biomedical engineer. The biomedical engineers with creative skills are able to make an advanced modern technology medical devices for helping medical peoples for solving the variety of health issues. Hence, the biomedical engineering has provided the public values though the production of medical appliances saves the most of lives using the futuristic technologies like artificial hips and knee, cardiac pacemaker, artificial kidney and heart lung machines, prosthetic implants, and 3-D printing of biological organs. For the development of medical sources or devices, few more disciplines come under the biomedical engineering that have an influence on human health such as medical imaging, tissue engineering, organ implantation, and stem cell and clinical engineering (Colas Fustero and Guillen Arredondo 2010; Frisch et al. 2014; Chien et al. 2015).

6.2.1 *Use of Biomedical Engineering*

Every year the biomedical engineering is changing the way of medical research to broaden the quality of life. Today, the major importance of the biomedical engineering is very helpful to develop the healthcare system. Some of examples of biomedical engineering applications are (Colas Fustero and Guillen Arredondo 2010; Domach 2004; Frisch et al. 2014; Chien et al. 2015):

1. Developing the bioinstrumentation in diagnosis and treatment of diseases
2. Discovering the future medicine in the form of nanocapsules to examine the biological system
3. Understanding the properties and behaviour of the living organism cells at microscopic level for making an appropriate implantation
4. Analysing the medical imaging with high-speed electronic data processing
5. Helping in enhancing the quality and ability of life of an individual with physical and cognitive impairment
6. Designing and developing the artificial biomaterials for the substitution of bones, ligaments, tendons and cartilages for proper functioning of joints and muscles in the body

6.3 Genome Editing

Alkaptonuria is the first identified human genetic disease by Sir Archibald Garrod in 1901. Presently, rapidly increasing ~5000 to 8000 disease day by day is associated with mutations in every single gene in human genome. Among these, very rare diseases have the effective therapeutic treatment at advanced level, but still, our scientific community is putting tremendous effort to uncover the potential treatment against all kinds of human disease. Achieving the specific changes in genes or organism, a highly modified fundamental tool in medicinal research is required. Thus, a newborn idea in the form of genome editing has been launched to develop a method for making corrections in genes, improving the genetic modification, and providing the cure for human disease. Recently the advances of genome editing are increasing our ability to creating perfect biological models for better understanding the genetics to disease. As per current status, the utilization of genome editing is more comfortable to correct the typographical errors at genomic level. The use of genome editing is able to uncover the treatment for both monogenic and infectious diseases. However, many of the difficulties of genetic disease have to be defeated through this potential technology for safer side. Currently, scientific discovery has been revolutionized by the easiest and most efficient programmable nuclease tools of genome editing like zinc finger nucleases (ZFN), transcription activator-like effector nuclease (TALEN), cluster regularly interspaced short palindromic repeat (CRISPR)/Cas systems and RNA-guided endonuclease (RGEN), which are developed to manipulate the site-specific genome editing. The two main purposes of

handling genome skilfully (Kim 2016; Yin et al. 2017; Gori et al. 2015; Gaj et al. 2013; Zhang 2014) are:

1. Understanding the functional and regulation roles of genes or genomic region in cells
2. Getting novel insights of therapeutic agents for understanding the mutational changes in several monogenic diseases

6.3.1 Applications of Genome Editing (Gaj et al. 2013; Zhang 2014)

(a) Genome editing has the ability to manipulate the genome sequence virtually for studying the human diseases.
(b) It has significant advantage to examine the gene function for discovering new therapeutic targets for developing new therapies.
(c) Genome editing tool, TALEN, has been broadly utilized towards the gene addition, correction and disruption.
(d) TALEN fusions help to recover the quality of oil products and also increase the potatoes' long-time storage capacity.
(e) An effective genome editing tool ZFN is widely used in mutational analysis and the clinical treatment for various diseases such as sickle cell diseases, alpha-1 antitrypsin deficiency, leukaemia, X-linked severe combined immune deficiency (SCID) and HIV infection.
(f) CRISPR/Cas is one of the potent tools, which is directly useful for delivering the Cas9 endonuclease and other crRNA to human genome cells.
(g) It helps in chromosome rearrangement and endogenous gene labelling.
(h) It is widely utilized in the agriculture field to make the genetic changes for improving the plant resistance to various diseases.

6.4 Development of Quorum Sensing Mechanism

The natural behaviour of living organism needs the signalling system and support between partner's individuals. Quorum sensing is the generic regulatory mechanism of gene expression through the cell-cell communication in response to fluctuations in cell-population density. Bacteria utilize the quorum sensing to yield a secret chemical signal molecules known as autoinducers (AIs) that increase in concentration as a function of cell density to regulate the gene expression. These signal molecules can either be related to cell density or simple signals generated by bacteria at various stages of growth. But, sometimes signals are modified based on utilization in normal or different phenotypes (Jiachuan and Dacheng 2009; Rutherford and Bassler 2012). When a solo bacteria releases autoinducers (AIs) into the environment, their

concentration is too low for the detection. But, a sufficient number of bacteria are present in the environment; AI concentrations reach a threshold level that allows the bacteria to activate or deactivate targeted genes, which means that the leading bacterial growth reveals the minimal threshold stimulatory concentration of AIs which may be responsive to a gene expression.

The basic concept of QS mechanism depends on three principle features:

(a) Primary, the generation of signalling molecules (autoinducers) by community members of bacteria
(b) Secondary, the detection of variation in concentration by signalling molecules in the cell
(c) Tertiary, the detection of signalling molecules in response of activation of gene expression during AI production

The changes in gene regulation expressions depend on the diffusion mechanism process, which means the signalling molecule may diffuse at low level of cell density and may exceed at high level of cell density (Waters and Quorum Sensing 2005; Antunes et al. 2010; Basu et al. 2004).

Thus, the quorum sensing has the ability to regulate the biological process like antibiotic resistance, biofilm formation, gene transfer, virulence factor secretion, and bioluminescence in bacteria. One great example of quorum sensing is biological nitrogen fixation used in symbiotic and cell growth process. These kinds of biological process are operative based on the key behaviour of quorum sensing mechanism. Positive and negative bacteria are able to regulate the expression activities by performing the quorum sensing, called as two-component system. For example, the gram-negative bacteria use the N-acyl-homoserine lactone (AHL) as signalling molecules catalysed by Luxl. At minimal threshold level of concentration, the LuxR binds to AHL and activates the specific Lux box genes. But, the gram-positive bacteria utilized autoinducing peptide (AIP) at high concentration in the environment, which is able to activate the gene expression by phosphorylating the transcription factor with the help of receptor kinase.

An emerging application of quorum sensing circuits is in synthetic biology. Synthetic biology involves the engineering of artificial networks of proteins and/or metabolites to impart new functions to cells. The use of modularity, abstraction and standardization allows generalized principles and designs to be applied to this field. Quorum sensing is relevant to synthetic biology as the development of synthetic circuits and networks draws heavily upon basic tenets of cell-cell communication, such as the specificity of interaction between a signal and its cognate receptor. The interaction between cell-cell communication and synthetic biology is clearly two-way, as synthetic biology approaches can also help to understand the complex phenomena and observations in partner's cell interactions (Rutherford and Bassler 2012; Waters and Quorum Sensing 2005; Antunes et al. 2010; Basu et al. 2004; Costerton et al. 1987).

6.5 Microbiome Engineering

> Synthetic biology may lead to the creation of smart microbes that can detect and treat disease.
>
> By Justin L. Sonnenburg (2015)

Over the last decennary, a primary research importance of microbiome engineering is to reveal the functional role of microbiome for human health. Microbiome means the combination of collective genomes of the microorganism residing inside the host. Since the beginning of century, the microorganisms have acclimatized to make an ecological society over the globe. The microbiome engineering is a powerful platform for scientists, who focus on the microorganism's relations with human diseases to understand the biologically relevant cause at molecular level. Microbiome engineering has great promise to create and redesign the biological organism in the field of synthetic biology. These newly discovered communities of microbes are able to live in the hosts. Thus, it is easy for microorganisms to form healthy relations to the host cell. Also, it will be useful to understand the physiology of the host in relation to the disease (Kali 2015; Grice and Segre 2012; Foo et al. 2017). Like human microbiome, the animal as well as plant microbiome has developed with various communities of microbes, which is having efficiency towards the healthy growth of host. The progress of synthetic biology knowledge on microbiome communities and the newly developed microbiota-based therapeutics can be useful for understanding the host-associated microbial consortia and helpful for improving agricultural productivity and also treating the biological diseases. Microbiome engineering is a vital method, in which societies of microbiomes have ability to reproduce an interacting microbiological process (Foo et al. 2017; Mueller and Sachs 2015; Waldor et al. 2015).

Microbiome engineering modifies the microbiomes through environmental and evolutionary processes. The environmental processes contain modification in the community diversity, the structure of host–microbe and microbe–microbe relation networks and familiar species abundances. The evolutionary processes contain the disappearance of microbial types, mutation, variation in allele frequencies, and transfer of horizontal gene. Both environmental and evolutionary modifications can be analysed with high-throughput DNA sequencing process which concludes taxon presence–absence and abundance and active microbial functions that are being expressed and allows mechanistic inferences of microbiome functions. Microbiome engineering offers an opportunity to make an improved understanding of industrially important genetically manipulated and engineered prokaryotic and eukaryotic cell systems. The applications of microbiome tools with enormous efficiency can be utilized for developing a novel diagnosis and medicinal therapeutics for unlocking plenty challenges that arise in clinical area (Justin 2015; Kali 2015; Grice and Segre 2012; Foo et al. 2017; Mueller and Sachs 2015; Waldor et al. 2015; De Vrieze et al. 2017).

Note

The recent advances of medicinal application of synthetic biology in research and clinical science will be useful to protect the public health and solve complex issues globally to make a better world with better life.

Acknowledgements SKS thanks the Department of Biotechnology (DBT), New Delhi, for providing financial support. UP gratefully acknowledges Alagappa University for AURF (No. Ph.D./1122/AURF FELLOWSHIP/2015).

Conflict of Interest The author(s) declare that there is no conflict of interest.

References

Antunes LC1, Ferreira RB, Buckner MM, Finlay BB (2010) Quorum sensing in bacterial virulence. Microbiology 156(Pt 8):2271–2282. https://doi.org/10.1099/mic.0.038794-0

Basu S, Mehreja R, Thiberge S, Chen M, Weiss R (2004) Spatiotemporal control of gene expression with pulse-generating networks. Proc Natl Acad Sci USA 101:6355–6360. https://doi.org/10.1073/pnas.0307571101

Cachat E, Davies JA (2011) Application of synthetic biology to regenerative medicine. J Bioeng Biomed Sci (S2):003. https://doi.org/10.4172/2155-9538.S2-003

Chien S, Bashir R, Nerem RM, Pettigrew R (2015) Engineering as a new frontier for translational medicine. Sci Transl Med 7(281):281fs13. https://doi.org/10.1126/scitranslmed.aaa4325

Colas Fustero J, Guillen Arredondo A (2010) The biomedical engineer as a driver for Health Technology innovation. Conf Proc IEEE Eng Med Biol Soc 2010:6844–6846. https://doi.org/10.1109/IEMBS.2010.5626454

Costerton JW, Cheng K-J, Geesey GG, Ladd TI, Nickel JC et al (1987) Bacterial biofilms in nature and disease. Annu Rev Microbiol 41:435–464

Davies JA (2016) Synthetic biology: rational pathway design for regenerative medicine. Gerontology 62(5):564–570. https://doi.org/10.1159/000440721

De Vrieze J, Christiaens MER, Verstraete W (2017) The microbiome as engineering tool: manufacturing and trading between microorganisms. New Biotechnol 39(Pt B):206–214. https://doi.org/10.1016/j.nbt.2017.07.001

Domach MM (2004) What is bioengineering? In: Introduction to biomedical engineering. Pearson Prentice Hall, Upper Saddle River, pp 3–15

Foo JL, Ling H, Lee YS, Chang MW (2017) Microbiome engineering: current applications and its future. Biotechnol J 12(3). https://doi.org/10.1002/biot.201600099

Frisch PH, Stone B, Booth P, Lui W (2014) New roles & responsibilities of hospital biomedical engineering. Conf Proc IEEE Eng Med Biol Soc 2014:3488–3491. https://doi.org/10.1109/EMBC.2014.6944374

Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31(7):397–405. https://doi.org/10.1016/j.tibtech.2013.04.004

Gori JL, Hsu PD, Maeder ML, Shen S, Welstead GG, Bumcrot D (2015) Delivery and specificity of CRISPR-Cas9 genome editing technologies for human gene therapy. Hum Gene Ther 26(7):443–451. https://doi.org/10.1089/hum.2015.074

Grice EA, Segre JA (2012) The human microbiome: our second genome. Annu Rev Genomics Hum Genet 13:151–170

Jiachuan P, Dacheng R (2009) Quorum sensing inhibitors: a patent overview. Expert Opin Ther Pat 19(11):1581–1601. https://doi.org/10.1517/13543770903222293

Justin LS (2015) Microbiome engineering. Nature 518:S10. https://doi.org/10.1038/518S10a

Kali A (2015) Human microbiome engineering: the future and beyond. J Clin Diagn Res 9(9): DE01–DE04. https://doi.org/10.7860/JCDR/2015/14946.6570

Kim JS (2016) Genome editing comes of age. Nat Protoc 11(9):1573–1578. https://doi.org/10.1038/nprot.2016.104

Mueller UG, Sachs JL (2015) Engineering microbiomes to improve plant and animal health. Trends Microbiol 23:606–617

Rutherford ST, Bassler BL (2012) Bacterial quorum sensing: its role in virulence and possibilities for its control. Cold Spring Harb Perspect Med 2(11). pii: a012427. https://doi.org/10.1101/cshperspect.a012427

Waldor MK, Tyson G, Borenstein E, Ochman H (2015) Where next for microbiome research? PLoS Biol 13:e1002050

Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. Annu Rev Cell Dev Biol 21:319–346

Yin H, Kauffman KJ, Anderson DG (2017) Delivery technologies for genome editing. Nat Rev Drug Discov 16(6):387–399. https://doi.org/10.1038/nrd.2016.280

Zhang Y (2014) Genome editing with ZFN, TALEN and CRISPR/Cas systems: the applications and future prospects. Adv Genet Eng 3:e108. https://doi.org/10.4172/2169-0111.1000e108

Chapter 7
Computational Tools for Applying Multi-level Models to Synthetic Biology

Roberta Bardini, Gianfranco Politano, Alfredo Benso, and Stefano Di Carlo

Abstract Synthetic Biology is characterized by a forward engineering approach to the design of biological systems implementing desired functionalities. The Synthetic Biology design cycle benefits from the understanding and the proper representation of the underling biological complexity, allowing predicting the behavior of the target system. Considering the intrinsic nature of the systems to be designed with a Systems Biology perspective is a key requirement to support the Synthetic Biology design cycle. In particular, good models for synthetic biological systems must express hierarchy, encapsulation, selective communication, spatiality, quantitative mechanisms, and stochasticity. Computational models in general not only properly handle such modeling requirements. They can also manage heterogeneous information in compositional processes, support formal analysis and simulation, and can further be exploited for knowledge interchange among the scientific community. In particular, the nets-within-nets formalism expresses all of these features providing high flexibility in the modeling task. The formalism is well suited to represent heterogeneous systems and in general provide an extraordinary expressivity. This is achieved thanks its capability of tuning the abstraction level in each part of the model.

Keywords Computational models · Nets-within-nets · Hierarchical models · Petri Nets

7.1 Introduction

Synthetic Biology, as a field of research and application, eludes a unique definition but generally concerns the modification or the construction of a biological system with the aim of artificially implementing a desired function (Cameron et al. 2014).

R. Bardini (✉) · G. Politano · A. Benso · S. Di Carlo
Control and Computer Engineering Department, Politecnico di Torino, Torino, Italy
e-mail: roberta.bardini@polito.it; gianfranco.politano@polito.it; alfredo.benso@polito.it; stefano.dicarlo@polito.it

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_7

This usually follows a top-down engineering approach (Benso et al. 2014). In fact, regardless of the specific goals, which range from basic research to industrial applications, Synthetic Biology considers biological structures as building blocks that can be exploited to obtain, by intentional design, a desired behavior or function from a living organism.

The goal of this chapter is to highlight the role of multi-level computational models and tools in studying and designing synthetic biological systems. To introduce the reader to this challenging topic, this section starts from the analysis of a set of high-level modeling requirements that the design of a synthetic biological system may pose.

Two main strategies can be used when designing complex synthetic biological systems. In the first case, the word synthetic refers to the fact that an existing system undergoes artificial modifications. This often corresponds to the modification of the genetic information. This introduces structural/functional alterations in all steps following the transcription, until a desired function or behavior becomes available (Esvelt and Whang 2013). In the second case, synthetic biological systems can take shape by construction. This in turn means that the design effort is directed at the organization and assembly of defined biological sub-parts to compose a desired scheme. This scheme, in its overall functional activation, is able to yield the desired behavior (Purnick and Weiss 2009). Both approaches, i.e., modification and construction, pose requirements for the corresponding design processes. These two approaches often exist together when designing a synthetic biological system, and the same holds for the respective requirements.

When engineering the modification of an existing living organism, the designer requires a sufficient amount of knowledge on the portion of the biological informational stream interested by the modification. Such knowledge must go beyond the sole functional description of causal relations and interactions among different actors. It must include quantitative descriptions of the considered mechanisms.

For instance, genetic engineering tools are able to alter a genetic sequence in different ways (Qaisar et al. 2017). Examples of modifications include the introduction of a specific mutation inside a coding portion from an open reading frame. This can be used to obtain a different functional activation via structural modifications of the protein product. Another example of modification is the movement of a fast promoter in front of a sequence of interest in order to boost the transcription rate without any intervention on the coding sequence itself. In both cases, to achieve the desired result, the designer requires precise knowledge of the quantitative relation between the genetic information and its structural-functional correlates.

While this requirement holds at least for the informational stream targeted by the introduced modification, a much better scenario would be to extend it to the whole biological system. This would allow the designer to consider the effect of the introduced modifications in their whole functional context, opening a path for better predictions. A possible way to achieve this goal is to study synthetic biological systems from a Systems Biology perspective, i.e., as holistic systems functioning according to nonlinear dynamics over regulative networks of quantitative mechanisms. This is definitively different than considering a set of qualitative descriptions

of causal relations running independently of each other. In particular, the nonlinearity is a critical aspect to take into account when designing a synthetic biological system.

It is now clear that the complexity is intrinsic, when modeling and designing biological systems. This becomes even more relevant when dealing with synthetic systems whose design is specified by construction. In fact, in this case, the knowledge of the quantitative relations between the genetic information and the available structural/functional complexes introduced within and between different informational streams becomes a mandatory requirement for the design process. In fact, in this specific design approach, the designer must explicitly focus on how the intertwine among different sub-parts of the system, seen as specific informational streams or as structural-functional building blocks, can affect the behavior of the engineered organism.

This holds for many different organizational levels. For example, when designing a gene regulation network, one may consider different genetic sequences as basic building blocks. At a different level, one may use whole cells as building blocks when setting up an artificial in vitro ecological system (Goers et al. 2014). The capability of using as building blocks different elements from different system levels stresses even more the necessity of considering the Synthetic Biology design problem under a systemic perspective. However, in this case, this does not only mean to account for the inherent complexity of the system in terms of nonlinearity and network structure. It also means taking into account the hierarchical organization coming with the biological system.

To add complexity, another aspect to consider when approaching the modeling and the design of biological systems with a top-down attitude is the role of the spatiality on the system's behavior (Bardini et al. 2017a). With spatiality, we intend how the system elements are organized in the physical space and how they mutually affect each other by their relative positions.

In synthetic biological systems designed by genetic modification of existing structures, the introduction of a modification may translate into design issues relative to genetic regulation through spatial segregation. For instance, in eukaryotic cells, transcription takes place inside the nucleus, and the physical positioning of the interesting DNA portion is relevant to the relative transcription rate. RNA transcripts may or may not reach the cytoplasm and undergo translation depending on their capability to get over the regulated physical restraint within the nuclear compartment. Protein products can reach their functional targets at specific locations inside or outside the cell via regulated and active transport mechanisms, and, if they do not, this impairs the overall process.

Spatiality assumes a critical role also for synthetic biological systems designed by construction. In this case, the relative positioning of the building blocks often assumes a functional relevance for their interactions, since communication between biological entities at every organizational level is mediated by structural changes requiring physical interactions to take place. At the level of biomolecules, for example, such physical interactions involve two or more biochemical entities

stochastically moving in space and uniformly diffusing in a volume.[1] The relative positioning of two biomolecules, in the simplest case, affects the probability of their interaction, given that they are potential interactors.

At a higher level, structural and spatial features of a biological system are often involved into regulative mechanisms of specific parts of the system. For instance, tissue formation and patterning rely on spatiality-dependent induction mechanisms using diffusion profiles of regulative factors named morphogens. At a certain time, such molecules will be highly concentrated near their source and less very far from there. The morphogen effect is dose-dependent. It works differently on cells living at different distances from the signal source. In this way, during development, through the mediation of a single molecule, spatial patterns of cells assuming different fates can emerge.

To summarize, as sketched in Fig. 7.1, the design of a synthetic biological system is a top-down process, aiming at constraining a living organism within the boundaries of desired behaviors. This in turn requires to deal with complex nonlinear systems whose behavior must be described by detailed quantitative information from analyzed processes, organized into a complex hierarchical structure including spatial information. All these aspects must be holistically taken into account during the design process without neglecting the fact that, overall, stochasticity is a common ground in the mechanics of all living systems. To achieve this goal, there is a need for suitable representational tools and models. In the remaining of this chapter, we will discuss how multi-level computational models constitute a viable solution to fulfill these requirements.

7.2 Multi-level Computational Models and Representations

Multi-level computational models describe a system considering at least two different hierarchical levels. Interactions take place within and between these levels (Uhrmacher et al. 2005). In the last decade, several methods have been proposed to properly represent and simulate complex biological systems using multi-level computational models (Bardini et al. 2017a).

Among them, the most addressed methods in literature are:

- Different classes of ordinary differential equation (ODE)
- π-calculus
- Rule-based languages
- Agent-based models
- Petri Nets

In the Synthetic Biology domain, given the complexity of the design process, choosing the best modeling methodology is a non trivial task.

[1] We are not considering here more complex active movement mechanisms or other regulations.

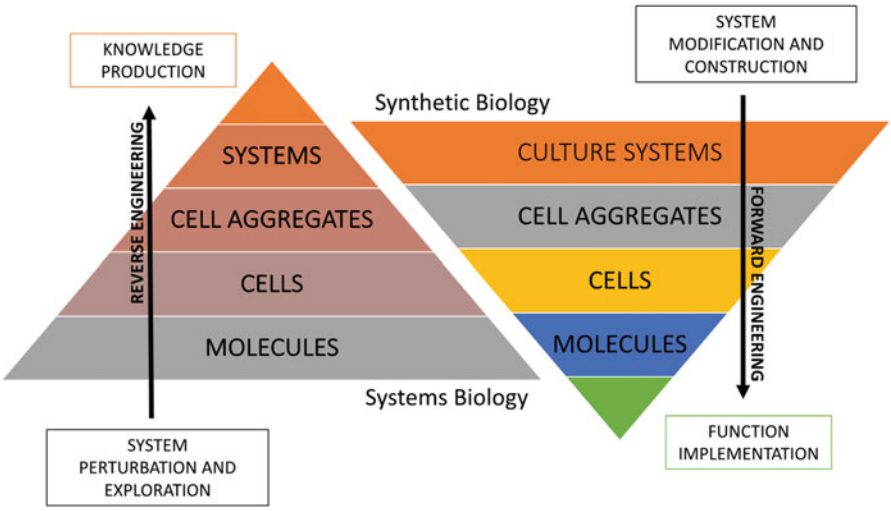

Fig. 7.1 A schematic overview of the relations between Systems and Synthetic Biology as approaches to study and design biological systems. Both disciplines take into account features intrinsic to biological systems. For example, as depicted in this figure, both disciplines deal with the hierarchical organization of the system into levels in similar ways. They consider biological complexity at a variety of temporal and spatial scales, ranging from molecules to cells and from cells to different types of cell aggregates. However, they are based on two different paradigms. On one hand, Systems Biology aims at producing information to organize into each level, at the scope of producing knowledge by means of experimental perturbation of the system. On the other hand, Synthetic Biology employs such knowledge at the different levels for the design of biological systems implementing desired functions either by modification of existing systems or by composition of existing systems or parts. The knowledge production that can arise from this procedure is a secondary goal. In other words, Systems Biology is mainly centered on a reverse engineering approach of the biological systems, while Synthetic Biology performs forward engineering actions using well-characterized biological components

Ordinary differential equations are historically one of the most common exploited approaches to create multi-level models of biological systems (Cull et al. 2005; Jones et al. 2009). They are in general used to numerically describe the continuous variations in the concentration of different substances or molecules. ODE-based models have been inherited from the biochemical domain and largely applied to gene regulatory and signaling networks. Nevertheless, although ODEs well fit stoichiometry-based chemical problems in which relations and gradients are well known, they may become over-complex and not efficient when applied to contexts with high degree of knowledge uncertainty like gene/protein interactions. Moreover, the complexity of solving complex ODE-based models strongly increases with the size of the system. This complexity becomes difficult to manage in real biological case studies with many levels of abstraction/compartmentalization and thousands of base entities (e.g., genes), all of which are concurrently interacting.

In order to reduce the computational complexity of ODEs, another multi-level modeling approach referred to as π-calculus has been proposed (Regev et al. 2001).

Although π-calculus overcomes other methods in terms of capability of modeling concurrency, communication, and stochasticity of cellular systems, it also shows several drawbacks. Its minimalism does not cope well with the description of systems that show complex dependencies and simultaneous exchange of information. High-level interaction patterns are in fact procedurally simplified into low-level ones, resulting into a set of basic modules that are difficult to write and understand (Duchier and Klutter 2006).

Another approach to create multi-level models of complex biological systems is the use of rule-based languages (Maus et al. 2011). The main advantage of these languages is their concise and compact multi-level representation of a complex system. Nonetheless, they cannot easily express downward and upward causation. In fact, an explicit notion of linkage is not provided unless they are coupled with hierarchical graphs with multiple edge types.

To overcome the limitation of the understandability of the π-calculus in multi-level modeling environments, many agent-based models have been proposed as well. They offer high degree of detail about the agent functions, which correspond to the rules governing the underling biological interactions. Moreover, they allow to manage computational parallelization, thus scaling up to very complex systems. However, one of their major drawbacks is their low accessibility to non-expert users. The majority of the agent-based models require in fact some programming skills to define a working model. Some models (e.g., Repast (North et al. 2006) and CompuCell3D (Swat et al. 2012)) offer graphical user interfaces, but these unfortunately lack significant features of relevance for synthetic biologists (Gorochowski 2016).

When facing multi-level hybrid modeling, Petri Nets appear as an overall good compromise among all the previously discussed methods (Bonzanni et al. 2014; Heiner et al. 2008). As graphical and mathematical tools, Petri Nets provide a uniform environment for modeling, formal analysis, and design of discrete-event systems. One of the major advantages of using Petri Net models is that the same model is used for the analysis of behavioral properties and performance evaluation, as well as for systematic construction of discrete-event simulators and controllers. This results in a graphical layout easily understandable also for life scientists. Moreover, it provides an unambiguous formalism that can be derived from other formal notations, such as stoichiometric matrices or ODEs. Eventually, the structure of Petri Nets is deeply based on causality, allowing to finely distinguish among concurrent and alternative behaviors (Heiner and Gilbert 2011). Several general purpose simulation tools that allow real-time inspection and network simulation using Petri Nets are available.

Among the several classes of Petri Nets presented in the literature, nets-within-nets (NWNs) are an interesting class of high-level Petri Nets. As will be better described later in this chapter, a NWN is a high-level Petri Net supporting nested architectures where complex information attached to tokens can recursively be specified with the Petri Net formalism (Valk 2004). NWNs implicitly enable to observe a system in a zoom-in/zoom-out fashion. NWNs can be used to model properties such as process synchronization, asynchronous events, concurrent

	ODE	π-Calculus	Rule-based	Agent-based	Petri Nets
1. Scalability with system's complexity	✗	✓	✗	✓	✓
2- Results Visualization/ Readability	✗	✗	✗	✓	✓
3. Step by Step Simulation	✗	✗	✓	✓	✓
4. Derivation from other formalisms	✗	✗	✓	✗	5. ✓
5. Modeling Easiness	✓	✗	✓	✗	✓
6. Management of complexity	✓	✗	✗	✓	✓

Fig. 7.2 Comparison of the main formalisms used to generate multi-level models of complex biological systems. For each formalism, the following features are considered: (1) scalability of the model with the complexity of the system, (2) capability of visualizing the model or the results of its analysis in a readable form, (3) ability to perform step-by-step simulation to inspect the behavior of the system, (4) possibility of deriving the model from other models, (5) easiness in the generation of the model from data and simulations, and (6) capability of handling and representing the complexity of the real system. Among the analyzed formalisms, Petri Net looks the most promising one to generate computation models for Synthetic Biology

operations, and conflicts or resource sharing. These properties characterize discrete-event systems and look promisingly coping with the Synthetic Biology complexity (Bardini et al. 2017a).

To wrap up, Fig. 7.2 graphically overviews the different multi-level modeling approaches presented in this section clearly highlighting how Petri Nets represent one of the most promising solutions that will be exploited in the reaming of this chapter in its NWN extension as preferred computational model for Synthetic Biology.

7.3 Nets-Within-Nets Basic Concepts and Definitions

As highlighted in the previous section, nets-within-nets are a modeling formalism belonging to the family of Petri Nets. They have peculiar characteristics for the creation of multi-level models to study and design synthetic biological systems. This section introduces the reader to the main concepts and notations required to

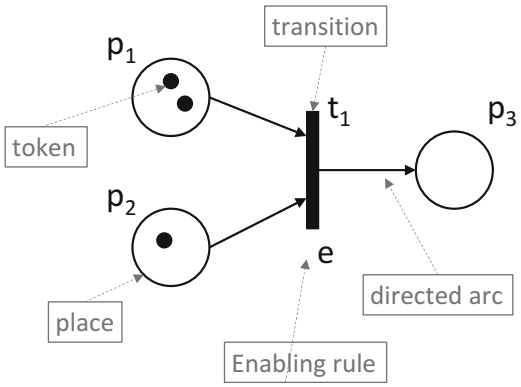

Fig. 7.3 Example of graphical representation of a Petri Net containing its basic elements. Places are depicted using circles and transitions are depicted using boxes. Places are connected to transitions and vice versa by means of directed arcs. Each transition is characterized by its enabling rule, while each place is characterized by the presence or absence of tokens depicted as black dots

understand this formalism, whose potential in Synthetic Biology will be discussed in the following section through its application on a realistic test case.

7.3.1 Petri Nets Definitions

A Petri Net (PN) is a graphical mathematical modeling tool named after Carl A. Petri who created the formalism in 1962 to study communication with automata.

As shown in Fig. 7.3, in its simplest definition, a Petri Net may be identified as a particular kind of bipartite directed graph consisting of two types of nodes: *places* represented by circles and *transitions* represented by boxes. These nodes are connected through a set of *directed arcs*. A place that has an outgoing arc toward a transition is an *input place* for the transition, whereas a place that has an incoming arc from a transition is an *output place* for the transition.

This elementary net may be used to model various aspects of a system. For instance, places may represent conditions and transitions may represent events. As another example, places may represent the availability of resources and transitions may represent their utilization.

To study the dynamic behavior of the modeled system, each state may hold a set of *tokens*, pictured as small solid dots in Fig. 7.3.

The tokens represent the basic information unit of the model that can be moved across places through the use of transitions. Tokens can represent discrete or continuous quantities of resources. At any given time, the distribution of tokens on places, called the Petri Net *marking*, defines the current state of the modeled system.

When a transition fires, the source influences the number of tokens assigned to the target.

In its basic definition, the directed arcs of a Petri Net are usually labeled with weights that represent the minimum tokens required in the input places to trigger the transition. When a transition fires, it removes a token from each input place and adds a token to each output place. Nevertheless, with the advent of modern Petri Net modeling tools (Cabac et al. 2016; Heiner et al. 2012), this stringent definition can be relaxed. It is now possible to associate every transition with a corresponding *enabling rule*. The rule can be written in a high-level programming language defining the conditions required for the transition to fire (e.g., a particular marking of the input places) and the effect of the transition (i.e., how tokens are moved when the transition fires).

Starting from the initial marking of the network, changes in the system can be simulated by executing the Petri Net, thus creating a time series of states.

The model presented so far can include more complexity to build what is usually referred to as *high-level* Petri Nets, in which each token contains complex information or data, providing much more potential for addressing real-world problems. An example of representation of this additional layer of information is given in *colored* Petri Nets where arbitrarily complex data structures represented as colors can be attached to a token. In this case, each place is described by a marking, which is a multi-set over the color set attached to the place (Jensen 2013).

Other variants of the basic Petri Nets have been proposed in the literature to extend the basic formalism adding additional expressiveness and simulation power. Interesting extensions include stochastic Petri Nets and timed Petri Nets (Bardini et al. 2017a).

7.3.2 Nets-Within-Nets Definition

Nets-within-nets (NWN) is a particular class of high-level Petri Nets that proposes to recursively represent the complex information associated with a token using the same Petri Net formalism (Valk 2004). NWNs have been already successfully exploited to model complex biological systems (Bardini et al. 2016, 2017b). As reported in Fig. 7.4, in a NWN a net contains other nets, thus generating a hierarchical organization into N different levels.

Given their definition, NWNs are particularly suited to model distributed systems which require hierarchy and encapsulation. Moreover, by construction, the NWN model perfectly fits the object-oriented design paradigm used in modern programming languages, in which each level represents a class of possible nets. This facilitates the integration of this model with several modern programming languages (Cabac et al. 2016).

When looking at the transition in a NWN, one must distinguish between *intra-layer* and *cross-layer* communications. Intra-layer communications involve elements available at a single layer, while *cross-layer* communications involve exchanges of information between two adjacent layers. Figure 7.5 proposes four

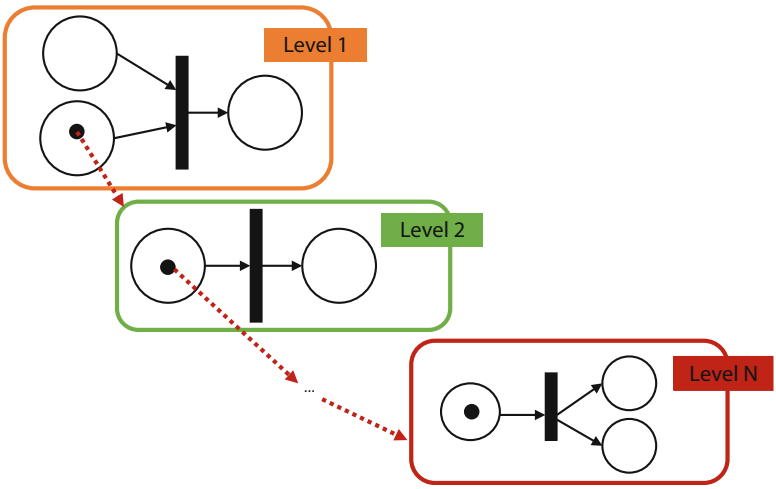

Fig. 7.4 Hierarchical organization of a NWN. Every token is recursively represented using the Petri Net formalism. This recursive modeling approach can be repeated several times, thus enabling the designer to model complex hierarchical organizations of a real system

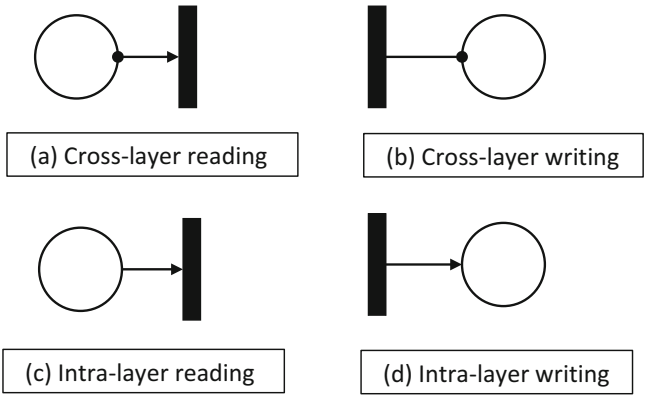

Fig. 7.5 NWN communication styles

communication basic blocks that can be used to compose complex communication structures in a NWN:

- Cross-layer reading
- Cross-layer writing
- Intra-layer reading
- Intra-layer writing

Each block is associated with a graphical symbol that will be used through the rest of this chapter to draw examples.

In a cross-layer reading, a transition at level *i* requires access to a set of information (i.e., tokens or markings) encoded in one or several nets (tokens) at level *i + 1*. Similarly, in cross-layer writing, a transition at level *i* can deliver information (tokens) to a set of lower-level nets at level *i + 1*. In a cross-layer communication, one important aspect is how to specify the target nets at the lower level. This must be specified in the definition of the transition, and, at a high-level, it can be implemented either with deterministic or stochastic enabling rules.

Intra-layer communications are the canonical Petri Net formalism in which transitions can be conditioned based on the marking of their input places and can produce results in their output places.

The four communication blocks presented in Fig. 7.5 can be freely combined together enabling the implementation of complex cross- and intra-layer communication mechanisms.

7.4 Nets-Within-Nets in Synthetic Biology

After introducing the NWN formalism, in this section we show, with a realistic example, how a NWN model can suit the requirements posed when modeling a synthetic biological system. The example has been chosen to recapitulate the different requirements posed by the definition of a synthetic system and also to be representative of what is considered as the second wave of Synthetic Biology. Synthetic Biology intends to go beyond prokaryote organisms as preferential systems in order to target the more extensive biological complexity of mammalian cells (Cameron et al. 2014).

7.4.1 Case Study Description

Among the variety of synthetic systems based on mammalian cells, we present in this section a case study based on a co-culture system including two different cell types. The choice of this system as case study implies to take into account during the modeling effort different aspects related to the experimental setup. These include the heterogeneity of different cell types and their communication via the culture medium. The culture medium, besides providing the right chemical environment to culture all the involved cell types, can also provide mechanical support for their growth and survival, possibly controlling their spatial organization and their intercellular interactions (Goers et al. 2014).

The system we want to model in this example is based on the work of Wang et al. (2008) and is graphically represented in a simplified form in Fig. 7.6. It includes two different cell types communicating through an intercellular nitric oxide (NO)-based signaling mechanisms. NO synthesis is induced in the *sender cells* and acts on the *receiver cells*, triggering the expression of an EGFP reporter. More specifically,

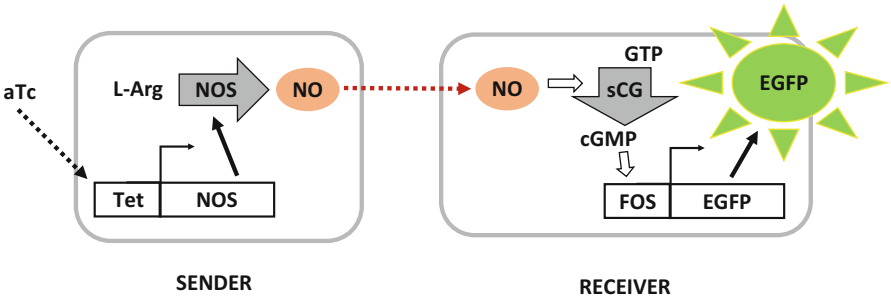

Fig. 7.6 The synthetic circuit modeled as an example of a NWN model application to Synthetic Biology, based on Wang et al. (2008). A synthetic genetic circuit is split in two complementary parts, each of them being inserted in two different cell recipients, sender and receiver cells. Sender cells produce a NO signal following the activation of an aTc-inducible transcriptional mechanism; receiver cells produce EGFP following the activation of a NO-mediated activation. NO signaling-based cell-cell communication connects these parts, enabling overall functional activation. The synthetic aspects of the system are implemented both by modification and by construction, considering the organizational level of cells. In fact, the genetic circuits, corresponding to the sender and receiver cells, are artificially introduced in cellular hosts, and this corresponds to their modification at the genetic level. The two differently modified cell types are then co-cultured, enabling the activation of the NO signaling-based cell-cell communication between them. This corresponds to a synthetic action by construction: different biological building blocks are combined to implement a desired function

sender cells are engineered to express constitutively or in TeT-on-inducible way the enzyme NO synthase (NOS) which, integrating with the c-fos promoter, catalyzes the production of NO. Receiver cells, on the other hand, are engineered to carry an EGFP-based gene reporter activated by pfos. The NO signal sent by the sender cells activates soluble guanylyl cyclase in the receiver cells, which in turn produces cytosolic GMP from GTP. Cytosolic GMP activates the reporter gene through fos, and EGFP reporter is produced. The two cell types are cultured together, and no specific geometry is imposed on their spatial organization other than the proximity between sender and receiver cells, allowing to exchange signals. This strategy relies on the use of the culture medium as a support to specify culture geometries, as well as a tunable dial to administer signals organized in space and time. Under this perspective, the culture system can be considered as an additional regulative layer for the cells, deriving from the artificial setup of the synthetic system.

The goal of this section is to show how NWNs can model the different layers involved in this design: the cell layer, the cell population layer, and the culture medium layer. Spatial organization and its corresponding effects on the cells regulative mechanisms can also be modeled at each level.

7.4.2 Model Hierarchy

The definition of the levels in the proposed case study is based on organizational aspects, comprising spatial and temporal scales of reference for the involved mechanisms and actors. In this case, it is possible to define three organizational levels of interest for the synthetic system:

- The *bottom level* models the regulatory mechanisms taking place within single cells.
- The *middle level* models populations of cells, with their spatial organization and interactions.
- The *top level* models the support medium, which organizes cells spatially and subsequently enables specific interactions between them.

A more detailed description of the NWN model representing the system of interest follows, comprising the specific use of the formalism adapted to the peculiarities of each organizational level is discussed in the next sections.

7.4.2.1 Bottom Level: The Cells

At the bottom level, the Petri Net formalism is used to implement the two regulative networks of the two cell types involved in the synthetic system. This model is based on a strong simplification of the actual regulatory network complexity since it only focuses on the synthetic circuit implementing the functions of interest. The two network classes defined at this level, one for the sender cells and the other for the receiver cells, are presented in Fig. 7.7.

Sender cells present a place for each molecular species of interest in the synthetic construct. NOS induction is modeled by specifying a place for the NOS gene and including the tet-inducible promoter in the transition regulating its transcription as an enabling rule. The transition is enabled as long as tokens are present in the place modeling the aTc signal. This transition produces tokens in the place modeling the NOS gene product. The marking in here determines the enabling of the transition modeling the catalytic action by NOS which, when active, consumes tokens in the L-Arg place for producing tokens in the NO place.

Receiver cells have a place for inactive sCG, and tokens are moved from there to the place modeling the active form of the enzyme by a transition requiring the presence of tokens in the NO place. The presence of tokens in the sCG place in turn enables the transition that moves tokens from the GTP place to the cGMP place. The presence of tokens in the cGMP place enables, via a rule for the transition modeling the functioning of the cGMP-dependant promoter, the transition that models the transcription of the EGFP gene, which produces tokens in the place for EGFP gene products.

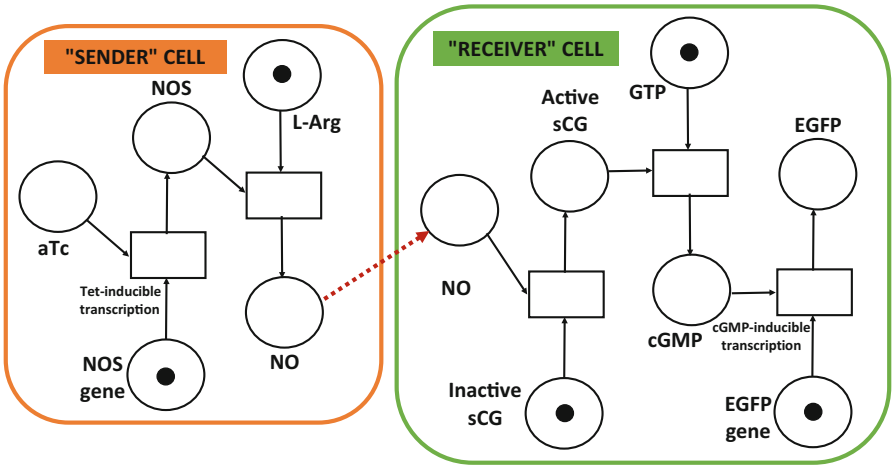

Fig. 7.7 The NWN model at the bottom level represents the cell types involved in the system. They are described using Petri Nets modeling the relevant synthetic functionalities that they implement. Sender cells model the aTc-dependant transcriptional activation of NO synthase with the enabling rule of the transition modeling the production of the NOS gene product. The marking of the NOS operates the transition modeling NO production from L-Arg. Receiver cells depend on the marking of the place for incoming NO molecules. These activate the sCG enzyme, which if present in its active form corresponds to a marking enabling the transition transforming GTP into cGMP. This activates through a dedicated enabling rule the transition for the transcription of the EGFP gene

7.4.2.2 Middle Level: The Cell Populations

This level, represented in Fig. 7.8, models a spatial grid where places function as positional slots for cells or molecular species, specifying their absolute position with respect to the overall grid and their spatial relations with the neighbor cells.

Each place can carry different types of tokens, each of them being an instance of a net class. At this level, tokens are either net instances of the bottom-level networks (sender or receiver cells) or net instances modeling molecular species present in the extracellular proximal environment of the cells (like the NO molecule or the aTc molecule).

Transitions at this level implement different mechanisms. Both a cross-layer reading and a cross-layer writing mechanisms are implemented to model the cell-cell communication system between sender and receiver cells. The cross-layer reading is used to:

- Consume tokens from the NO place of a sender cell net to produce tokens being instances of a simple net class corresponding to the NO molecule in the proximities of the sender cells.
- Consume tokens from the EGFP place of a receiver cell to produce tokens being instances of a simple net class corresponding to the EGFP fluorescent signal.

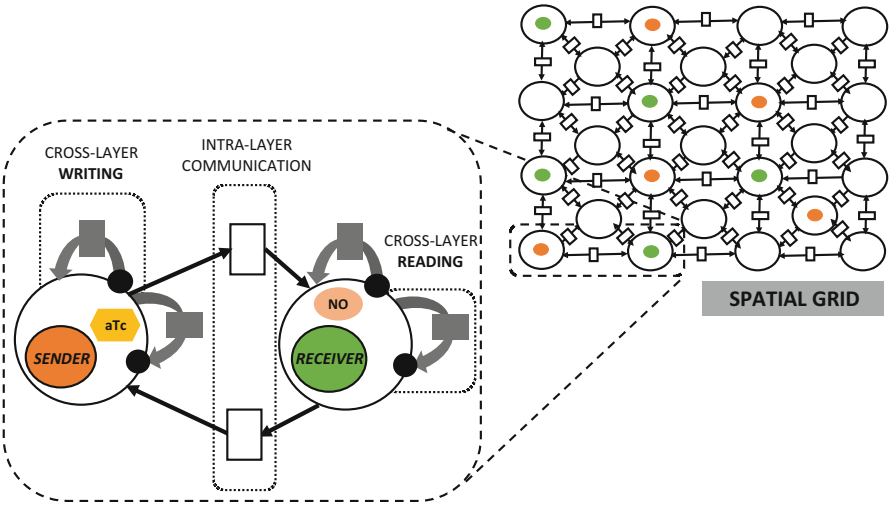

Fig. 7.8 The NWN model at the middle level represents the population of cells from the co-culture. Instances of both the sender and the receiver net classes can exist in the places at the middle level. Intra-level transitions able to move simple tokens modeling diffusible molecules, such as the NO and aTc signals, link each place with its neighbors. This provides an adjacency-based representation of the spatial grid the cells live into, making it possible to model spatiality-dependent processes, such as the diffusion of signals among different places. Cross-layer communication mechanisms are available for each place from the grid. Cross-layer writing can impose a marking on the net instance at the lower level, based on the signals (molecules) available in the same place as the cell net. Cross-layer reading can generate a signal (molecule) based on the marking of the cell net from the lower level existing in that same place

In both cases, the NO and EGFP molecules share on the grid the same location with the cell that produced them.

On the other side, the cross-layer writing channel is used to model the following mechanisms:

- When an NO molecule token appears in the same place of a receiver cell, to write the marking to the NO place of the receiver cell net instance.
- When an aTc inducer signal token is in the same place of a sender cell, to write a marking into the aTc place of the sender cell net instance.

A third category of transitions at the middle level models the intra-level movement of resources among places. For instance, the diffusion process of NO or aTc inducer signal. At this level, instances of NO molecule tokens or aTc inducer signal tokens simply model the presence of those molecules on the grid. Such tokens can randomly move from a place to one of its neighbor places through intra-level transitions.

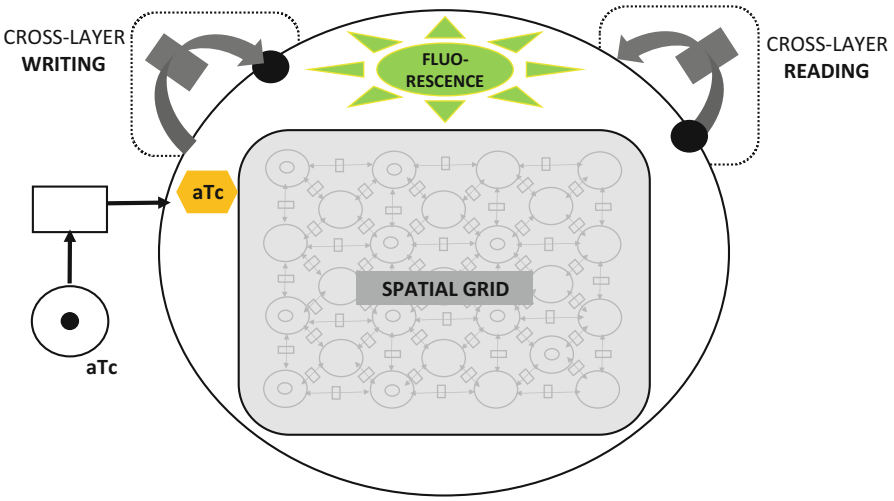

Fig. 7.9 The NWN model at the top level has a main place modeling the overall co-culture system. It considers the culture environment as a control system for the heterogeneous, spatially organized cell population from the middle level. Reservoir places hold chemical cues of interest, such as the aTc signal, which can be administered to the culture system when the dedicated intra-layer transition is activated. Cross-layer writing mechanisms pass such signals to the spatial grid net. Cross-layer reading mechanisms based on the spatial grid net's marking for EGFP signals are instead able to produce a fluorescence signal net instance at the top level

7.4.2.3 Top Level: The Culture Environment

This level, represented in Figs. 7.8 and 7.9, models the role of the culture environment in regulating and evaluating the cell culture conditions and possibly organizing the cells into specific spatial geometries. A single place models the overall cell culture, holding the net instance of the spatial grid defined by the middle level. Intra-level transitions allow to move stimuli of interest from dedicated reservoir places to the culture environment place (e.g., the aTc net instance).

Such place has cross-layer communication channels to administer stimuli to, or detecting signals from, the lower level. Intra-layer transitions can produce tokens labeled as signals, such as the aTc molecular inducer. Cross-layer transitions are used here to move aTc molecular inducers to the middle level or to read from it the presence of EGFP tokens and mark the top-level place with a fluorescence token.

7.5 Challenges and Future Work

Through this chapter, we tried to review the importance that computational models and in particular multi-level models such as NWNs will play in the near future for the design and development of complex synthetic biological systems.

Contributions from diverse fields and realities make the design effort and knowledge production in Synthetic Biology a collaborative action. This determines the necessity of speaking a common language both for descriptive and operative purposes. This translates into the need of choosing representations for the components, their manipulation, and the resulting system designs which are universally understandable. Multi-level computational models, and NWNs among them, are tools supporting this goal.

However, when investigating different modeling approaches, it is necessary to consider that Synthetic Biology basically deals with the organization of existing knowledge into top-down designs. Portions of a model could be already available and described using different formalisms or generated ad hoc for a particular design. This makes standardization a key requirement toward the definition of a globally accepted modeling solution in this domain.

Finally, while the primary goal of a computational model is to enable automatic analysis and simulation, it is also important to assure that each part of the model could be straightforwardly referred to a human-readable representation of the system. Therefore, the readability of the formalism used to describe the model is another key aspect to take into account.

References

Bardini R, Benso A, Di Carlo S, Politano G, Savino A (2016) Using nets-within-nets for modeling differentiating cells in the epigenetic landscape. Lect Notes Comput Sci 9656:315–321

Bardini R, Politano G, Benso A, Di Carlo S (2017a) Multi-level and hybrid modelling approaches for systems biology. Comput Struct Biotechnol J 15:396–402

Bardini R, Politano G, Benso A, Di Carlo S (2017b) Using multi-level Petri nets models to simulate microbiota resistance to antibiotics. IEEE Int Conf Bioinforma Biomed

Benso A, Di Carlo S, Politano G, Savino A, Bucci E (2014) Alice in "Bio-Land": engineering challenges in the world of life sciences. IT Prof 16:38–47

Bonzanni N, Feenstra KA, Fokkink W, Heringa J (2014) Petri nets are a biologist's best friend. In: Fages F, Piazza C (eds) Formal methods in macro-biology. FMMB 2014. Lecture notes in computer science, vol 8738. Springer, Cham

Cabac L, Haustermann M, Mosteller D (2016) Renew 2.5–towards a comprehensive integrated development environment for petri net-based applications. In: International conference on applications and theory of Petri nets and concurrency. pp 101–112

Cameron DE, Bashor CJ, Collins JJ (2014) A brief history of synthetic biology. Nat Rev Microbiol 12:381–390

Cull P, Flahive M, Robson R (2005) Difference equations: from rabbits to chaos. Springer, New York

Duchier D, Klutter C (2006) Biomolecular agents as multi-behavioural concurrent objects ENTCS, p 150

Esvelt KM, Whang HH (2013) Genome-scale engineering for systems and synthetic biology. Mol Syst Biol 9:641

Goers L, Freemont P, Polizzi KM (2014) Co-culture systems and technologies: taking synthetic biology to the next level. J R Soc Interface 11

Gorochowski TE (2016) Agent-based modelling in synthetic biology. Essays Biochem 60 (4):325–336

Heiner M, Gilbert D (2011) How might petri nets enhance your systems biology toolkit. In: Applications and theory of Petri nets, pp 17–37
Heiner M, Gilbert D, Donaldson R (2008) Petri nets for systems and synthetic biology formal methods for computational. Syst Biol 5016:215–264
Heiner M, Herajy M, Liu F, Rohr C, Schwarick M (2012) Snoopy – a unifying Petri net tool. Lect Notes Comput Sci 7347:398–407
Jensen K (2013) Colored Petri nets: basic concepts, analysis methods and practical use, vol 1. Springer, Heidelberg
Jones DS, Plank M, Sleeman BD (2009) Differential equations and mathematical biology. CRC, London
Maus C, Rybacki S, Uhrmacher AM (2011) Rule-based multi-level modeling of cell biological systems. BMC Syst Biol 5:166
North MJ, Collier NT, Vos JR (2006) Experiences creating three implementations of the repast agent modeling toolkit. ACM Trans Model Comput Simul 16-1:1–25
Purnick PEM, Weiss R (2009) The second wave of synthetic biology: from modules to systems. Nat Rev Mol Cell Biol 10:410–422
Qaisar U, Yousaf S, Rehman T, Zainab A, Tayyeb A (2017) Transcriptome analysis and genetic engineering. In: Cirillo P (ed) Applications of RNA-Seq and omics strategies – from microorganisms to human health. InTech. https://doi.org/10.5772/intechopen.69372
Regev A, Silverma W, Shapiro E (2001) Representation and simulation of biochemical processes using the pi-calculus process algebra. Pac Symp Biocomput 6:459–470
Swat M, Thomas GL, Belmonte JM, Shirinifard A, Hmeljak D, Glazier JA (2012) Multi-scale modeling of tissues using CompuCell3D. Comput Methods Cell Biol, Methods Cell Biol 110:325–366
Uhrmacher AM, Degenring D, Zeigler B (2005) Discrete event multi-level models for systems biology. Springer, Berlin, pp 66–89
Valk R (2004) Object Petri nets: using the nets-within-nets paradigm. Lect Notes Comput Sci 3098:819–848
Wang W, Chen Z, Kang B, Li R (2008) Construction of an artificial intercellular communication network using the nitric oxide signaling elements in mammalian cells. Exp Cell Res 314:699–706

Chapter 8
Computational Techniques for a Comprehensive Understanding of Different Genotype-Phenotype Factors in Biological Systems and Their Applications

Abhishek Subramanian and Ram Rup Sarkar

Abstract Identification and analyses of the discrete genotype and phenotype components of a biological system are complex, multi-scale problems. Computational techniques are widely used to extract some meaningful information from the underlying heterogeneous, raw biological data. Here, we review the use of different approaches to mathematically model biological data at different levels of complexities. At the raw sequence level, we discuss about the various techniques that model biological sequences based on frequency, geometry and spectral representation of nucleotides/amino acids. We also discuss about techniques that can be used to capture quantitative patterns of codons and briefly present an application in identification of evolutionary mechanisms that preserve codon usage patterns in trypanosomatids. Lastly, we discuss about the integration of the genotype and phenotype components into a systems-level mathematical representation of functional interactions in the form of a reconstructed genome-scale metabolic network and discern its role in governing variations in metabolic behaviours under different environmental conditions.

Keywords Homology · Alignment-free sequence comparison · Codon usage · Functional annotation · Genome-scale reconstruction and constraint-based modeling

A. Subramanian · R. R. Sarkar (✉)
Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, Pune, Maharashtra, India

Academy of Scientific & Innovative Research (AcSIR), CSIR-NCL Campus, Pune, India
e-mail: rr.sarkar@ncl.res.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_8

8.1 Introduction

Biological outcomes like survival, growth or proliferation and evolution depend upon the synchronous activity of multiple processes governed by a set of biological macromolecules, which do not function in isolation, but as a system of elements. These processes work at different levels of biological organization, where each level can be understood by different physicochemical principles. The biological whole is typically greater than sum of its parts (Fridolin and Green 2017). To analyse these processes, a systems biology approach which integrates a multitude of biological data at various levels of organization to extract meaningful information is largely required (Snoep and Westerhoff 2005). The genotype and the phenotype represent discrete components of a biological system which work in unison within an organism to achieve the cellular outcomes (Orgogozo et al. 2015). Genotype refers to the static genetic complement that encompasses the arrangement of coding and non-coding elements within an organism's genome, which is inherited from generation to generation. The phenotype demonstrates itself as a dynamic complement of interactions between the coding genotype elements, which are observable properties like variable expression of genes/proteins or the coherent behaviours of biological pathways.

High-throughput sequencing of genomes has helped to provide raw data about the genotype of an organism. Many sequence-based bioinformatics approaches are continuously being developed to analyse this data to extract meaningful information. Many alignment-based and alignment-free approaches of sequence comparison are employed to predict the intronic and exonic regions within the genome, genic and intergenic regions, annotate gene structure and function and identify evolutionary events and relatedness between genes across species (Mount 2004; Zielezinski et al. 2017). Once the genes are identified, their corresponding protein sequences are annotated for their function with respect to the presence of functional domains or motifs within the sequence, using functional genomics approaches. These functions can be assembled and visualized as networks that represent the actual functional interactions between the parts that make the whole, along with integration of data from high-throughput omics experiments (Henry et al. 2011).

The aim of this chapter is to introduce the reader to different mathematical and computational approaches that analyse the genotype and phenotype complements of an organism, at different levels of biological data organization. We present a brief summary of the popular sequence alignment-based approaches used for biological sequence comparison and profoundly review the alignment-free approaches for visualization and comparison of biological sequences emphasizing their use in this context. Next, after identification of primary coding sequences, one can use the comparison of its coding elements or codons to understand the variations in the codon usage as a second level of complexity. Using a large-scale codon usage comparative study in trypanosomatids, we demonstrate the use of sequence-based measures in identifying evolutionary mechanisms that can constrain the choice of codons and thereby regulate translation. Finally, we discuss about a system-level approach of integrating heterogeneous

biological data to reconstruct the metabolic network of a species *Leishmania infantum* JPCM5 and thereby identify the variations in the metabolic pathway phenotype, under different environmental conditions.

8.2 Computational Techniques for Biological Sequence Visualization and Comparison

8.2.1 Homology-Based Sequence Comparison

Homology refers to the common evolutionary ancestry that two biological entities (DNA/RNA/protein) share with each other. Homology is either an outcome of speciation (orthology) or gene duplication (paralogy) events. Homology between nucleotides or proteins can be discovered by computing the percentage of identity or similarity between their sequences. For this calculation, it is necessary to perform an alignment of sequences and then calculate similarities. The principle of evolutionary conservation between sequences can be used as a tool for identification of proteins of similar function, reconstruction of phylogenetic trees of proteins/protein families and identification of conserved structural elements within genes/proteins (Mount 2004).

Alignment-based techniques can be broadly classified as global or local alignment-type approaches. The Needleman-Wunsch algorithm is one of the most popular global alignment techniques, which uses dynamic programming for comparison (Needleman and Wunsch 1970). As opposed to this, the Smith-Waterman alignment uses dynamic programming to identify the most optimal local alignment between two sequences (Smith and Waterman 1981). Both these algorithms are implemented within the EMBOSS software package available as standalone or server versions (Rice et al. 2000). For aligning multiple sequences, these algorithms fail due to quadratic and cubic complexities in space and time. To efficiently align multiple sequences, progressive alignments are used as an alternative to dynamic programming approaches, where the relationships are represented as a guide tree followed by the addition of sequences to this growing multiple sequence alignment (MSA). CLUSTAL-W is a software package which implements progressive alignment-based MSA (Thompson et al. 1994). A more efficient alignment based on usage of substitution matrices can be obtained by performing a pairwise BLAST between two protein sequences, which is a local heuristic approach to find the best aligning stretches and then perform local alignment extensions, accounting for gaps. The BLAST pairwise alignment is extensively used at a database level by searching a query gene/protein of unknown function against a sequence database like GenBank (Benson et al. 2010), UniProt (Apweiler et al. 2013), etc. to find appropriate sequence homologs with considerable sequence identity, higher coverage and low E-values. Another replacement to the substitution matrix-based alignment are the Hidden Markov Model (HMM)-based techniques, which generate a profile HMM using multiple sequence alignment and compare this profile with a given sequence.

Sequences performing better than a null model are considered to be statistically significant homologs. HMMER is one of the popularly implemented HMM-based sequence alignment techniques (Eddy 2001).

8.2.1.1 Limitation

The sequence identity of a gene with its homolog decreases with increasing evolutionary divergence. As the homology-based methods largely rely on evolutionary conservation of a signature in the pair of sequences being compared, they fail to identify homologs for a gene that possess a higher level of sequence divergence due to events of domain swapping, recombination and duplication of peptide motifs (Zielezinski et al. 2017).

8.2.2 Genomic Context-Based Methods

In genomic context-based methods, the assumption is that proteins having similar contexts in the genome are guaranteed to have similar function (Huynen 2000). The term 'context' as suggested depends on the context of the gene within a genome. The genomic context is defined with respect to:

(a) Conserved gene neighbourhood, where the co-occurrence of neighbouring genes (Overbeek et al. 1999) is compared across genomes
(b) Conserved gene order: the order in which the unknown gene is placed in relation with the neighbouring genes (Dandekar et al. 1998)
(c) Conserved phylogenetic profiles (Pellegrini et al. 1999), where the assumption is proteins having similar functions are likely to evolve in a correlated manner
(d) Identification of domain/gene fusion events (Enright et al. 1999), where few genes are predicted to fuse together if they demonstrate similar selective pressures

STRING (Jensen et al. 2008), MCScanX (Wang et al. 2012) and GeneMANIA (Mostafavi et al. 2008) are some software/web servers based on genomic context methods.

8.2.2.1 Limitation

The genomic context-based methods might fail in genomes showing higher degree of exon shuffling and genome rearrangements (Wolf et al. 2001).

8.2.3 Ab Initio Gene Structure Prediction

These set of methods are highly applied to eukaryotic genomes where the intrinsic information of the gene sequence is used for prediction of gene structure. These methods indirectly suggest a function for a gene. A lot of statistical approaches from information theory (Lin 1991) to Markov models (Eddy 1996; Burge and Karlin 1997; Stanke and Waack 2003) have been dominating this category. Similarly, a number of machine learning-based methods also extract features from a DNA sequence in order to classify a gene according to protein structure/function (Yang 2004; Dror et al. 2005). But these methods are dataset dependent and rely on a training dataset of very similar sequences. The accuracy of these methods is optimized in accordance with the given training data set, making it highly dataset specific.

8.2.3.1 Limitation

The ab initio methods are training set dependent and are restricted to a very specific or local application of function prediction and, hence, require good quality and large quantity information for training.

8.2.4 Alignment-Free Sequence Comparison

Apart from the above popular techniques, a new class of techniques that compare and quantify sequence similarity without the use of sequence-based alignment (residue/base to residue/base correspondence) can also be used for annotation of sequences (Zielezinski et al. 2017). These approaches are called as the alignment-free sequence comparison methods. As we will see further, alignment-free methods largely involve numerical characterization of biological sequences as word frequency, geometry or spectral dynamics. A significantly large number of methods that belong to this class of techniques exist in literature.

8.2.4.1 Word Frequency-Based Methods

Such methods are also called as the feature vector-based comparison methods. In these methods, word combinations representing a total possible set of unique words of a k-length present within the pair of sequences are computed from all given sequences (Vinga and Almeida 2003; Zielezinski et al. 2017). These word combinations are generated using an overlapping window moving from the start to the end of each sequence. Using this fixed set of words, number vectors corresponding to their counts of occurrence within a given pair of sequences are computed. Hence, each sequence can be represented as a number vector representing frequencies of *k*-mer

words. For example, consider two sequences s_1 = ATGTGTAGT and s_2 = ATGCGTTTT. For a k-mer length of 3, a union of unique triplet combinations can be computed using the overlapping window for these sequences. Therefore, we obtain W_u = {AGT, ATG, CGT, GCG, GTA, GTG, GTT, TAG, TGC, TGT, TTT}, where W_u is the union of words between the two sequences. Representing the sequences as number vectors, we compute the number vector of s_1 corresponding to counts of combinations in W_u as n_1 = (1, 1, 0, 0, 1, 1, 0, 1, 0, 2, 0) and s_2 as n_2 = (0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 2). Computing the Euclidean distance between these number vectors, it is possible to calculate the similarities between sets of sequences.

CAFÉ is an important software that implements k-mer word frequency-based methods to explore relationships among multiple DNA sequences through a graphical user interface (Lu et al. 2017). CVTree is another method that implements a modification of the k-mer composition vector approach to generate whole genome-based phylogenetic trees (Qi et al. 2004). Word frequency-based methods have been applied for annotation of G protein-coupled receptor sequences which could not be assigned to known receptor families previously. Also, functional similarities among regulatory sequences were identified in flies and mammals using composition vector methods (Zielezinski et al. 2017).

8.2.4.2 Frequency Chaos Game Representation

In 1990, Jeffrey proposed a new way to visualize nucleotides at different positions within a DNA sequence as points within a square of unit area whose vertices correspond to the four DNA bases (Jeffrey 1990). The change of nucleotides from one position to another within a sequence is represented as an iterated function system (IFS), which can be used for representing points within a square. Such a representation came to be known as the chaos game representation (CGR) of DNA sequences (Fig. 8.1). Such representations use DNA random walk models which distinguish the real pattern of a given sequence as compared to a random pattern in a scale-independent manner.

Although the CGRs can be compared visually, so as to compare a large set of sequences quantitatively, frequency CGRs (FCGRs) which give a quantitative frequency-based representation of DNA sequences were developed (Almeida et al. 2001; Wang et al. 2005). A *k*th-order FCGR of a sequence *s* is denoted by $FCGR_k(s)$, which is a $2^k \times 2^k$ matrix where *k* represents the length of subsequences, whose occurrence within the sequence is calculated. FCGR is, thus, a numerical matrix that can be computed by plotting a CGR for the sequence *s*, dividing the CGR plot into $2^k \times 2^k$ grid and calculating the number of points within each grid. It is also suggested that the FCGR matrix can be computed without plotting of an actual CGR by counting the total number of occurrences of a k-length subsequence within a sequence and place it in an appropriate position within the FCGR matrix, according to the correspondence between the k-mer subsequence and the CGR grid square.

Hence, depending upon the subsequence length of *k*, the FCGR can be of different orders:

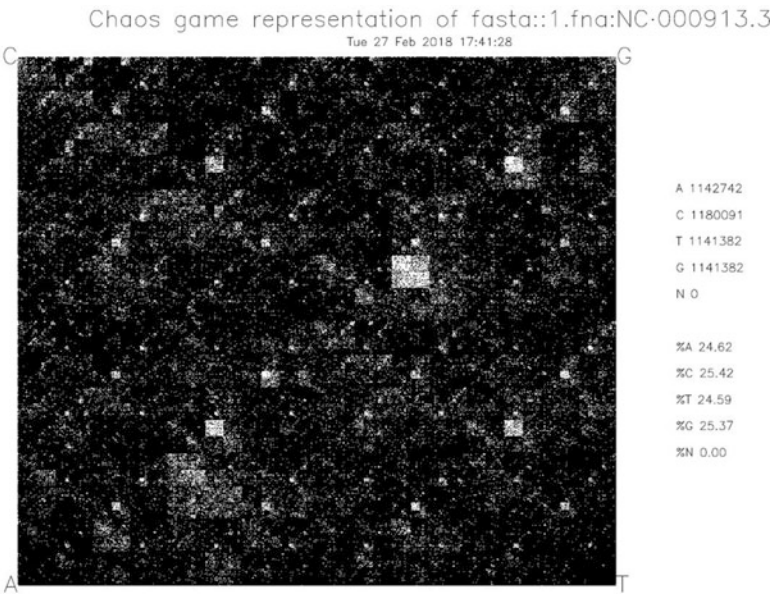

Fig. 8.1 Chaos game representation of the *E. coli* K-12 MG 1655 genome generated using the 'chaos' subroutine within the EMBOSS package (Rice et al. 2000)

$$\text{First order FCGR} : \text{FCGR}_1(s) = \begin{bmatrix} N_A & N_T \\ N_G & N_C \end{bmatrix} \tag{8.1}$$

$$\text{Second order FCGR} : \text{FGCR}_2(s) = \begin{bmatrix} N_{AA} & N_{TA} & N_{AT} & N_{TT} \\ N_{CA} & N_{GA} & N_{CT} & N_{GT} \\ N_{AC} & N_{TC} & N_{AG} & N_{TG} \\ N_{CC} & N_{GC} & N_{CG} & N_{GG} \end{bmatrix} \tag{8.2}$$

Instead of calculating just counts, the values in the FCGRs can also be replaced by relative frequencies. Such FCGRs are called as relative FCGRs. A second-order relative FCGR is also called as a dinucleotide relative abundance profile or DRAP. It is given by:

$$\text{Second order relative FCGR or DRAP} : \text{rFCGR}_2(s) = \begin{bmatrix} \sigma_{AA} & \sigma_{TA} & \sigma_{AT} & \sigma_{TT} \\ \sigma_{CA} & \sigma_{GA} & \sigma_{CT} & \sigma_{GT} \\ \sigma_{AC} & \sigma_{TC} & \sigma_{AG} & \sigma_{TG} \\ \sigma_{CC} & \sigma_{GC} & \sigma_{CG} & \sigma_{CG} \end{bmatrix} \tag{8.3}$$

where the relative frequency of the chosen subsequence is given by $\sigma_{XY} = \frac{f_{XY}}{f_X f_Y}$. f_x and f_Y represent the mononucleotide frequencies of nucleotides X and Y in sequence

s. f_{xY} represents the number of occurrences of the dinucleotide XY in sequence s. X and Y can be any of the four DNA sequence nucleotides.

kth order FCGRs of any two given sequences can be compared by calculating geometric distances like the Euclidean, Hamming, etc. or statistical distance like the Pearson correlation between the computed FCGRs. The chaos game representation has many applications. The fractal properties of a CGR have been successfully applied to large-scale comparisons of bacterial genomes obtained from metagenomics and HIV-1 group subtyping (Joseph and Sasikumar 2006; Pandit and Sinha 2010). A web server for plotting the CGR of a given DNA sequence is also available (Arakawa et al. 2009). A program called 'chaos' is available within the EMBOSS software for plotting of a chaos game representation of a given DNA sequence (Rice et al. 2000). A computer application called GENSTYLE was designed to compare the genomic signature between a given set of sequences by calculating the oligonucleotide frequencies of a given length within a sequence and computing its frequency CGR (Fertil et al. 2005).

8.2.4.3 Graphical Representations of DNA Sequences

Graphical representations of biological sequences are most interesting ways of sequence visualizations. In such methods, the biological sequences are imagined similar to atoms of a small molecule, superimposed on a 2-D or a 3-D grid. The quantitative information of the bases occurring in specific positions of the sequence can be understood using the points on the grid. The 2-D/3-D descriptors can thus be used to quantitate DNA sequences and can be further used for comparison purposes. These sets of representations were made popular by Randic et al. in the early 2000s (Randić et al. 2000, 2003a, b). These methods follow the structure-property similarity principle which suggests that similar graphical structures from DNA sequences will have similar properties.

A very simple form of graphical information is a binary plot or the zigzag curve where the possibility of occurrence of one of the four nucleotides can be plotted for every position of the sequence against the nucleotide occurring at a position within a given subsequence (Randić et al. 2003a). Here, it can be assumed that the first position as the origin with coordinates (0, 0) and the points corresponding to nucleotide position can be assumed to fall on grid points in the positive real X-Y plane (Fig. 8.2). The number of edges traversed between each nucleotide to every other nucleotide within a sequence (topological distance) and also the Euclidean distance between the vertices of the grid points (geometrical distance) are calculated to identify a numerical representation of the sequence. By calculating the quotient of geometrical distance to the topological distance between all vertices, it is possible to construct an M/M symmetric matrix. Computing the quotient of the geometrical distance between two vertices representing the nucleotides at every pair of positions and the sum of geometrical lengths of edges between the two vertices, it is also possible to construct a L/L symmetric matrix. The dimensions of this symmetric matrix increase with the length of the biological sequence. Similarly, it is possible to

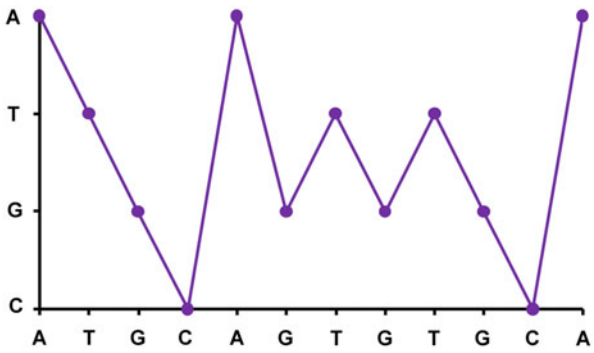

Fig. 8.2 2-D graphical representation of the DNA sequence – ATGCAGTGTGCA

calculate the higher-order $^{k}L/^{k}L$ or $^{k}M/^{k}M$ matrices by performing a Hadamard multiplication of the M/M or L/L matrix with itself *k*-times. Likewise, once enough matrices are generated, it is possible to calculate the highest eigenvalue of each matrix. This set of eigenvalues can be used to represent a DNA sequence (Randić et al. 2003a, b).

Although there is no specific physical interpretation of these eigenvalues, it may facilitate comparison of DNA sequences, as shown for the globin gene sequences across species (Randić et al. 2003a). This method was further extended for a 3-D representation of protein sequences using the triplet codon mapping while generating zigzag curves in 3-D (Randić et al. 2004). This method is very similar to the CGR representation of DNA sequences. Once represented, the leading eigenvalue-based descriptors are calculated for each sequence and compared. Slight variations to the 3-D representations have also been presented by Yuan et al. (2003) to avoid loss of information while transitioning from a DNA sequence to a numerical characterization, thereby avoiding the overlaps of the zigzag curves in 3-D space.

Another easy way of implementing graphical representations is to imagine a DNA sequence as a directed graph with four vertices each representing a base (Qi et al. 2011). An adjacency matrix of this graph with weights between every pair of nucleotides corresponding to the frequency of paired nucleotide transitions (directed from the previous nucleotide to the all the future nucleotides) can be calculated. Comparing adjacency matrices of the different DNA sequences, it is possible to compute the similarity/dissimilarity between sequences. Such adjacency matrix-based representations are shown to be highly resistant to cases of exon shuffling.

8.2.4.4 Linear Regression Model Representation

A Markov-based modification of Randic's 2-D graphical representation is formulated to model DNA sequence as a random process relevant to time (Dai et al. 2007). Let $a = a_1a_2a_3...a_n$ be a given DNA sequence with n bases. In other words, for a sequence a of length 10, the entire sequence a can be formed by sampling a base ten

times. From basic molecular biology, it is known that the four DNA bases can be classified into $R = \{A,G\}$ and $Y = \{C, T\}$, $M = \{A,C\}$ and $K = \{G, T\}$, $W = \{A, T\}$ and $S = \{G,C\}$ according to their chemical properties. Hence, it is possible to formulate distribution functions for particular nucleotides (PNDF) using these chemical characteristics. Let us use the purine-pyrimidine (R-Y) chemistry. With respect to the R-Y chemistry, the PNDFs can be represented as:

$$\mathrm{PNDF}_A^R(t) = \frac{\sum_{k=1}^{t} \sigma_A(a(k))}{\sum_{k=1}^{t} \sigma_A(a(k)) + \sum_{k=1}^{t} \sigma_G(a(k))} \tag{8.4}$$

$$\mathrm{PNDF}_G^R(t) = \frac{\sum_{k=1}^{t} \sigma_G(a(k))}{\sum_{k=1}^{t} \sigma_A(a(k)) + \sum_{k=1}^{t} \sigma_G(a(k))} \tag{8.5}$$

$$\mathrm{PNDF}_T^R(t) = \frac{\sum_{k=1}^{t} \sigma_T(a(k))}{\sum_{k=1}^{t} \sigma_T(a(k)) + \sum_{k=1}^{t} \sigma_C(a(k))} \tag{8.6}$$

$$\mathrm{PNDF}_C^R(t) = \frac{\sum_{k=1}^{t} \sigma_C(a(k))}{\sum_{k=1}^{t} \sigma_T(a(k)) + \sum_{k=1}^{t} \sigma_C(a(k))} \tag{8.7}$$

where $\sigma_X(a(k)) = \begin{cases} 1 & if\ a(k) = X \\ 0 & else \end{cases}$.

For mathematical purposes, R represents the union of the occurrences of R and Y classes of nucleotides. Using these distribution functions, the DNA sequences can be graphically represented as a linear regression of a purine (R) onto the pyrimidine (Y) within the sequence. Similarly, it can be constructed for other classes as well. The generalized linear regression model can be written as:

$$\mathrm{PNDF}_P^R(t) = a_{\mathrm{PQ}}^R + b_{\mathrm{PQ}}^R \mathrm{PNDF}_Q^R(t) + \varepsilon_{\mathrm{PQ}}^R \tag{8.8}$$

where P and Q are any of the four nucleotides regressed over one other.

It is easy to estimate the parameters a_{PQ}^R, b_{PQ}^R and $\varepsilon_{\mathrm{PQ}}^R$ by performing a linear regression (Dai et al. 2007). Hence, each sequence numerically can be represented as parameter matrices, namely, A_{PNDF}^Z and B_{PNDF}^Z, where each matrix value is a parameter value obtained after regression between the distributions of any two nucleotides belonging to the R-Y class. Parameter matrices for two DNA sequences

can be compared using a distance function based on condensed parameter matrices (DCMP) which is given as:

$$\mathrm{DCMP}^{Z}(X|Y) = \sqrt{\sum_{u=1}^{4}\sum_{o=1}^{4}\left(\left(A_X^Z\right)_{\mathrm{uo}} - \left(A_Y^Z\right)_{\mathrm{uo}}\right)^2} + \sqrt{\sum_{u=1}^{4}\sum_{o=1}^{4}\left(\left(B_X^Z\right)_{\mathrm{uo}} - \left(B_Y^Z\right)_{\mathrm{uo}}\right)^2} \tag{8.9}$$

where $Z \in \{M, R, W\}$.

A further extension of this method was applied for k-words where the distribution of all k-words was generated and the linear regression between any two k-words in a sequence was performed (Yang and Wang 2013). Apart from DNA sequences, this method was also applied on protein sequences to generate distributions using their physicochemical properties (Qi and Jin 2016).

8.2.4.5 Symbolic Dynamics

The protein coding regions of the DNA sequences tend to exhibit a period 3 pattern, because the triplet codon structure typically codes for a protein (Tsonis et al. 1991). Hence, the three-period property is a good indicator of position of a gene within a genome. Period 3 is also known to imply chaos. Hence, a chaotic dynamics property can exist within the exons of a gene. Similar to the numerical representations of a biological sequence as observed in graphical representations of sequences, using symbolic dynamics, it is possible to map biological sequences into chaotic sequences, which contain the direct information existing within a biological sequence (Wang et al. 2009a). These mappings numerically convert each nucleotide within a sequence into a 3-D representation where the three coordinates signify the physicochemical properties of that nucleotide (Wang et al. 2009b). The values of these coordinates are obtained from a binary operator that corresponds to these properties.

Let us assume that the given DNA sequence is represented by a set of characters $\{u_m\}$. Each character can be mapped onto three symbolic sequences given by:

$$\begin{aligned} s_m^1 &= \begin{cases} 1 & u_m = R \\ 0 & u_m = Y \end{cases}, \\ s_m^2 &= \begin{cases} 1 & u_m = M \\ 0 & u_m = K \end{cases}, \\ s_m^3 &= \begin{cases} 1 & u_m = W \\ 0 & u_m = S \end{cases}. \end{aligned} \tag{8.10}$$

where R represents the purine nucleotides A and G; Y represents the pyrimidine nucleotides T and C; M represents nucleotides that switch between tautomers of imino/amino forms, namely, A and C; K represents nucleotides switching between tautomers of enol/keto forms, namely, G and T; W represents nucleotides associated with weak H-bonds, namely, A and T, and S represents nucleotides with strong H-bonds, namely, G and C.

Thus, with respect to the property of a given nucleotide, each nucleotide within the sequence can be numerically represented as symbolic sequences $s_m\left(s_m^1, s_m^2, s_m^3\right)$. For example, if we consider a triplet nucleotide sequence d => ATG, it can be numerically represented as $d = \{(1, 1, 1), (0, 0, 1), (1, 0, 0)\}$.

Now, we map the set of symbolic sequences for a stretch of nucleotides which is a subsequence of the entire DNA sequence. This mapping is given by:

$$x_m^{\mathrm{i}} = \sum_{k=m}^{m+N-1} s_k^i M^{-(k-m+1)} \tag{8.11}$$

where $i = 1, 2$ or 3, m is a given position within a sequence, N is the truncated length (length of a subsequence) and M represents the number of partitions within the entire space of the chaotic system. Simply speaking, it represents the number of groups used to disintegrate a given nucleotide into its numerical features, as given in (8.10). From (8.10), the number of groups for representing each feature is 2. Hence, (8.11) can be rewritten as:

$$x_m^{\mathrm{i}} = \sum_{k=m}^{m+N-1} s_k^i 2^{-(k-m+1)} \tag{8.12}$$

The subsequences of a given truncation length can be generated using an overlapping window moving through a sequence. Such a method is called continuous method of mapping. This method is recommended for DNA sequences with the truncated length N in multiples of 3 to capture three periodicities of coding sequences. The subsequences of a given truncation length can be also be generated discontinuously without overlaps. This method is recommended for mapping protein sequences. It is called as the discontinuous mapping method. Once each sequence is represented by sets of x_m mappings, the similarity or dissimilarity between sequences can be calculated. As it follows the three-period pattern, this representation can be coupled with signal processing techniques to predict gene signals within genome sequences (Wang et al. 2009a). Another application of symbolic representations is to identify tandem repeats within sequences by mapping the symbolic dynamic sequences of nucleotides onto a frequency domain thereby identifying a global repeat map (Glunčić and Paar 2012). A computer program called Spectral Repeat Finder is available for this purpose (Sharma et al. 2004).

8.2.4.6 Spectral Dynamic Representation

As mentioned previously, performing graphical representation of DNA sequences, it is possible to extract numerical descriptors unique to a biological sequence. Similar to these representations, one can also represent a DNA sequence as a spectrum where the distribution of a particular base within a DNA sequence is represented as a series of lines resembling atomic or molecular spectra and extract descriptors in the form of moments of inertia constructed from the spectrum. Bielińska-Wąż and Wąż (2017) define a DNA sequence as the B-spectrum (a spectrum of bases) with B = A, T, G, C. The position of the *i*th line within the B-spectrum denoted by v_i^B, called as the frequency and the length of the spectral line representing a base at a particular position, is denoted by $I_B\left(v_i^B\right)$, called as the intensity of a line, where $i = 1,2,\ldots.N_B$. For simplicity, the length of all the lines within the spectrum can be taken as equal to 1. As N_B is the total number of lines within a DNA sequence, it is the sum of number of lines representing all the four bases. For example, a sequence ATGCGT can be represented by the lines:

$$\begin{aligned} &A : I_A\left(v_1^A\right) = 1, v_1^A = 1 \\ &T : I_T\left(v_1^T\right) = 1, I_T\left(v_2^T\right) = 1, v_1^T = 2, v_2^T = 6 \\ &G : I_G\left(v_1^G\right) = 1, I_G\left(v_2^G\right) = 1, v_1^G = 3, v_2^G = 5 \\ &C : I_C\left(v_1^C\right) = 1, v_1^C = 4 \end{aligned} \tag{8.13}$$

These four lines can be represented as vertical spectral lines connecting the following points and a horizontal axis representing the positions within a sequence.

$$\begin{aligned} &A : \left(v_1^A, 0\right), \left(v_2^A, 0\right), \ldots, \left(v_{N_A}^A, 0\right) \\ &T : \left(v_1^T, 0\right), \left(v_2^T, 0\right), \ldots, \left(v_{N_T}^T, 0\right) \\ &G : \left(v_1^G, 0\right), \left(v_2^G, 0\right), \ldots, \left(v_{N_G}^G, 0\right) \\ &C : \left(v_1^C, 0\right), \left(v_2^C, 0\right), \ldots, \left(v_{N_C}^C, 0\right) \end{aligned} \tag{8.14}$$

Details can be found in the representative figures of the B-spectrum in (Bielińska-Wąż and Wąż 2017).

Assigning a mass m_i^B to each point $\left(v_i^B, 0\right)$, we obtain four massive bodies composed of point masses distributed along the horizontal axis. The moments of inertia of these bodies are equal to:

$$M_B = \sum_{i=1}^{N_B} m_i^B \left(\tilde{v}_i^B\right)^2, \tag{8.15}$$

where $\tilde{v}_i^B = v_i^B - v_i^A$, and v_a^B is the centre of mass of the bodies with $\left(v_a^B, 0\right)$ as the coordinates given by:

$$v_a^B = \frac{1}{N_B} \sum_{i=1}^{N} v_i^B. \tag{8.16}$$

Setting $m_i^B = 1$ for each point, the total mass of the spectrum becomes equal to the number of points. Hence,

$$\sum\nolimits_{i=1}^{N_B} m_i = N_B. \tag{8.17}$$

Hence, the normalized moments of inertia can be defined as:

$$r_B = \sqrt{\frac{M_B}{N_B}}. \tag{8.18}$$

The coordinates of the centre of mass divided by the normalized moments of inertia given by

$$D_B = \frac{v_a^B}{r_B} \tag{8.19}$$

can be used as descriptors of a DNA sequence. Thus, a vector of these descriptors for a DNA sequence can be given by $D_B = [D_A, D_T, D_G, D_C]$. Computing the standard distance between the vectors of a pair of sequences, one can possibly calculate the similarity between any two given sequences. The derived moments of inertia in this case are called as 1-D moments of inertia. As performed in other cases, this method was tested on the sequences of β globin genes available from mammals (Bielińska-Wąż and Wąż 2017). 2-D or 3D moments of inertia were also used as descriptors for dynamic graph of protein sequences (Yao et al. 2008, 2014; Hou et al. 2016).

Advantages

All alignment-free methods have few common advantages.

1. Mathematically well-founded in nature and calculate the pairwise dissimilarity or distance between a pair of sequences.
2. They do not depend on the assumptions of optimization based on evolutionary relatedness between sequences.
3. Hence, these techniques are resistant to domain shuffling and recombination events which produce a low conservation between sequences.
4. As these techniques do not rely on dynamic programming or greedy algorithms, they are computationally efficient as compared to alignment-based methods and can be useful for comparison of whole genomes.

Disadvantages

1. Although alignment-free methods are computationally inexpensive, they are largely memory inefficient.

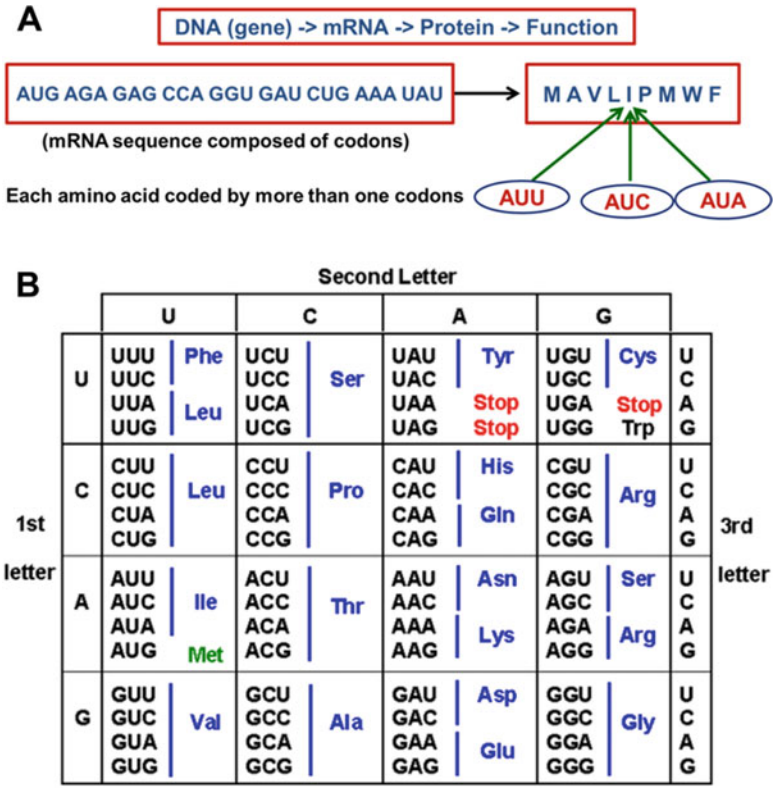

Fig. 8.3 Degeneracy in the genetic code – (**a**) central dogma of molecular biology; (**b**) standard genetic code. (Courtesy: https://openwetware.org/wiki/CH391L/S12/TranslationRBSandCodons)

2. Fails to detect complex levels of sequence organization as seen in protein sequences with long-range interactions between amino acid residues playing an important role.
3. Considers all the existing information within sequences without filtering based on prior knowledge for sequence conservation. Hence, appropriate comparisons can be noisy.

8.3 Comparative Codon Usage Analysis

With respect to the central dogma of molecular biology (Fig. 8.3), the gene is transcribed into mRNA by transcription (Crick 1970). This mRNA contains the encoded information of the function of the gene. By the process of translation, the encoded information within the mRNA is translated into a sequence of amino acids,

which further form the protein that governs that function. Codons are triplet nucleotides present within the coding sequence/mRNA of a gene that code for an amino acid. It is often observed that there is degeneracy in the choice of the codons coding for a particular amino acid. There are amino acids that are either encoded by 2, 3, 4 or 6 codons. This association of the amino acid and the codon is proposed by a codon usage table, which is given by the standard genetic code (Fig. 8.3). Furthermore, it is also commonly observed that certain codons are chosen with a higher frequency to code for an amino acid as compared to other alternative codons (Plotkin and Kudla 2010). This phenomenon is broadly termed as codon usage bias (CUB). Evolutionary events like mutation pressure and translation selection are the most explored causes of CUB in many organisms. Mutation pressure suggests that codons with a high occurrence of specific nucleotides are chosen within a genome (Sueoka 1988). Translation selection suggests codons that benefit translational efficiency are chosen within a genome (Hershberg and Petrov 2008; Plotkin and Kudla 2010). To study the roles of mutation pressure and translation selection, a number of sequence-based measures can be calculated and compared across genes within an organism or genes across species.

8.3.1 Sequence-Based Measures of Codon Usage

There are numerous sequence-based measures of codon usage that can be used for identification of mutation pressure and translation events within genes.

8.3.1.1 Mutation Pressure

The mutation pressure towards maintaining AT or GC content in a given gene can be calculated by counting the number of substitutions between the given gene in the first genome and its ortholog in its closely related genome (Sueoka 1988; Alonso et al. 1992). The mutation pressure of the given gene is calculated as:

$$\text{Mutation pressure} = \frac{p}{q}, \tag{8.20}$$

where p is the total number of substitutions of G or C to A or T between the pair of orthologous genes and q is the total number of substitutions of A or T to G or C between the same pair of genes. If p/q is greater than 1, mutation pressure towards AT is expected for the second gene as compared to first gene. If p/q is less than 1, mutation pressure towards GC is expected for the second gene as compared to first gene.

8.3.1.2 Relative Synonymous Codon Usage (RSCU)

The relative synonymous codon usage (RSCU) is the observed frequency of a particular codon coding for an amino acid within a gene scaled by the number of synonymous codons for that amino acid (Sharp and Li 1987). RSCU is given by:

$$\mathrm{RSCU}_{ij} = \frac{x_{ij}}{\frac{1}{k_i}\sum_{j=1}^{k_i} x_{ij}}, \tag{8.21}$$

where RSCU_{ij} is the relative synonymous codon usage for the *j*th codon coding for the *i*th amino acid, x_{ij} is the total count of codon '*j*' coding for amino acid '*i*' and k_i is the number of alternate codons coding for that amino acid.

8.3.1.3 Codon Adaptation Index (CAI)

Codon adaptation index is a widely used measure of codon usage that assesses the degree of translation selection acting upon a gene (Sharp and Li 1987). CAI is calculated for every gene relative to a known reference set of highly expressing genes (for example, set of ribosomal genes) as a geometric mean of RSCU values. CAI is given by:

$$\mathrm{CAI} = \exp\frac{1}{L}\sum_{i=1}^{18}\sum_{j=1}^{k_i} x_{ij} \ln\left(w_{ij}\right), \quad \text{where } w_{ij} = \frac{\mathrm{RSCU}_{ij}}{\mathrm{RSCU}_{i\max}} \tag{8.22}$$

$\mathrm{RSCU}_{i\max}$ is the relative synonymous codon usage of the most frequently used codon for the *i*th amino acid, and *L* is the length of the gene. The weight w_{ij} is calculated from a reference set of highly expressing genes. The values of CAI are scaled between 0 and 1. Further, CAI >0.5 also indicates a high degree of translation selection.

8.3.1.4 Effective Number of Codons (ENC)

The ENC index is a nondirectional measure of codon usage bias which calculates the sum of contributions of codons for two-, three-, four- and sixfold degenerate amino acids to codon usage bias within a gene (Wright 1990). The ENC index has a strong relationship with the nucleotide composition at the third synonymous position of the codon. The ENC index is calculated using,

$$\mathrm{ENC} = 2 + \frac{9}{\overline{\widehat{F}_2}} + \frac{1}{\overline{\widehat{F}_3}} + \frac{5}{\overline{\widehat{F}_4}} + \frac{3}{\overline{\widehat{F}_6}}, \tag{8.23}$$

where $\overline{\widehat{F}_i}$ is the average codon homozygosity estimate for amino acids having degeneracy of 'i' codons. The values of the ENC index are scaled between 20 and 61. A gene with ENC value of 20 indicates a high codon usage bias in the gene, whereas an ENC value of 61 suggests an equal contribution of each codon within the gene to code for an amino acid.

8.3.2 *Applications of Codon Usage to a Large-Scale Cross Species Comparison Across Trypanosomatids*

Codon usage bias was speculated to be an important mechanism for global translation regulation in *Trypanosoma* species, a close evolutionary relative of the *Leishmania* parasite. The roles for mutation pressure and translation remained unknown for *Leishmania* species. To investigate the causes and consequences of codon usage bias, a large-scale comparative codon usage analysis across 13 species of trypanosomatids was performed using sequence-based measures of codon usage (Subramanian and Sarkar 2015). Comparisons indicated directional mutation pressure and translation selection to be evolutionary mechanisms that reinforce the choice of codons ending with a G or C in *Leishmania* genomes. Further, frequent codons were found to form energetically less stable secondary structures at the 5′ end of the mRNA and, hence, can be a putative mechanism of translation regulation. The occurrence of codon context pairs follows codon usage bias and is related to efficient translation elongation. A high codon adaptation is observed in energy metabolism, translation and stress-related genes of *L. infantum* and *L. donovani* as opposed to other species, suggesting species specificity in codon usage.

8.4 Applications to Functional Annotation of Enzymes for Genome-Scale Metabolic Reconstruction and Analysis

The alignment-based or the alignment-free methods can be used to annotate enzymes that assist in the conversion of a substrate metabolite into a product within cellular metabolism from fully sequenced genomes. Once annotated, the identified genes can be associated to metabolic reactions that form functional pathways. The association of genes, enzymes and reactions into a functional framework of interacting metabolic reaction components is termed as genome-scale reconstruction.

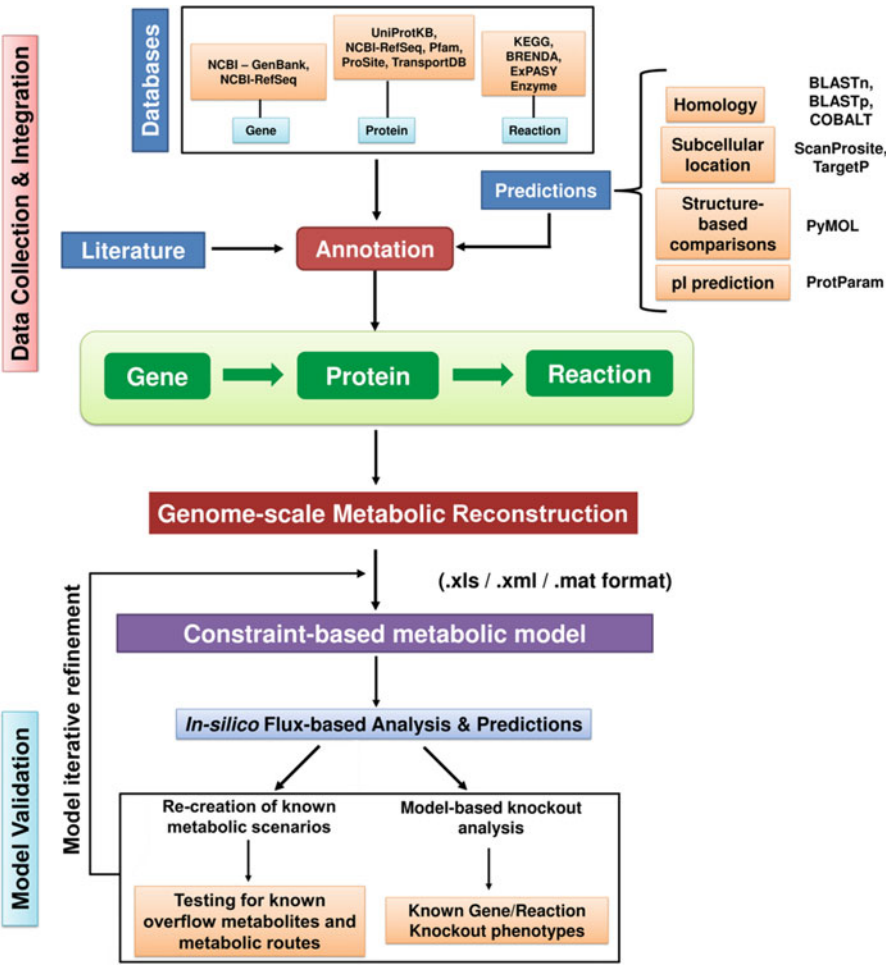

Fig. 8.4 Strategy for reconstruction of the genome-scale metabolic reconstruction

As described in literature (Subramanian et al. 2015; Subramanian and Sarkar 2017), genome-scale metabolic reconstruction (Fig. 8.4) involves the following important steps:

8.4.1 Annotation of Enzyme Function

The molecular function of each metabolic enzyme, both at the nucleotide and protein sequence level, needs to be annotated through primary sequence search using the aforementioned alignment-based or alignment-free methods against databases like

NCBI GenBank (Benson et al. 2008), UniProt (UniProt Consortium 2014) and KEGG (Kanehisa et al. 2013); motif/domain search against domain databases, like Pfam (Finn et al. 2013) and PROSITE; or by using other annotation techniques mentioned in the previous section. This also needs to be supplemented with experimental proof of existence. After the identification of genes/proteins, they can be associated with reactions catalysed by the metabolic enzymes and their corresponding enzyme classification numbers (EC Nos.) from databases like BRENDA (Schomburg et al. 2012), ExPasy Enzyme (Bairoch 2000) and KEGG (Kanehisa et al. 2013). Annotation can also be extended to account for substrate specificity of the enzyme by comparison with homologs of known function in other species.

8.4.2 Annotation of Subcellular Location of Enzymes

As a first priority, reactions need to be assigned to their appropriate subcellular compartments on the basis of available evidence of occurrence in a particular compartment for that species. Experimental evidence of subcellular location from mass spectrometry, subcellular fractionation, fluorescent labelling, etc. is ideal for annotating the subcellular location of an enzyme (Huh et al. 2003). Along with literature evidence, the corresponding protein sequence can be perused for sequence-based subcellular targeting signals to affirm the subcellular location of an enzyme specific to that species. Tools like TargetP and SignalP can be used for predicting the presence of either a mitochondrial target peptide or an endoplasmic reticular signal within a protein sequence (van Weelden et al. 2005; Emanuelsson et al. 2007). For transport of a protein to specific organelles like peroxisomes or glycosomes, ScanProsite is an important tool that searches for the sequence motifs within protein sequences. It can search for a peroxisomal targeting sequence (PTS) signal either at the C-terminal of a peptide sequence (PTS-1 type) or at the N-terminal of the protein sequence (PTS-2 type) (Borst 1986; Opperdoes and Szikora 2006). The PTS-1 signal is a tripeptide signalling sequence present at the end of a protein sequence, given by the PROSITE signature PS00342 (De Castro et al. 2006; Opperdoes and Szikora 2006). The last three peptides in a sequence can also be analysed for partial PTS-1 signals by considering the condition that any of the three peptides can be altered by other optional amino acids. The PTS-2 signal is a variable pattern sequence that can be specified with respect to the type of peroxisome and organism. For example, to identify a N-terminal signalling sequence targeting the glycosome (a special form of the peroxisome), one can use the PROSITE pattern <M-x(0,20)-[RK]-[LVI]-x(5)-[HKQR]-[LAIVFY] (Opperdoes and Szikora 2006). As proteins present at different subcellular locations possess different isoelectric points (pI), the subcellular localization of an enzyme can also be annotated with respect to the isoelectric point of the protein. Calculation of isoelectric point (pI) of enzymes is an important part of subcellular location annotation. Isoelectric point for every protein can be predicted in silico using the ProtParam Tool (Gasteiger et al. 2005).

In *Leishmania*, for example, almost all glycosomal proteins have isoelectric point (pI) in the range of 8.8–10.2 (Misset et al. 1986; Michels 1988; Guerra-Giraldez et al. 2002).

8.4.3 Incorporation of Transport Reactions

The transports of previously reported carbon and nitrogen sources, fermentation products, vitamins, ions/protons and other overflow metabolites need to be included within the metabolic network reconstruction. Transporters for various organisms are available in the TransportDB database (Ren et al. 2006), which can either be used directly or by annotating proteins within an organism of interest by performing a sequence comparison.

8.4.4 Model Iterative Refinement for Filling up Missing Metabolic Gaps

The raw metabolic network reconstructions formed in the previous step need to be iteratively refined to identify missing gaps by comparing with known experimental observations for a given organism. This can be done in conjugation with flux balance analysis (FBA) and the GapFill algorithm, a method for computational prediction of reactions fluxes within the metabolic network under given environmental transport conditions. By doing so, the most optimal pathways activated to utilize a particular environmental metabolite can be predicted and compared with known experiments like targeted metabolomics and gene knockout/RNA interference studies. This step is required as it identifies gaps and dead-end metabolites within the network and new enzymes can thus be searched within the genome to fill the gaps and improve the reconstruction. The modified network can be again refined through the above procedure until the network is sufficiently able to reproduce the observations reported for the organism of interest. By an iterative refinement procedure, reactions in appropriate subcellular reactions (e.g., novel placement of fatty acid oxidation enzymes in glycosome for NAD redox coupling with glycolysis) and novel reactions (like threonine aldolase [UniProt ID: A4HRH1]) were identified in *Leishmania infantum* JPCM5.

8.4.5 Flux Balance Analysis

The model structure contains a stoichiometric matrix S, a mathematical representation of the reconstructed network. The S matrix consists of m reactions and

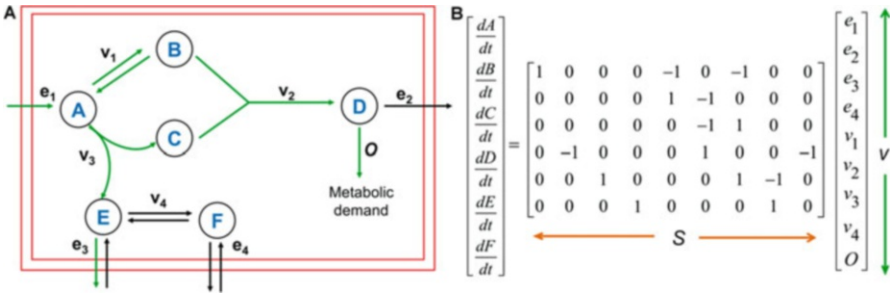

Fig. 8.5 Stoichiometric representation of the metabolic network – (**a**) a toy metabolic network demonstrating the interconversion of metabolites A, B, C, D, E and F through intracellular reactions. *O* represents the flux through the metabolic demand reaction, which represents the objective function; e_1, e_2, e_3 and e_4 represent the transport/exchange fluxes that either consume metabolite from environment or secrete metabolite into the environment; v_1, v_2, v_3 and v_4 represent the intracellular reaction fluxes which are assumed to be catalysed by specific metabolic enzymes. The green coloured arrows represent the pathway of uptake and utilization of metabolite A through multiple reactions to satisfy the objective of maximizing metabolic demand; (**b**) decomposition of the rate of change of metabolites within the network into the *S* (stoichiometry) and *v* (flux) components. The stoichiometric matrix containing the participation of *m* metabolites (rows) in *n* reactions (columns) of the network

n metabolites. Each element in *Sij* represents the stoichiometric coefficient of metabolite *i* in reaction *j*. The coefficients are positive, if metabolite *i* is produced in reaction *j* and negative, if it is consumed. With respect to the metabolic information known for *Leishmania*, flux-based constraints can be applied to the reactions of the network. Apart from the applied exchange constraints, the inherent stoichiometry and reversibility of the reactions within the network and subcellular compartmentalization provide an important constraint to distribution of fluxes through reactions. A toy model and its stoichiometric representation are given in Fig. 8.5.

The rate of change of concentration of every metabolite in a metabolic reconstruction can be represented as a function of individual reaction fluxes:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = S.v \tag{8.24}$$

where M = vector of metabolite concentrations; S = stoichiometric matrix of m rows of metabolites and n columns of reactions; v = reaction flux vector for n reactions (refer to Fig. 8.5 for an example).

Assuming the system to function at steady state, from Eq. (8.24),

$$\frac{\mathrm{d}M}{\mathrm{d}t} = 0 \quad \text{and} \quad S.v = 0 \tag{8.25}$$

As the nature of this system is underdetermined, FBA uses linear programming (LP) techniques to solve the above system of equations and identify a solution vector

(flux distribution) in a particular constrained situation that would optimize a specific cellular objective. Note that each reaction flux v_i in the flux vector is constrained between bounds a_i and b_i such that, the LP optimization problem formulated for performing FBA was given by

$$\text{Maximize } O, \tag{8.26}$$

subject to constraints $S.v = 0$ and $a_i \leq v_i \leq b_i$,
where O = objective function to be maximized and $O = c_1v_1 + c_2v_2 + \ldots + c_nv_n$ where $c_1, c_2, \ldots, c_n$ are stoichiometric coefficients of metabolites in demand reaction and $v_1, v_2, \ldots v_n$ are the fluxes of model reactions that maximize the objective a_i = lower bound of flux through every reaction i in the model and b_i = upper bound of flux through every reaction i in the model.

8.4.6 *The iAS556 Genome-Scale Reconstruction of* Leishmania infantum *JPCM5 Metabolism*

Using the above strategy, the iAS556 *Leishmania infantum* JPCM5 metabolic network was reconstructed (Subramanian and Sarkar 2017). The network consisted of 556 genes, 1260 reactions and 1160 metabolites occurring in 8 subcellular compartments. Subjecting this reconstruction to flux-based analysis, the metabolic organization in *Leishmania* is adapted to utilize different metabolite alternatives, depending upon the redox balance constraints within compartments generated for different combinations of input metabolites. The observations indicated that inherent stoichiometry and reversibility coerced by occurrence of enzymes in multiple subcellular locations are itself enough to explain changes between the metabolic states specific to different developmental stages of the parasite. Flux coupling between reactions formed a functional subnetwork, which remained hierarchical and modular in organization ensuring unhindered production of biomass metabolites while remaining robust to accidental errors thus supporting the parasite to achieve optimal survival.

8.5 Summary

The above chapter provides a detailed understanding of the mathematical and computational principles in understanding large-scale, heterogeneous biological data and also provides an overview of their explored applications. The above methodology-oriented discussion substantiates the strength of mathematical models to detect patterns occurring within the genotype and phenotype so as to explain the relationship between the discrete components within and across different levels of

biological organization. Although many of these techniques seem to be strong in principle, they still have not become popular in use. Once standardized for a large number of datasets, these methods can be integrated into a simplified computer package that can provide the user with an ability to switch between methods to predict the unknown relationships between the genotype and phenotype.

References

Almeida JS, Carrico JA, Maretzek A et al (2001) Analysis of genomic sequences by Chaos Game Representation. Bioinformatics 17:429–437

Alonso G, Guevara P, Ramirez JL (1992) Trypanosomatidae codon usage and GC distribution. Mem Inst Oswaldo Cruz 87:517–523

Apweiler R, Martin MJ, O'Donovan C et al (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res 41:D43–D47

Arakawa K, Oshita K, Tomita M (2009) A web server for interactive and zoomable Chaos Game Representation images. Source Code Biol Med 4:6

Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28:304–305

Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2008) GenBank. Nucleic Acids Res 36:D25–D30

Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2010) GenBank. Nucleic Acids Res 38:D46–D51. https://doi.org/10.1093/nar/gkp1024

Bielińska-Wąż D, Wąż P (2017) Spectral-dynamic representation of DNA sequences. J Biomed Inform 72:1–7

Borst P (1986) How proteins get into microbodies (peroxisomes, glyoxysomes, glycosomes). Biochim Biophys Acta (BBA)-Gene Struct Expr 866:179–203

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA1. J Mol Biol 268:78–94

Consortium U et al (2014) Activities at the universal protein resource (UniProt). Nucleic Acids Res 42:D191–D198

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

Dai Q, Liu X-Q, Wang T-M, Vukicevic D (2007) Linear regression model of DNA sequences and its application. J Comput Chem 28:1434–1445

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23:324–328

De Castro E, Sigrist CJA, Gattiker A et al (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34:W362–W365

Dror G, Sorek R, Shamir R (2005) Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics 21:897–901. https://doi.org/10.1093/bioinformatics/bti132

Eddy SR (1996) Hidden markov models. Curr Opin Struct Biol 6:361–365

Eddy SR (2001) HMMER: profile hidden Markov models for biological sequence analysis

Emanuelsson O, Brunak S, Von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90. https://doi.org/10.1038/47056

Fertil B, Massin M, Lespinats S et al (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. Nucleic Acids Res 33:W512–W515

Finn RD, Bateman A, Clements J et al (2013) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Fridolin G, Green S (2017) The sum of the parts: large-scale modeling in systems biology. Philos Theory Biol 9:1–26
Gasteiger E, Hoogland C, Gattiker A et al (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook. Humana press, Totowa, pp 571–607
Glunčić M, Paar V (2012) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Res 41:e17–e17
Guerra-Giraldez C, Quijada L, Clayton CE (2002) Compartmentation of enzymes in a microbody, the glycosome, is essential in Trypanosoma brucei. J Cell Sci 115:2651–2658
Henry CS, Overbeek R, Xia F et al (2011) Connecting genotype to phenotype in the era of high-throughput sequencing. Biochim Biophys Acta (BBA) Gen Subj 1810:967–977
Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287–299
Hou W, Pan Q, He M (2016) A new graphical representation of protein sequences and its applications. Phys A Stat Mech Appl 444:996–1002
Huh W-K, Falvo JV, Gerke LC et al (2003) Global analysis of protein localization in budding yeast. Nature 425:686
Huynen M (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10:1204–1210. https://doi.org/10.1101/gr.10.8.1204
Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Res 18:2163–2170
Jensen LJ, Kuhn M, Stark M et al (2008) STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37:D412–D416
Joseph J, Sasikumar R (2006) Chaos game representation for comparison of whole genomes. BMC Bioinforma 7:243
Kanehisa M, Goto S, Sato Y et al (2013) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–D205
Lin J (1991) Divergence measures based on the Shannon entropy. Inf Theory IEEE Trans 37:145–151
Lu YY, Tang K, Ren J et al (2017) CAFE: accelerated alignment-free sequence analysis. Nucleic Acids Res 45:W554–W559. https://doi.org/10.1093/nar/gkx351
Michels PAM (1988) Compartmentation of glycolysis in trypanosomes: a potential target for new trypanocidal drugs. Biol Cell 64:157–164
Misset O, Bos OJM, Opperdoes FR (1986) Glycolytic enzymes of Trypanosoma brucei. Eur J Biochem 157:441–453
Mostafavi S, Ray D, Warde-Farley D et al (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol 9:S4
Mount DW (2004) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press, New York
Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
Opperdoes FR, Szikora J-P (2006) In silico prediction of the glycosomal enzymes of Leishmania major and trypanosomes. Mol Biochem Parasitol 147:193–206
Orgogozo V, Morizot B, Martin A (2015) The differential view of genotype – phenotype relationships. Front Genet 6:179
Overbeek R, Fonstein M, D'Souza M et al (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896–2901
Pandit A, Sinha S (2010) Using genomic signatures for HIV-1 sub-typing. BMC Bioinforma 11:S26
Pellegrini M, Marcotte EM, Thompson MJ et al (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285–4288
Plotkin JB, Kudla G (2010) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12:32–42

Qi Z-H, Jin M-Z (2016) An intuitive graphical method for visualizing protein sequences based on linear regression and physicochemical properties. Match Commun Math Comput Chem 75:463–480
Qi J, Luo H, Hao B (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res 32:W45–W47
Qi X, Wu Q, Zhang Y et al (2011) A novel model for DNA sequence similarity analysis based on graph theory. Evol Bioinforma 7:EBO–S7364
Randić M, Vracko M, Nandy A, Basak SC (2000) On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inf Comput Sci 40:1235–1244
Randić M, Vračko M, Lerš N, Plavšić D (2003a) Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett 368:1–6
Randić M, Vračko M, Lerš N, Plavšić D (2003b) Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem Phys Lett 371:202–207
Randić M, Zupan J, Balaban AT (2004) Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem Phys Lett 397:247–252
Ren Q, Chen K, Paulsen IT (2006) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 35:D274–D279
Rice P, Longden I, Bleasby A et al (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277
Schomburg I, Chang A, Placzek S et al (2012) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res 41:D764–D772 gks1049
Sharma D, Issac B, Raghava GPS, Ramaswamy R (2004) Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics 20:1405–1412
Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295
Smith TF, Waterman MS (1981) Comparison of biosequences. Adv Appl Math 2:482–489
Snoep JL, Westerhoff HV (2005) From isolation to integration, a systems biology approach for building the Silicon Cell. In: Alberghina L, Westerhoff HV (eds) Systems biology: definitions and perspectives. Springer Berlin Heidelberg, Berlin, pp 13–30
Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:ii215–ii225
Subramanian A, Sarkar RR (2015) Comparison of codon usage bias across Leishmania and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions. Genomics 106:232–241
Subramanian A, Sarkar RR (2017) Revealing the mystery of metabolic adaptations using a genome scale model of Leishmania infantum. Sci Rep 7:10262. https://doi.org/10.1038/s41598-017-10743-x
Subramanian A, Jhawar J, Sarkar RR (2015) Dissecting Leishmania infantum energy metabolism – a systems perspective. PLoS One 10:e0137976. https://doi.org/10.1371/journal.pone.0137976
Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci 85:2653–2657
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
Tsonis AA, Elsner JB, Tsonis PA (1991) Periodicity in DNA coding sequences: implications in gene evolution. J Theor Biol 151:323–331
Vinga S, Almeida J (2003) Alignment-free sequence comparison—a review. Bioinformatics 19:513–523
Wang Y, Hill K, Singh S, Kari L (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene 346:173–185
Wang S-Y, Tian F-C, Liu X, Wang J (2009a) A novel representation approach to DNA sequence and its application. IEEE Signal Process Lett 16:275–278

Wang S, Tian F, Feng W, Liu X (2009b) Applications of representation method for DNA sequences based on symbolic dynamics. J Mol Struct THEOCHEM 909:33–42

Wang Y, Tang H, DeBarry JD et al (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49–e49

van Weelden SWH, van Hellemond JJ, Opperdoes FR, Tielens AGM (2005) New functions for parts of the Krebs cycle in procyclic Trypanosoma brucei, a cycle not operating as a cycle. J Biol Chem 280:12451–12460

Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 11:356–372. https://doi.org/10.1101/gr.161901

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29

Yang ZR (2004) Biological applications of support vector machines. Brief Bioinform 5:328–338

Yang X, Wang T (2013) Linear regression model of short k-word: a similarity distance suitable for biological sequences with various lengths. J Theor Biol 337:61–70

Yao Y-H, Dai Q, Li C et al (2008) Analysis of similarity/dissimilarity of protein sequences. Protein Struct Funct Bioinforma 73:864–871

Yao Y, Yan S, Han J et al (2014) A novel descriptor of protein sequences and its application. J Theor Biol 347:109–117

Yuan C, Liao B, Wang T (2003) New 3D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett 379:412–417

Zielezinski A, Vinga S, Almeida J, Karlowski WM (2017) Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol 18:186

Chapter 9
Alignment-Free Analyses of Nucleic Acid Sequences Using Graphical Representation (with Special Reference to Pandemic Bird Flu and Swine Flu)

Ashesh Nandy, Antara De, Proyasha Roy, Munna Dutta, Moumita Roy, Dwaipayan Sen, and Subhash C. Basak

Abstract The exponential growth in database of bio-molecular sequences have spawned many approaches towards storage, retrieval, classification and analyses requirements. Alignment-free techniques such as graphical representations and numerical characterisation (GRANCH) methods have enabled some detailed analyses of large sequences and found a number of different applications in the eukaryotic and prokaryotic domain. In particular, recalling the history of pandemic influenza in brief, we have followed the progress of viral infections such as bird flu of 1997 onwards and determined that the virus can spread conserved over space and time, that influenza virus can undergo fairly conspicuous recombination-like events in segmented genes, that certain segments of the neuraminidase and hemagglutinin surface proteins remain conserved and can be targeted for peptide vaccines. We recount in some detail a few of the representative GRANCH techniques to provide a glimpse of how these methods are used in formulating quantitative sequence descriptors to analyse DNA, RNA and protein sequences to derive meaningful results. Finally, we survey the surveillance techniques with a special reference to how the GRANCH techniques can be used for the purpose and recount the forecasts made of possible metamorphosis of pandemic bird flu to pandemic human infecting agents.

A. Nandy · A. De · P. Roy · D. Sen
Centre for Interdisciplinary Research and Education, Kolkata, India

M. Dutta · M. Roy
Dinabandhu Andrews College, Kolkata, India

S. C. Basak (✉)
Department of Chemistry and Biochemistry, University of Minnesota Duluth, Duluth, MN, USA
e-mail: sbasak@nrri.umn.edu

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_9

Keywords DNA/RNA sequence · BLAST · Numerical characterisations of DNA/RNA/protein sequences · Graphical representation and numerical characterisation (GRANCH) · Alignment-free graphical representation methods · Genes and genomic sequences · Molecular sequence similarity/dissimilarity · Beta-globin sequences · Mathematical descriptors · Graph theory · Graph-theoretical (topological) distance · Euclidean distance · Adjacency matrix · Distance matrix · D_E/D_G matrix · Graph invariant · L/L matrix · M/M matrix · 3D plot · Protein sequences · Spanish influenza · H1N1 flu virus · Pandemics · H5N1 avian flu · Negative-sense strand RNA virus · Hemagglutinin (HA) · Neuraminidase (NA) · Sialic acid residue · Antigenic drift · Antigenic shift · Mutations through recombination · Flu vaccines · Surveillance of flu · Ethics in surveillance

9.1 Introduction

Since the availability of large nucleic acid sequence data in the late 1980s, followed by the subsequent exponential growth of such data in the databases, the need for supportive tools to store, view, retrieve and analyse such sequences became and continues to be of urgent necessity. These tasks require methods to characterise the information contained in these sequences with a view to compare and contrast different sequences to determine their functions and place in the biological domain. The symbolic DNA sequence representation has been the legacy mainstay of such analyses carried over from the oligonucleotide sequence days, falling far short of the requirements of the broad sweeping overview of local and global base distributions, multiple sequence similarities and dissimilarities, molecular evolution and other studies necessary for an overarching assessment of the relationship of any DNA/RNA sequence to the biome through the accumulated data in the nucleic acid databases.

Sequence alignment has been the primary tool of choice for such comparison of generic sequences. However, in the conventional methods, the time complexity of sequence alignment varies as $\sim O(l^2)$ for two sequence segments of length l (Kobori and Mizuta 2015). In the realm of genomic sequences, this clearly consumes an enormous amount of computation time, which grows even worse when multiple alignments are necessary. The reason behind such sequence alignment is the fact that mutational changes in nucleic acid sequences through individual nucleotide mutations and rearrangements in nucleotide distributions brought on by various biological processes are reflected in the DNA/RNA sequences we observe today. This constitutes one of the main considerations facing biologists trying to understand the variations between different sequences of the same genera and their evolution (Qi et al. 2011).

Numerical characterisation of the DNA/RNA sequences is one of the main tools in undertaking this task, and a number of methods were developed to analyse specific attributes of DNA sequences. For example, the dot plot (Gibbs and McIntyre 1970) is used to visualise identical sequence segments between two proteins by the

diagonal runs in a matrix of the constituents of the two sequences. Needleman and Wunsch (1970) used dynamic programming techniques to align protein or nucleotide sequences; another dynamic programming scheme was later proposed by Smith and Waterman (1981) (Smith et al. 1985), for improved scoring local sequence alignment, but the computational complexity and time requirements have limited its use. For protein secondary structure prediction, Chou and Fasman (1974a, b) developed an empirical technique, which, however, has been superseded by new machine learning techniques. In particular, dynamic programming techniques that analyse DNA sequences in terms of discrete small segments to eventually merge into the full sequence require time commitments of $\sim O(l^k)$, where k is the number of sequences to be aligned and is expensive in the use of computing resources, whereas fast Fourier transform (FFT) requiring resource time of $\sim O(l \log l)$ is only slightly better (Hanson, thesis 2003). Multiple alignment fast Fourier transform (MAFFT) with a simplified scoring system to reduce the computational overhead (Katoh et al. 2002), discrete Fourier transforms (Yin et al. 2014) and wavelet transforms (Dodin et al. 2000) that displays regularity in patterns in DNA sequences have been some of the other methods deployed to extraction of information of interest from the DNA databanks. BLAST (Basic Local Alignment Search Tool) is one of the most popular search tools for local alignments; it uses a heuristic algorithm that takes shortcuts to identify, very rapidly, sequences in a library that most closely resemble the query sequence (Altschul et al. 1990, 1997).

A more general approach that permits local and global sequence comparisons, generation of phylogenetic trees, numerical characterisations of DNA/RNA/protein sequences, quantitative estimation of sequence divergences and other properties and rational design of peptide vaccines is offered by novel, alignment-free schemes of graphical representation and numerical characterisation of biomolecular sequences (see review Nandy et al. 2006). The basic procedure in all these methods is to identify a vector with each type of nucleotide and plot each vector successively until the end of the sequence, thus drawing out a trajectory in the space of the model; associating a numerical value with the curve results in a descriptor of the sequence. This allows for quantitative comparisons between sequences and forms a powerful tool for analysis of similarities and dissimilarities between genes on a local and global scale. These novel approaches dispense with the need for alignments, which, for large sequences, become less reliable due to divergence of species and subsequent rearrangements (e.g. reversal, transposition or block exchange) occurring over evolutionary time scales; corrections introduced to take care of these divergences depend upon significant homologies between the sequences and the genes to be compared and thus become restrictive (Qi et al. 2011).

However, as we discuss later, these alignment-free graphical representation methods are not without their problems either. One of these is related to computation of the sequence descriptors. In geometric approaches (Raychaudhury and Nandy 1999; Liao and Ding 2006; Bielinska-Waz et al. 2007), computational closure in limited time is possible; however, in some of the approaches where sequence descriptors are defined through matrices of distances between the bases or amino acids of the sequence (Randic et al. 2000a, b; Song and Tang 2005; Wang and Zhang

2006; Li et al. 2016; Randic et al. 2004; Bai and Wang 2006), generating a final value for large sequences remains a significant problem, and approximations have to be resorted too. In some instances, defining a suitable matrix to characterise and an invariant to associate with a gene sequence presents some difficulties (Qi et al. 2011).

In this chapter, we present a concise summary of a selection of graphical representation methods to demonstrate the basics of this class of approaches of numerically characterising a biomolecular sequence. There have been a very large number of methodologies proposed to quantitatively estimate the differences between selected groups of genes and genomic sequences (see review Nandy et al. 2006). Extensions of these approaches have been made to take into account protein sequences (see review Randic et al. 2011) where the problem is compounded by the fact that the basic building blocks in protein sequences are constituted by 20 amino acids (Nandy et al. 2009) in comparison to nucleotide sequences where the basic building blocks constitute four bases; in terms of the methodologies, we consider a few in a short section and refer the readers to the original papers for greater details.

These graphical representation methods are found ready for acceptance in research. The visual representation of base distribution in DNA sequences was used by Larionov et al. (2008) to compare the chromosomal DNAs of human and mouse to find several examples of palindromic runs in almost the same manner in the two species. Wiesner and Wiesnerova (2010), in an innovative application of 2D graphical techniques to multiallelic marker loci from *Begonia* x *tuberhybrida* Voss, suggested that the DNA walks could be correlated with genetic diversity to predict new allele-rich loci and improve new DNA germplasm identificators. Applications of the graphical representation and numerical characterisation techniques were made by Liao and his group (Liao et al. 2005, 2006) to analyse the genomic sequences of the viruses in SARS (severe acute respiratory syndrome) pandemic of 2003 and generate their phylogenetic relationship tree. Humberto Gonzalez-Diaz and his group applied the underlying concepts in an analysis of protein sequences for drug discovery research through numerical parameters analogously to quantitative structure-activity relationship (QSAR) topological indices (Estrada and Uriarte 2001; González-Díaz et al. 2007), determination of mass spectral data of proteins, toxicoproteomics, toxicity prediction and diagnosis of cancer patients (Cruz-Monteagudo et al. 2008; Gonzalez-Diaz et al. 2008a, b; Aguero-Chapin et al. 2009).

In a number of interesting applications of the original 2D graphical representation of DNA/RNA sequences, one of the current authors of this chapter showed that introns and exons of mammalian genes differed significantly in base distribution characteristics (Nandy 1996a), that these properties can be used to determine coding regions in newly identified DNA sequences (Nandy 1996b) and that mutational changes in genetic sequences are guided by some restrictions which indicate an intricate relationship between intra-purine and intra-pyrimidine content of the sequences (Nandy 2009), among others.

A number of applications have been made over the years of some of these graphical representation and numerical characterisation methods, molecular sequence similarity/dissimilarity and phylogenetic relationships being among the

ones with the most abiding interest. For example, Randic et al. (2000a, b) and Liao and Ding (2006) used a novel 3D graphical representation model to examine the base distributions in the first exons of beta-globin genes of 11 species giving results that matched closely with those obtained by other graphical representation methods. In another approach, Zhang et al. (2010) used a spectral representation of DNA sequences defining a 24-component vector as sequence descriptor and illustrated their method using complete beta-globin sequences of seven species. Li et al. (2016) introduced a novel method of considering dinucleotides in a 3D graphical representation model and obtained phylogenetic trees of 4 datasets: the coding sequences of the beta-globin genes of 18 species, the mitochondrial cytochrome oxidase subunit I (COI) genes of 9 butterflies, the S segments of 32 hantaviruses (HVs) and 70 complete mitochondrial genomes, all of which gave evolutionary relationships that matched with standard data.

In a slightly different application area, we refer in some detail to the applications of graphical representations and numerical characterisation schemes to viral genetics, focusing our attention more on the influenza virus in view of the several pandemics that it has caused, but we do mention in passing other viral pandemics also. It will be seen that some of the issues related to vaccine design and drug discovery are more easily and readily approached using graphical techniques, and these approaches with their visual representations also give fresh insights into base distribution and composition characteristics. These have led to determination of vaccine targets in the virions, tracking viral progress across geographical and time boundaries, computation of phylogenetic relationships and possible regulation of base composition and distribution in DNA/RNA sequences. Brief discussions of these applications and methods of surveillance of viral developments and progress are covered in another section of this chapter.

9.2 Graphical Representation and Numerical Characterisation (GRANCH)

Graphical representation and numerical characterisation (GRANCH) of nucleotide and amino acid sequences provides a very powerful tool for basic biological research. Mathematical descriptors based on graphical representations are used as a compact method for sequence comparison. The novelty, utility and low complexity of these representations have spawned a wide variety of approaches, some of which have seen many applications. A selection of such methods is given here in order to elucidate the many ways in which this technique can be formulated.

While most applications to date have been done in graphical representations in 2D and 3D spaces, several authors (e.g. Tang et al. 2010) have proposed representations in higher orders of dimensionality. However, there are some advantages and disadvantages to them; in a 4D graphical representation of nucleic acid sequences, while overlaps and intersections of the sequence curve with itself can be avoided,

graphical visualisation and the ability to directly compare two sequences are lost which, on the other hand, are easily achieved in 2D or 3D plots. Liao and Wang (2004) proposed a 6D representation, while Randic et al. (2005) suggested a novel four-colour map representation.

Although graphical representation provides visual clues to the characteristics of base distribution in a sequence and enables visual comparison of similarities and dissimilarities between two or more sequences, quantitative estimation of such characteristics or differences between sequences is required for any significant comparisons. All authors of graphical representations have, therefore, recommended sequence descriptors to quantify sequence similarity/dissimilarity through the use of such descriptors. There are two approaches to define mathematical descriptors: geometrical and graph-theoretical methods.

9.2.1 Quantitative Estimation Methods

Various numerical descriptors for the numerical characterisation of sequences have been formulated during the past few decades. A particular descriptor maps the set of sequences (S) into the set of real numbers (R). In some cases, a vector may be extracted instead of a single number. Given below is a representative sample of the important and widely used descriptors known thus far. However, it should be noted that the list is not exhaustive.

9.2.1.1 Geometrical Methods

This method was first described by Raychaudhury and Nandy (1999), where they used the graphical representation of DNA sequence on a 2D rectangular grid, as explained below, using the (x, y) coordinate to derive the descriptor values. The first-order moments (μ_x, μ_y) and a graph radius g_R are defined for each sequence as:

$$\mu_x = \frac{\sum x_i}{N}, \mu_y = \frac{\sum y_i}{N} \text{ and } g_R = \sqrt{{\mu_x}^2 + {\mu_y}^2}$$

where (x_i, y_i) are the coordinates of each point on the plot and N is the total number of bases in the segment. g_R is the base distribution index which is dependent on the specific positions of the bases in a given sequence and has been found to be characteristic of the specific sequence base distributions. Thus, if two sequences yield the same g_R value, the two sequences will be found to be identical for all practical purposes (Nandy and Nandy 2003). Now, if μ_1 and μ_2 refer to two different DNA sequences, then the quantity Δg_R defined by:

$$\Delta g_R = \sqrt{(\mu_{1x} - \mu_{2x})^2 + (\mu_{1y} - \mu_{2y})^2}$$

provides an estimation of the difference between the two sequences. Thus, g_R and Δg_R are very important measures of sequence composition and distribution.

9.2.1.2 Graph-Theoretical Methods

In this method, DNA/RNA sequence is represented and characterised by a two-step process: representation of the sequences using graphs and characterisation of the graphs by graph invariants. A graph invariant is a graph-theoretic property which is preserved by isomorphic graph (Harary 1969; Janežič et al. 2007). A graph G is defined as an ordered pair consisting of two sets, V and $G=(V(G), R)$, where $V(G)$ represents a finite nonempty set of points and R is a binary relation defined on the set $V(G)$. The elements of V are called vertices, and the elements of R also symbolised by $E(G)$ or E are called edges. Such an abstract graph is commonly visualised by representing elements of $V(G)$ as points and by connecting each pair (u, v) of elements of $V(G)$ with a line or edge if and only if $(u, v) \epsilon R$. The vertex, v, and edge, e, are incident with each other, as are u and e. Two vertices in G are called adjacent if $(u, v) \epsilon R$, i.e. they are connected by an edge. A walk of a graph is a sequence beginning and ending with vertices in which vertices and edges alternate and each edge is incident with vertices immediately preceding and following it. A walk of the form $v_0, e_1, v_1, e_2, \ldots, v_n$ joins vertices v_0 and v_n. The length of a walk is the number of edges in the walk. A graph G is connected if every pair of its vertices is connected by a path. The distance (u, v) between vertices u and v in G is the length of the shortest path connecting u and v.

In a molecular graph corresponding to DNA/RNA sequences, V represents the set of nucleic acid bases (A, T/U, G, C) and E represents the set of bonds connecting the adjacent bases. For any pair of bases (i, j) in the sequence, $(i, j) \epsilon R$, they are either connected (adjacent) or not. Such a graph may be represented by an adjacency matrix $A = \{\alpha_{ij}\}$ where:

$$a_{ij} \begin{cases} = 1 & \text{if } i,j \text{ are connected} \\ = 0 & \text{otherwise} \end{cases}$$

Now the graph-theoretical distance (topological) D_G and the Euclidean distance D_E between the vertices are measured, and a particular type of matrix, D_E/D_G, is formed from which the leading eigenvalues are calculated. The leading eigenvalue is used to characterise a sequence, while the difference between the eigenvalues forms an estimate of the similarity/dissimilarity between the sequences. Thus, the leading eigenvalues of D_E/D_G matrix and the associated eigen matrices are considered to be descriptors of the DNA sequences.

To clarify the process of graphical representations and the differences between the multiple approaches, a short sequence of 10 bases, ATGAACACTC, is used as a

standard example to generate plots using each of the methods mentioned here. The sequence denotes the stretch of first ten bases of the fourth segment of the influenza genome and the hemagglutinin (HA) gene, belonging to influenza A virus (A/Shenzhen/SP139/2014(H7N9)).

9.2.2 Graphical Representation and Numerical Characterisation Methods for Nucleotide Sequences

9.2.2.1 2D Graphical Representations

2D Rectangular Plot

A simple 2D Cartesian coordinate system is used to represent the nucleic acid sequences. Gates (1986), Nandy (1994) and Leong and Morgenthaler (1995) independently proposed the four cardinal directions on a 2D grid to represent the four bases. According to Nandy, starting from the origin, a point is plotted by moving one step in negative x-direction if the base is an adenine (A) and in the opposite direction if the base is a guanine (G) and a walk of one step in positive y-direction if the base is a cytosine (C) and in the opposite direction if it is a thymine (T) (or uracil, U). Plotting the points successively in this manner draws a graph of the sequence of bases. Gates and Leong and Morgenthaler proposed a similar graphical representation but with different assignments of the bases to the cardinal directions. The graph according to the Nandy prescription is given in Fig. 9.1.

Table 9.1 represents the coordinates, the centre of mass of all the points and the graph radius g_R of the sequence as described in Sect. 9.2.1.1 (Raychaudhury and Nandy 1999).

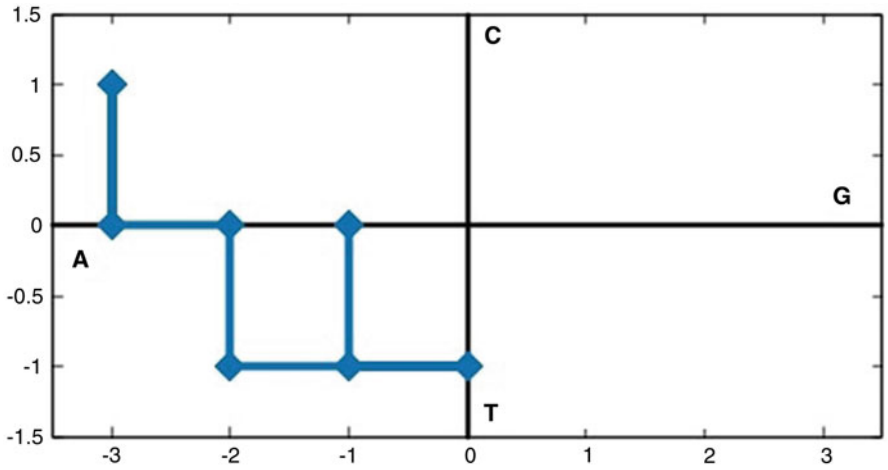

Fig. 9.1 Nandy plot of ATGAACACTC

Table 9.1 The co-ordinates, the centre of mass of all the points, and the graph radius g_R of the sequence as per Nandy rectangular plot

Base	Coordinates		Centre of mass		
	X	Y	μ_x	μ_y	g_R
A	−1	0	−1.9	−0.2	1.91
T	−1	−1			
G	0	−1			
A	−1	−1			
A	−2	−1			
C	−2	0			
A	−3	0			
C	−3	1			
T	−3	0			
C	−3	1			

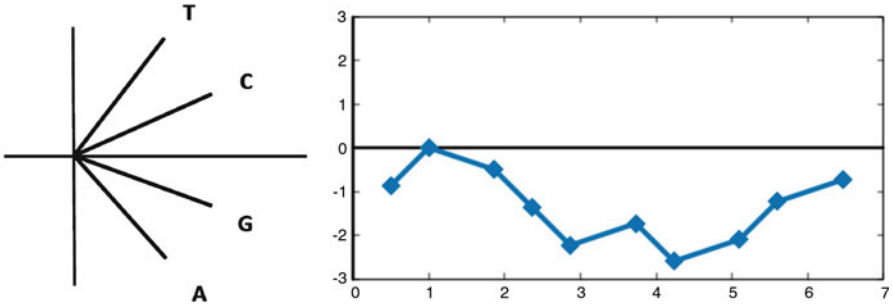

Fig. 9.2 Yau plot coordinate system and the graph of the sample sequence ATGAACACTC

Yau Plot

Although the 2D rectangular plots use a simple technique to easily visualise the sequence, it involves visual degeneracy that leads to an apparent loss of genetic information where repeated sequences, such as GAGAGAGAGAG, show overlapping paths. To overcome this, several authors proposed different modifications to the method which would reduce degeneracy (see review Nandy et al. 2006). Yau et al. (2003) proposed using two quadrants instead of four; cytosine (C) and thymine (T) are assigned to the first quadrant and adenine (A) and guanine (G) to the second quadrant, with the coordinates of the four bases defined as $(1/2, -\sqrt{3}/2)$ for A, $(\sqrt{3}/2, -1/2)$ for G, $(\sqrt{3}/2, 1/2)$ for C and $(1/2, \sqrt{3}/2)$ for T (Fig. 9.2).

While they did not prescribe a precise method for computing descriptors for the plots arising from their graphical method, we can, analogously to the Raychaudhury and Nandy (1999) method, define a centre of mass and a graph radius as descriptors. The results for our sequence in the Yau et al. (2006) plot are given in Table 9.2.

Table 9.2 The co-ordinates, centre of mass of all points and the graph radius as per the Yau et al plot

Base	Coordinates		Centre of mass		
	X	Y	μ_x	μ_y	g_R
A	0.5	−0.866	3.372	−1.336	3.627
T	1	0			
G	1.866	−0.5			
A	2.366	−1.366			
A	2.866	−2.232			
C	3.732	−1.732			
A	4.232	−2.598			
C	5.098	−2.098			
T	5.598	−1.232			
C	6.464	−0.732			

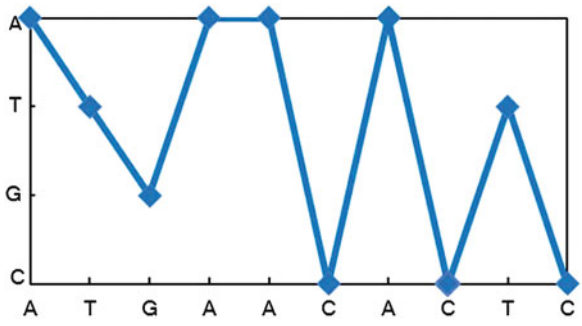

Fig. 9.3 Randic 2D plot for the sequence ATGAACACTC

Randic 2D Plot

There are many 2D approaches which do not involve the Cartesian coordinate system to represent nucleic acid sequence. Randic et al. (2003) proposed four horizontal equidistant parallel lines labelled as A, T, G and C from top to bottom. To plot the nucleic acid sequence, the bases are labelled along the x-axis, and straight lines are drawn from individual points on the four parallel lines according to the bases occurring in the sequence (Fig. 9.3).

A matrix method is used for determining a descriptor by constructing an M/M matrix for the given sequence ATGAACACTC. The off-diagonal entries of the M/M matrix are determined by dividing the Euclidean distance between two vertices of the zigzag curve by the graph-theoretical distance, i.e. the number of edges, between the two vertices. Here, if we take the first base A and the sixth base C of our sequence as an example, then the corresponding matrix element is obtained by taking the ratio of Euclidian distance between the first base (A) and the sixth base (C), i.e. $\sqrt{34}$, to the number of edges between A and C, which is 5 here. Proceeding in this way, we obtain the M/M matrix elements as shown below and the leading eigenvalue as the descriptor of the sequence (Tables 9.3, 9.4 and 9.5).

The leading eigenvalue in this case turns out to be 12.2388.

Table 9.3 The D_E matrix in Randic 2D model

	A	T	G	A	A	C	A	C	T	C
A	0	1.4142	2.8284	3	4	5.831	6	7.6158	8.0623	9.4868
T		0	1.4142	2.2361	3.1623	4.4721	5.099	6.3246	7	8.2462
G			0	2.2361	2.8284	3.1623	4.4721	5.099	6.0828	7.0712
A				0	1	3.6056	3	5	5.099	6.7082
A					0	3.1623	2	4.2426	4.1231	5.831
C						0	3.1623	2	3.6056	4
A							0	3.1623	2.2361	4.2426
C								0	2.2361	2
T									0	2.2361
C										0

Table 9.4 The D_G matrix in Randic 2D model

	A	T	G	A	A	C	A	C	T	C
A	0	1	2	3	4	5	6	7	8	9
T		0	1	2	3	4	5	6	7	8
G			0	1	2	3	4	5	6	7
A				0	1	2	3	4	5	6
A					0	1	2	3	4	5
C						0	1	2	3	4
A							0	1	2	3
C								0	1	2
T									0	1
C										0

Table 9.5 The M/M matrix in Randic 2D model

	A	T	G	A	A	C	A	C	T	C
A	0	1.4142	1.4142	1.0000	1.0000	1.1662	1.0000	1.0880	1.0078	1.0541
T		0	1.4142	1.1181	1.0541	1.1180	1.0198	1.0541	1.0000	1.0308
G			0	2.2361	1.4142	1.0541	1.1180	1.0198	1.0138	1.0102
A				0	1.0000	1.8028	1.0000	1.2500	1.0198	1.1180
A					0	3.1623	1.0000	1.4142	1.0308	1.1662
C						0	3.1623	1.0000	1.2019	1.0000
A							0	3.1623	1.1181	1.4142
C								0	2.2361	1.0000
T									0	2.2361
C										0

Song-Tang Plot

In a slight variation, Song and Tang (2005) proposed a three-horizontal line method where the central line represents two bases and the two peripheral lines denote each of the remaining bases. Counting of the nucleotides is done along the x-axis, and the

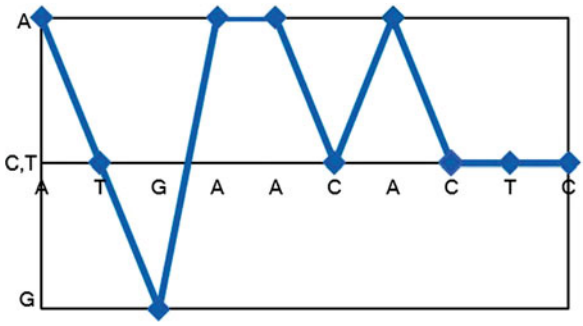

Fig. 9.4 One plot as per Song and Tang's representation of the sequence ATGAACACTC

bases are plotted and connected by straight lines as in the case of the Randic 2D plot. In the case of DNA primary sequences, the four bases A, C, G and T can be grouped as per their characteristics:

- Purine R = (A, G) and pyrimidine Y = (C, T)
- Amino M = (A, C) and keto K = (G, T)
- Weak H-bond W = (A, T) and strong H-bond S = (C, G)

Keeping one of such groups, such as purine, as central line, and cytosine (C) and thymine (T) as the peripheral lines, we can plot a graph. Repeating the same process for the other groups, we get a total of six ($^4C_2$) combinations of plots. For each of them, the two peripheral lines can also be exchanged. In total, 12 (6×2) plots can be obtained.

In one such graph, where cytosine (C) and thymine (T) are assigned to the central line, adenine (A) to the uppermost line and guanine (G) to the lowermost line, the plot for the sequence ATGAACACTC is shown in Fig. 9.4.

Here, the matrix elements and the descriptor are obtained by using the Randic 2D approach as discussed earlier. Like in the Randic 2D plot, two new matrix relations, as defined below, M/M and L/L, are constructed. The leading eigenvalues are calculated from the M/M and L/L matrices which form descriptors of the corresponding DNA sequence. The matrix elements are formed as follows:

$$\begin{aligned}(\mathrm{ED})_{\mathrm{ij}} &= \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}\\ (M/M)_{\mathrm{ij}} &= (\mathrm{ED})_{\mathrm{ij}}/(\mathrm{GD})_{\mathrm{ij}} i \neq j\\ (M/M)_{\mathrm{ij}} &= 0\end{aligned}$$

$$\begin{aligned}(\mathrm{PD})_{\mathrm{ji}} = (\mathrm{PD})_{\mathrm{ij}} &= (\mathrm{ED})_{\mathrm{ii+1}} + (\mathrm{ED})_{i+1,i+2} + \ldots, + (\mathrm{ED})_{j-1,j} i < j\\ (L/L)_{\mathrm{ij}} &= (\mathrm{ED})_{\mathrm{ij}}/(\mathrm{PD})_{\mathrm{ij}} \quad i \neq j\\ (L/L)_{\mathrm{ij}} &= 0\end{aligned}$$

where ED, GD and PD are the Euclidian distance matrix, graph-theoretical distance matrix and path distance matrix, respectively (Tables 9.6 and 9.7).

The leading eigenvalue value in this case turns out to be 10.0552.

Table 9.6 M/M matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1.4142	1.4142	1.0000	1.0000	1.0200	1.0000	1.0102	1.0078	1.0062
T		0	1.4142	1.1180	1.1180	1.0000	1.0200	1.0000	1.0000	1.0000
G			0	2.2361	1.4142	1.0541	1.1180	1.0200	1.0138	1.0102
A				0	1.0000	1.1180	1.0000	1.0308	1.02000	1.0138
A					0	1.4142	1.0000	1.0541	1.0308	1.0200
C						0	1.4142	1.0000	1.0000	1.0000
A							0	1.4142	1.1180	1.0541
C								0	1.0000	1.0000
T									0	1.0000
C										0

Table 9.7 L/L matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1	1.2649	0.6708	0.7310	0.7405	0.6161	0.7278	0.7524	0.7729
T		0	1	0.6125	0.6800	0.7898	0.7870	0.7601	0.7871	0.8086
G			0	1	0.8740	0.6800	0.7374	0.6818	0.7174	0.7460
A				0	1	0.9262	0.7836	0.7864	0.8168	0.8398
A					0	1	0.7071	0.7453	0.7864	0.8168
C						0	1	0.7071	0.7836	0.8284
A							0	1	0.9262	0.9262
C								0	1	1
T									0	1
C										0

In this method the leading eigenvalues of M/M and L/L matrices are used as DNA descriptors. These numerical parameters will facilitate the comparison of two DNA sequences. A vector can be constructed to characterise a DNA sequence from these data obtained from different characteristic curves. For example, if we consider only L/L matrix and since two symmetrical characteristic curves have the same L/L matrices, we can construct a six-component vector as a DNA descriptor by using the leading eigenvalues of these matrices associated with six characteristic curves.

Wang-Zhang Plot

In the Wang-Zhang method (Wang and Zhang 2006), there are three configurations, namely, non-A, non-C and non-G. A binary method is employed where the presence or absence of bases in the sequence is assigned 0 and 1 according to the specific configuration, for example, 0 for A and 1 for the others if it is a non-A plot. A similar protocol is implemented for the non-C and non-G plots. The graphs for the sequence ATGAACACTC are shown in Fig. 9.5a–c.

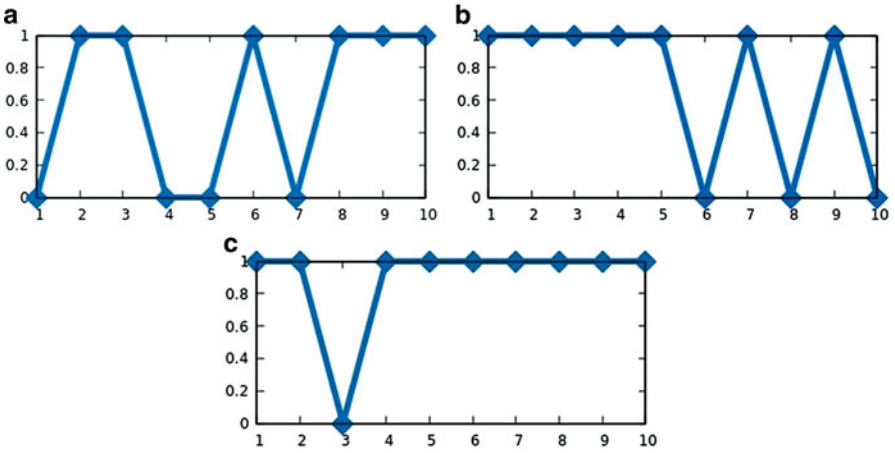

Fig. 9.5 (**a**) Non-A, (**b**) non-C and (**c**) non-G

Table 9.8 The co-ordinates and the leading eigenvalues calculated as per the Wang-Zhang model

	Coordinates			
Base	Non-A	Non-C	Non-G	Leading eigenvalues
A	0	1	1	Non-A = 0.3201 Non-C = 2.1605 Non-G = 4.1949
T	1	1	1	
G	1	1	0	
A	0	1	1	
A	0	1	1	
C	1	0	1	
A	0	1	1	
C	1	0	1	
T	1	1	1	
C	1	0	1	

In this model, the L/L matrices are constructed first from the three 2D graphs shown in Fig. 9.5 as a typical example to characterise the DNA sequence. Then the eigenvalues are computed from the matrices, and lastly, $\lambda_{\text{non-N}}$ are calculated which is $\lambda_{\text{non-N}}$ = maxeig (maximal eigenvalue) + mineig (minimal eigenvalue), where N = A, G, C. Proceeding in the same way, we obtain Table 9.8 for our sequence ATGAACACTC with the leading eigenvalues tabulated.

Yu-Hua Yao, Xu-Ying Nan and Tian-Ming Wang Plot

Yao et al. (2006) proposed a new 2D graphical representation of sequences, grouped as defined in section "Song-Tang plot", viz. W-S, M-K and R-Y. Here, we have plotted our sequence ATGAACACTC based on their system where the coordinates for the three curves are:

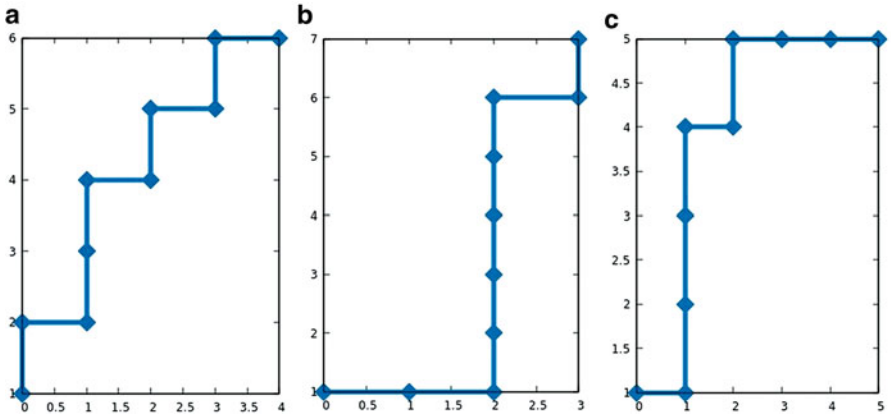

Fig. 9.6 (**a**) W-S curve, (**b**) M-K curve, (**c**) R-Y curve

Table 9.9 W-S curve

	Coordinates		Centre of mass		
Base	X	Y	μ_x	μ_y	g_R
A	0	1	1.7	3.8	4.1639
T	0	2			
G	1	2			
A	1	3			
A	1	4			
C	2	4			
A	2	5			
C	3	5			
T	3	6			
C	4	6			

The value of ρ for W-S curve = 1.4849

$$W - S\,\text{curve}\begin{cases} x = G_i + C_i \\ y = A_i + T_i \end{cases}$$

$$M - K\,\text{curve}\begin{cases} x = G_i + T_i \\ y = A_i + C_i \end{cases}$$

$$R - Y\,\text{curve}\begin{cases} x = C_i + T_i \\ y = A_i + G_i \end{cases}$$

The degeneracy is completely avoided in such representations. For our sample sequence, we get the plots as shown in Fig. 9.6a–c.

In this method, two sets of DNA descriptors are obtained. One is graph radius, subtending an angle θ with the x-axis, and the other is relative departure ρ which represents the difference between contents of bases of strong H-bonds and weak H-bonds in DNA sequence. Taking W-S curve as an example, the bisector is 50%(A, T)−50%S(G, C) line. A mean distance between the points in the DNA curve and this bisector is the new descriptor (Tables 9.9, 9.10 and 9.11):

Table 9.10 M-K curve

Base	Coordinates		Centre of mass		
	X	Y	μ_x	μ_y	g_R
A	0	1	1.9	3.6	4.0706
T	1	1			
G	2	1			
A	2	2			
A	2	3			
C	2	4			
A	2	5			
C	2	6			
T	3	6			
C	3	7			

The value of ρ for M-K curve = 1.3435

Table 9.11 R-Y curve

Base	Coordinates		Centre of mass		
	X	Y	μ_x	μ_y	g_R
A	0	1	2.0	3.5	4.0311
T	1	1			
G	1	2			
A	1	3			
A	1	4			
C	2	4			
A	2	5			
C	3	5			
T	4	5			
C	5	5			

The value of ρ for R-Y curve = 1.0607

$$\rho = 1/N \sum_{i=1}^{N} \sqrt{(x_i - i/2)^2 + (y_i - i/2)^2}$$
$$= \sqrt{2}/2N \sum_{i=1}^{N} |x_i - y_i|$$
$$= \sqrt{2}/2N \sum_{i=1}^{N} \sqrt{|(A_i + T_i) - (G_i + C_i)|}$$

Li and Ji TB Curve

In the 2D graphical method reported by Ji and Li (2006), let $X{=}X_1X_2\ldots X_n$ be a DNA primary sequence with n bases, and define a homomorphic map φ_1 by $\varphi_1(X) = \varphi_1(X_1)\varphi_1(X_2)\ldots\varphi_1(X_n)$ as:

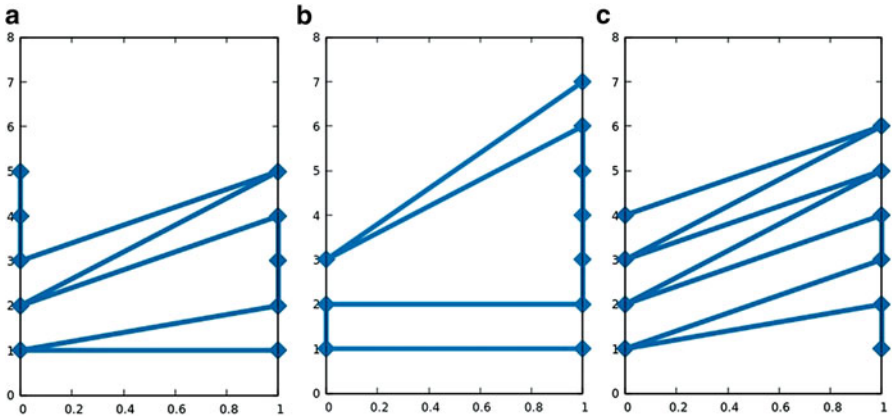

Fig. 9.7 (**a**) R-Y curve, (**b**) M-K curve, (**c**) W-S curve

Table 9.12 ED matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1	1	2	3	1.4142	4	2.2360	3.1622	4.1231
T		0	1.4142	2.2360	3.1622	1	4.1231	2	3	4
G			0	1	2	1	3	1.4142	2.2360	3.1622
A				0	1	1.4142	2	1	1.4142	2.2360
A					0	2.2360	1	1.4142	1	1.4142
C						0	2.2360	1	2	3
A							0	2.2360	1.4142	1
C								0	1	2
T									0	1
C										0

$$\varphi_1(X_i) = \begin{cases} 1(1, R_i) \text{ if } X_i \in \text{R} \\ 0(0, Y_i) \text{ if } X_i \in \text{Y} \end{cases}$$

$$\text{where } (i = 1, 2, \ldots, n)$$

where R_i (Y_i) is the cumulative occurrence of the numbers of bases ϵ $R(Y)$ in the first i bases. The DNA sequence is mapped into a series of nodes $P_i^{'}s$. Connecting the adjacent nodes, an R-Y curve is obtained. In a similar fashion, the other two maps are defined, and the M-K and W-S-TB curves for the sequence ATGAACACTC are obtained (Fig. 9.7a–c).

Here, three matrices, namely, ED matrix, M/M matrix and L/L matrix, are formed as before from which the leading eigenvalues are calculated. These are the descriptors of the corresponding DNA sequence. The formation of the matrix elements has already been discussed in section "Song-Tang plot" (Tables 9.12, 9.13 and 9.14).

Table 9.13 M/M matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1	0.5000	0.6667	0.7500	0.2828	0.6667	0.3194	0.3953	0.4581
T		0	0.7071	0.7453	1.0541	0.2000	0.6872	0.2857	0.3750	0.4444
G			0	0.3333	0.6667	0.2000	0.5000	0.2020	0.2795	0.3514
A				0	0.3333	0.2828	0.3333	0.1429	0.1768	0.2484
A					0	0.4472	0.1667	0.2020	0.1250	0.1571
C						0	0.3727	0.1429	0.2500	0.3333
A							0	0.3194	0.1768	0.1111
C								0	0.1250	0.2222
T									0	0.1111
C										0

Table 9.14 L/L matrix

	A	T	G	A	A	C	A	C	T	C
A	**0**	1	0.4142	0.5858	0.6797	0.2127	0.4076	0.1856	0.2433	0.2935
T		0	1	0.4142	0.7144	0.1770	0.4679	0.1810	0.2490	0.3065
G			0	1	1	0.2361	0.4055	0.1468	0.2103	0.2718
A				0	1	0.4370	0.3126	0.1158	0.1468	0.2103
A					0	1	0.1852	0.1852	0.1158	0.1468
C						0	1	0.1852	0.3125	0.4055
A							0	1	0.4370	0.2361
C								0	1	1
T									0	1
C										0

Zhao, Qi and Yang 2D Plot (Based on Nucleotide Triplets)

Zhao et al. (2015) proposed a 2D graphical representation method on the basis of nucleotide triplets (codons), as opposed to individual bases, in the DNA coding sequence strand. There are 64 codons including the 3 termination codons. Let $S = s_1 s_2 \ldots s_n$ represent a random DNA sequence. On the basis of standard genetic codes, each codon in the sequence can be defined with the aid of a mapping ϕ:

$$\phi(s_i s_{i+1} s_{i+2}) \begin{cases} (i,0) \text{if } s_i s_{i+1} s_{i+2} = \text{TAA, TAG, TGA} \\ (i,1) \text{if } s_i s_{i+1} s_{i+2} = \text{ATG} \\ (i,2) \text{if } s_i s_{i+1} s_{i+2} = \text{GCT, GCC, GCA, GCG} \\ (i,3) \text{if } s_i s_{i+1} s_{i+2} = \text{CGT, CGC, CGA, CGG, AGA, AGG} \\ (i,4) \text{if } s_i s_{i+1} s_{i+2} = \text{AAT, AAC} \\ (i,5) \text{ if } s_i s_{i+1} s_{i+2} = \text{GAT, GAC} \\ (i,6) \text{if } s_i s_{i+1} s_{i+2} = \text{TGT, TGC} \\ (i,7) \text{ if } s_i s_{i+1} s_{i+2} = \text{CAA, CAG} \\ (i,8) \text{ if } s_i s_{i+1} s_{i+2} = \text{GAA, GAG} \\ (i,9) \text{if } s_i s_{i+1} s_{i+2} = \text{GGT, GGC, GGA, GGG} \\ (i,10) \text{ if } s_i s_{i+1} s_{i+2} = \text{ATT, ATC, ATA} \\ (i,11) \text{if } s_i s_{i+1} s_{i+2} = \text{CAT, CAC} \\ (i,12) \text{if } s_i s_{i+1} s_{i+2} = \text{TTA, TTG, CTT, CTC, CTA, CTG} \\ (i,13) \text{ if } s_i s_{i+1} s_{i+2} = \text{AAA, AAG} \\ (i,14) \text{ if } s_i s_{i+1} s_{i+2} = \text{TTT, TTC} \\ (i,15) \text{ if } s_i s_{i+1} s_{i+2} = \text{CCT, CCC.CCA, CCG} \\ (i,16) \text{if } s_i s_{i+1} s_{i+2} = \text{TCT, TCC, TCA, TCG, AGT, AGC} \\ (i,17) \text{if } s_i s_{i+1} s_{i+2} = \text{ACT, ACC, ACA, ACG} \\ (i,18) \text{if } s_i s_{i+1} s_{i+2} = \text{TGG} \\ (i,19) \text{if } s_i s_{i+1} s_{i+2} = \text{TAT, TAC} \\ (i,20) \textit{ if } s_i s_{i+1} s_{i+2} = \text{GTT, GTC, GTA, GTG} \end{cases}$$

where $i = i^{th}$ is the base of sequence S. Applying this technique to our sequence ATGAACACTC, the resulting plot set is {(1, 1), (2, 0), (3, 8), (4, 4), (5, 17), (6, 11), (7, 17), (8, 12)}, which gives a 2_D curve termed as the considering codon degeneracy curve or CCD curve (Fig. 9.8).

Here the CCD curve is associated with a 21-dimensional characteristic vector V.

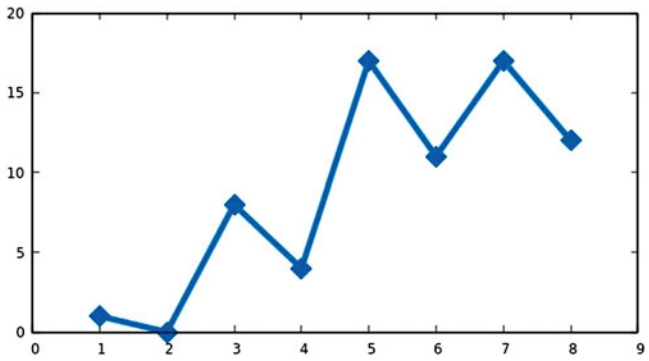

Fig. 9.8 CCD curve of the sequence ATGAACACTC using the method proposed by Zhao et al. (2015)

$$
\begin{aligned}
v_0 &= 0,\\
v_1 &= \underbrace{1+1+\cdots+1}_{n_1},\\
v_2 &= \underbrace{2+2+\cdots+2}_{n_2}, \ldots,\\
v_{20} &= \underbrace{20+20+\cdots+2}_{n_{20}}\\
V &= (v_0, v_1, \ldots, v_{20})
\end{aligned}
$$

where the parameters n_1, n_2, …, n_{20} indicate the sequential trinucleotides of the mapping ϕ, respectively. For example, for the formula:

$$v_2 = \underbrace{2+2+\cdots+2}_{n_2}$$

v_2 is the summation of the y-coordinate in the CCD curve which equals to 2. The n_2 denotes the number of the trinucleotides, GCT, GCC, GCA and GCG. Extending the characteristic vector V, they propose two characteristic vectors V_a and V_b for an evolutionary distance computing scheme whose formula is:

$$
\begin{aligned}
D(V_a, V_b) &= \frac{1 - V_a V_b / ||\, V_a||\, ||V_b||}{2}\\
&= \frac{1 - \sum_{i=1}^{21} v_a(i)\ \mathrm{x}\ v_b(i) / \sqrt{\sum_{i=1}^{21} (v_a(i))^2 \mathrm{x} \sum_{i=1}^{21} (v_b(i))^2}}{2}
\end{aligned}
$$

The distance of two CCD curves is represented by the distance $D(V_a, V_b)$ of two vectors, V_a and V_b, which can be used to quantify the evolutionary distance between sequences S_a and S_b. The two DNA sequences would be relatively similar if the $D(V_a, V_b)$ was small. The smaller the distance, the more similar (relatively) are the two sequences.

For our sample sequence ATGAACACTC, the descriptor, vector V_a, can be defined as:

$$V_a = (0, 1, 0, 0, 4, 0, 0, 0, 8, 0, 0, 11, 12, 0, 0, 0, 0, 17, 0, 0, 0, 0),$$

whereas, for another sequence, ATGGTGCACC, the descriptor, V_b, is:

$$V_b = (0, 1, 2, 0, 0, 0, 6, 0, 0, 9, 0, 11, 0, 0, 0, 0, 0, 17, 18, 0, 20)$$

The distance can then be computed as:

$$D(V_a, V_b) = 0.26995$$

The small distance value implies that the two sequences are relatively similar.

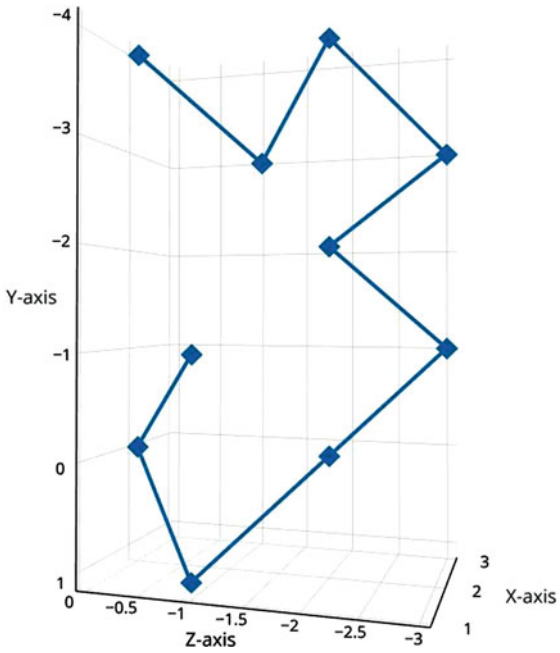

Fig. 9.9 3D plot as per Randic et al.'s representation of the sequence ATGAACACTC

9.2.2.2 3D Graphical Representations

Randic 3D Plot

The 3D graphical representation of DNA sequence was proposed by Randic et al. (2000). They associated the 2D method with a 3D graph by assigning each of the four bases to the corners of a tetrahedron where the points are as follows: (+1, −1, −1) for A, (−1, +1, −1) for G, (−1, −1, +1) for C and (+1, +1, +1) for T. The graph is plotted by placing the first base say A (for our mini sequence ATGAACACTC) at the corner position, i.e. at (+1, −1, −1), and the next base, i.e. T at (2, 0, 0). Proceeding in this manner, we obtain a plot of the sequence ATGAACACTC as shown in Fig. 9.9.

In this graphical plot of our mini sequence ATGAACACTC, a D_E/D_G matrix (from the Euclidean pairwise distance matrix D_E and a pairwise graph-theoretical distance matrix D_G of all the points on the graph) is constructed from which the leading eigenvalues are calculated to evaluate the descriptor of the sequence. In the following table, the coordinates of the bases of the sequence and the elements of the D_E/D_G matrix constructed from these coordinates are shown. Each matrix element is calculated as follows: Taking the (1,6) element, i.e. the first base A to the sixth base C as an example, the Euclidian distance is calculated from the coordinates shown in Table 9.15 and is divided by the minimum distance between two consecutive points

Table 9.15 The co-ordinates as per Randic 3D plot

Base	Coordinates		
	X	Y	Z
A	1	−1	−1
T	2	0	0
G	1	1	−1
A	2	0	−2
A	3	−1	−3
C	2	−2	−2
A	3	−3	−3
C	2	−4	−2
T	3	−3	−1
C	2	−4	0

Table 9.16 D_E/D_G matrix in Randic 3D model

	A	T	G	A	A	C	A	C	T	C
A	0	1	0.5774	0.3333	0.4082	0.2000	0.3333	0.2736	0.2041	0.2128
T		0	1	0.5774	0.6383	0.4083	0.5033	0.4303	0.2736	0.2887
G			0	1	1.000	0.6383	0.2000	0.6000	0.4303	0.4286
A				0	1	0.5774	0.6383	0.5774	0.3830	0.4303
A					0	1	0.5774	0.6383	0.4082	0.5033
C						0	1	0.5774	0.3333	0.4082
A							0	1	0.5774	0.6382
C								0	1	0.5774
T									0	1
C										0

which is $\sqrt{3}$. The division by the minimum distance is performed in order to normalise the distance scale, so that the Euclidian distance between adjacent vertices equals 1 and not $\sqrt{3}$ (due to taking the side of the cube to be 1). Now this value is divided by the number of edges before the sixth base, i.e. 5 here. Thus, the value obtained is $= (\sqrt{3}/\sqrt{3})/5 = 0.200$ (Table 9.16).

The descriptor value, i.e. the leading eigenvalue, in the Randic 3D model for this sequence is 5.38566.

Li et al. 3D Plot

A novel 3D representation of DNA sequences was proposed by Li et al. (2016). If the bases are represented by a set {A, G, C, T}, they graphically characterised a sequence in terms of nucleotide pairs by taking two combinations of the set in terms of a multiset $\{\infty.\,A, \infty.\,G, \infty.\,C, \infty.\,T\}$. The possible number of such pairs can be ten (Table 9.17).

Let V be a tetrahedron with its centre represented by $O(0, 0, 0)$. The bases A, C, G and T are plotted at the vertices $V_1(1, 1, 1)$, $V_2(-1, -1, 1)$, $V_3(1, -1, -1)$

Table 9.17 Two combinations of multiset $\{\infty. A, \infty. G, \infty. C, \infty. T\}$

Bases	A	G	C	T
A	{A, A}	{A, G}	{A, C}	{A, T}
G		{G, G}	{G, C}	{G, T}
C			{C, C}	{C, T}
T				{T, T}

Table 9.18 The co-ordinates of the Li et al 3D plot

Point	Dinucleotide	X	Y	Z
1	AT	0	1	0
2	TG	0	1	−1
3	GA	1	1	−1
4	AA	1.5774	1.5774	−0.4226
5	AC	1.5774	1.5774	0.5774
6	CA	1.5774	1.5774	1.5774
7	AC	1.5774	1.5774	2.5774
8	CT	0.5774	1.5774	2.5774
9	TC	−0.4226	1.5774	2.5774

and $V_4(-1,\ 1,\ -1)$, respectively. For all the line segments formed between each pair of vertices, a midpoint is regarded as follows:

- M is the midpoint of AC.
- K is the midpoint of GT.
- R is the midpoint of AG.
- Y is the midpoint of CT.
- W is the midpoint of AT.
- S is the midpoint of CG.

Ten directions $\underset{\text{OA}}{\rightarrow}, \underset{\text{OC}}{\rightarrow}, \underset{\text{OG}}{\rightarrow}, \underset{\text{OT}}{\rightarrow}, \underset{\text{OM}}{\rightarrow}, \underset{\text{OK}}{\rightarrow}, \underset{\text{OR}}{\rightarrow}, \underset{\text{OY}}{\rightarrow}, \underset{\text{OW}}{\rightarrow}, \underset{\text{OS}}{\rightarrow}$ are obtained from which ten-unit vectors are produced as:

$$r_A = \underset{\text{OA}}{\rightarrow} \Big/ \left\| \underset{\text{OA}}{\rightarrow} \right\|, r_C = \underset{\text{OC}}{\rightarrow} \Big/ \left\| \underset{\text{OC}}{\rightarrow} \right\| \ldots \text{ and so on.}$$

They associated each of the two combinations with the ten-unit vectors as:

$$\begin{aligned} &\{A,A\} \leftarrow r_A, \{A,G\} \leftarrow r_R, \{A,C\} \leftarrow r_M, \{A,T\} \leftarrow\ \ r_W, \\ &\{G,G\} \leftarrow r_A, \{G,C\} \leftarrow r_S, \{G,T\} \leftarrow r_K, \\ &\{C,C\} \leftarrow r_C, \{C,T\} \leftarrow r_Y, \{T,T\} \leftarrow\ \ r_T \end{aligned}$$

To plot our sequence ATGAACACTC with this method, we take two nucleotides at a time. Commencing from the origin, we move to the first dinucleotide AT, r_W and arrive at P_1, the first point on the 3D curve. Proceeding from P_1, we move towards the dinucleotide TG, r_K and reach the second point P_2. Progressing in this manner (Table 9.18), we get the curve as shown in Fig. 9.10.

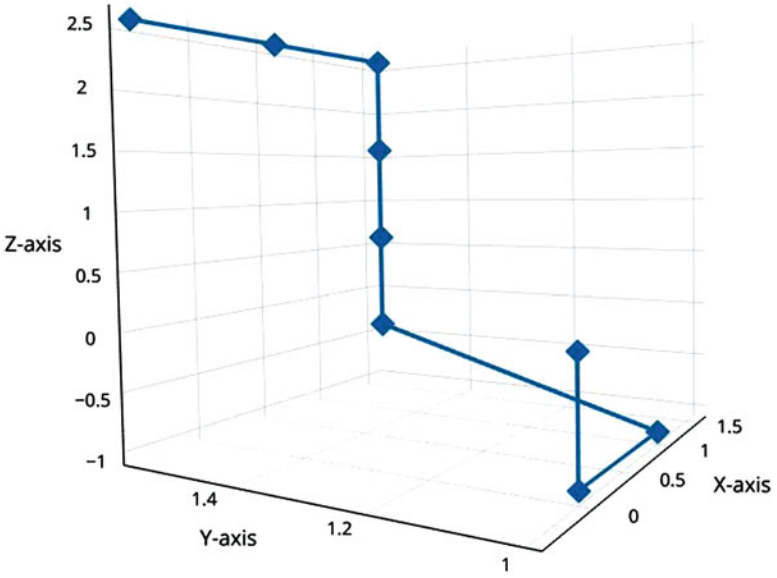

Fig. 9.10 3D dinucleotide plot method of Li et al. of the sequence ATGAACACTC

A common way to find the descriptor of DNA/RNA sequence is to construct a matrix from which the leading eigenvalue is calculated that serves as a descriptor for that sequence. But in this method, a different technique is introduced to calculate the DNA descriptor which is called 'piecewise function'. A 3D representation having n vertices, in order of their appearance in the curve, is partitioned into K parts; each part is termed as a cell. Therefore, the i^{th} cell:

$$\overrightarrow{U_{i-1}U_i} = (x_i - x_{i-1}, y_i - y_{i-1}, z_i - z_{i-1})$$

where $U_0 = (0, 0, 0)$

A DNA sequence can be numerically characterised by a $3K$ dimension vector V_{tp}:

$$\begin{aligned} V_{\text{tp}} = (&x_1 - x_0, x_2 - x_1, \ldots, x_k - x_{k-1}, \\ &y_1 - y_0, y_2 - y_1, \ldots, y_k - y_{k-1}, \\ &z_1 - z_0, z_2 - z_1, \ldots, z_k - z_{k-1}) \end{aligned}$$

which in the case of our sample sequence will be:

$$V_{\text{tp}} = (0, 1, 0.5774, 0, 0, 0, -1, -1, \\ 0, 0, 0.5774, 0, 0, 0, 0, 0, \\ -1, 0, 0.5774, 1, 1, 1, 0, 0)$$

Now, for any two sequences S_a, S_b, if their descriptor values are $a = (a_1, a_2, \ldots, a_{3K})$ and $b = (b_1, b_2, \ldots, b_{3K})$,respectively, their similarity can be calculated by the Euclidean distance given by:

$$d(a,b) = \sqrt{\sum_{j=1}^{3k} \left(a_j - b_j\right)^2}$$

whereby, the smaller the $d(a, b)$, the more similar are the two DNA sequences.

9.2.2.3 Comparative Study of Results from Different Graphical Methods

The following table contains the list of values of the descriptors prescribed for each of the methods described above for comparative study based on the coding sequences of four human globin genes. Some of the result sets had to be trimmed to accommodate in the list. Brief descriptions are given below for the results listed of some of the methods to understand the computation.

Yao-Nan-Wang plot (section "Yu-Hua Yao, Xu-Ying Nan and Tian-Ming Wang plot"): Three curves had been proposed as part of this method (i.e. RY, MK and WS). Ideally each curve is described by a 3D vector (of θ, g_R and ρ), among which only the results for the WS curve have been listed.

Li-Ji plot (section "Li and Ji TB curve"): The method proposes three curves the same as in Yao-Nan-Wang one. Only the leading eigenvalues for the L/L matrices corresponding to the RY family are shown in here.

Zhao-Qi-Yang plot (section "Zhao Qi Yang 2D plot (based on nucleotide triplets)"): The method proposes a 21-dimensional vector, of which the 0th set, given no weight, always ends up with a value of zero. Thus, it can be considered as a 20-dimensional vector, effectively. Values for only positions 1st, 2nd, 19th and 20th are listed for representational purpose.

Li et al. plot (section "Li et al. 3D plot"): This method prescribes a unique descriptor using 'piecewise function'. It depends mainly on the average number of bases per sequence in a given dataset (in our case, this is 440.25) and a factor 'k' (logarithmically dependent on the aforementioned 'average' factor, which is ~4 in our case,). As per the proposed representation, the length of the representative vectors for our dataset would have been 12 for each sequence. Hence, for the sake of simplicity and a biting space crunch in hand, we did away with the critical representation and presented the endpoints of the prescribed 3D curve as a numerical descriptor for each sequence instead (Table 9.19).

Table 9.19 Sequence descriptors and sample parameters computed for four human globin genes for all the GRANCH methods discussed above

	Alpha-globin HBA1-mRNA (NM_000558)	Beta-globin-HBB-mRNA (NM_000518)	Delta-globin-HBD-mRNA (NM_000519)	Gamma-globin HBG1-mRNA (NM_000559)	Remarks
Nandy plot section	47.310	31.864	29.092	11.132	g_R value
Yau plot section	158.900	157.402	155.728	155.263	g_R value
Randic 2D plot section	430.861	445.751	445.723	445.992	D/D – leading eigenvalue
Song-Tang plot section	429.202	444.274	444.300	444.380	D/D – leading eigenvalue
Wang-Zhang plot section	22.840168	29.7082743	27.8319458	24.365207	L/L – leading eigenvalue
Yao-Nan-Wang plot section	[0.296, 94.274, 46.555]	[0.614, 82.260, 20.930]	[0.697, 80.733, 14.521]	[0.736, 80.213, 10.986]	[θ, g_R, ρ] of WS curve
Li-Ji plot section	88.328	78.169	66.411	35.411	L/L – leading eigenvalue for RY curve
Zhao-Qi-Yang plot section	[4,90. . .57,460]	[7,70. . .95,640]	[10,70. . .95,620]	[11,62. . .76,420]	1st, 2nd, 19th and 20th position values of the 20-dimensional vector
Randic 3D plot section	101.177	61.562	59.945	46.591	D/D – leading eigenvalue
Li et al. plot section	[−23.97, −103.064, 22.053]	[−3.762, −38.939, −41.309]	[0.97, −28.743, −52.66]	[15.083, −24.011, −6.691]	The 3D coordinates of the endpoints for the prescribed 3D curve

9.2.3 Graphical Representation and Numerical Characterisation Methods for Protein Sequences

9.2.3.1 Introduction

Graphical representation schemes for protein sequences came about as a natural consequence of the development and applications of graphical representations of nucleotide sequences. In contrast to that of nucleic acid sequences, graphical representation of protein sequences is complicated, given there are 20 amino acid residues as the basic units of proteins, while nucleic acids can be defined by only 4 bases. Initial efforts were made to portray the amino acids in terms of four properties so they could be represented in a 2D coordinate system like for the DNA sequences and later extended to multidimensional representations (see review Randic et al. 2011). We proceeded to simplify the representation intuitively by projecting the primary sequence in a 20-dimensional Euclidean space which can be used to understand the phylogenetic relationship, provide mutational insight and identify conserved regions for groups of proteins. We briefly mention just this one method in this chapter on nucleic acid research for completeness.

9.2.3.2 Design of the 20D Graphical Representation Algorithm

A 20-dimensional Cartesian coordinate system is used to model the protein sequences. The 20 amino acids are assigned to 20 axes in the Cartesian framework (Nandy et al. 2009); all the axes are equivalent, and the choice of association of amino acids with the axes can be made arbitrarily but once made will remain constant for the duration of the exercise. The computation starts from the origin and moves a step in the appropriate axis direction for the first amino acid in the sequence, another step in the respective axial direction for the next amino acid in the sequence and so on. Tracing the path for the occurrences of all the amino acids in the sequence in order results in a series of points in the abstract 20D space generating a 20D curve. An index for the protein sequence can be considered as a vector from the origin of the coordinate system to the endpoint of the graph. In order to eliminate the disadvantage that different distributions of the residues with the same composition in the sequences will lead to degeneracy in these vectors, the algorithm is designed to choose the characteristics of a sequence by a weighted centre of mass approach, first used for DNA sequence (Raychaudhury and Nandy 1999) and defined by:

$$\mu_1 = \frac{\sum x_1}{N}, \mu_2 = \frac{\sum x_2}{N}, \ldots\ldots\ldots, \mu_{20} = \frac{\sum x_{20}}{N}$$

where the x_i's are the coordinate values of each point on the curve and N, a normalisation factor for the μ_is, is the number of amino acids in the protein sequence. Using these weighted averages, a protein graph vector $\vec{p_R}\,(\mu_1, \mu_2, \mu_3, \ldots\ldots\ldots, \mu_{20})$ and a protein graph radius p_R can be obtained:

$$p_R = \sqrt{\left(\mu_1^2 + \mu_2^2 + \ldots\ldots\ldots + \mu_{20}^2\right)}$$

Identical protein sequences have the same p_R values, which change with any alteration in the amino acid composition and distribution in the sequence. Δp_R is used to compare the differences between two protein sequences quantitatively as the Euclidean distance between two protein graph radii:

$$\Delta p_R = \sqrt{\left\{\left(\mu_1 - \mu_1'\right)^2 + \left(\mu_2 - \mu_2'\right)^2 + \ldots\ldots\ldots \left(\mu_{20} - \mu_{20}'\right)^2\right\}}$$

For multiple sequences, this method allows the comparison of pairwise differences, irrespective of sequence lengths. Thus, it can generate a distance matrix that could be used to construct phylogenetic trees, the advantage being that it does not require any multiple sequence alignments nor make any other model-dependent assumptions (Nandy 2009).

The 20D algorithms have been applied in phylogenetic analysis (Nandy 2009), mutational analysis of protein sequences on conserved surface accessible regions and drug target finding in influenza neuraminidase (Ghosh et al. 2010), rotavirus VP7 (Ghosh et al. 2012), Zika virus (Dey et al. 2017a) and others (Dey et al. 2016).

9.3 Applications of the New Methodology

We have briefly recounted in the introduction the various applications that have been made of alignment-free comparisons of DNA/RNA/protein sequences using the new methodology of graphical representation and numerical characterisation techniques. An area of particular concern that has seen many applications has been viral pandemics; specifically, in the twenty-first century, successive waves of epidemics and pandemics have swept across the globe starting with SARS (severe acute respiratory syndrome) pandemic of 2002–2003 that killed over 8000 people, the swine flu of 2009 (>100,000 deaths), the Ebola epidemic of 2014–2016 in West Africa (over 15,000 deaths) and the Zika virus pandemic that struck the Americas in 2015–2017 (>200,000 deaths). Understanding the characteristics of the viral spread and genetics would thus appear of urgent necessity. In this section we concentrate on the influenza pandemics as one of the most virulent killer diseases and summarise some of the specific applications of graphical representation methods to understand the character of the influenza virus, especially in relation to the pandemic varieties.

9.3.1 A Brief History of Flu Epidemiology

Viral outbreaks and epidemics have been recorded in historical archives from as far back as 400 BCE (Martin and Martin-Granel 2006). There have been written records of influenza outbreaks since the fifteenth century. The first cases of influenza in North America were possibly brought on by the European settlers in the late 1400s. The sixteenth century witnessed a flu pandemic spread from Russia to Europe via Africa (Pyle 1986), while England, Ireland and the USA were affected by three prominent outbreaks in the seventeenth century (Knobler et al. 2005).

The twentieth century saw one of the most horrific viral disasters of any era, viz. the 1918 'Spanish influenza'. This was a type A H1N1 virus that brought about an onslaught of worldwide pandemic, spreading to far corners of the world, including the Arctic and isolated islands in the Pacific (Mamelund 2011). By the end of the pandemic that had come in three waves between 1918 and 1919 (Taubenberger and Morens 2006), it had reportedly killed 50–100 million people (Patterson and Pyle 1991), mostly affecting the demographic age group of 20–40 years (Simonsen et al. 1998). During the peak of the pandemic in USA, more than 100,000 deaths were occurring in a week (Reid et al. 1999). It is interesting to note that post first infection wave, herd immunity within the population should have immunologically protected a second wave of infection in the same year. But on the contrary, the second wave drastically increased mortality rates. A key enigma, genetically, is the occurrence of the three infection waves in rapid succession within a period of 6–12 months: the H1N1 virus had to accumulate immense mutational changes or acquire key genetic alterations for it to become so fatal so rapidly which normally requires years of circulation within a population. However, in this case the virulence had increased over a matter of weeks (Reid et al. 1999).

Pandemics generally spread through trade routes when infected people travel or migrate from one country to another. The unprecedented proportions at which the Spanish influenza spread were only possible due to incidental military movements from one continent to the other for World War I (Erkoreka 2009).

This was followed by two more influenza epidemics in 1957 and 1968. The two viruses arose by reassortment of human and avian strains. The avian strain was from wild waterfowl found in Eurasia, and the human strain originated from the H1N1 strain that was already present in the population. It is speculated that the surface proteins of hemagglutinin and neuraminidase that had adapted to humans were replaced by those found in the avian strains (Reid et al. 2004). The 1957 epidemic, 'Asian influenza', originated in China and resulted in two million deaths caused by the type A H2N2 virus (Pyle 1986). The Asian flu virus next underwent antigenic shift to give rise to the H3N2 virus that brought about the 1968 'Hong Kong influenza' epidemic; in a report one million people died from the viral infection (Knobler et al. 2005). Towards the end of the century, in 1996, H5N1 virus was identified in China in geese which subsequently jumped the species barrier to infect poultry; some humans were incidentally infected in the 1997 wave and caused serious concern of another pandemic (De Jong et al. 1997). In the next few years,

the H5N1 avian flu spread across various countries around the world and millions of wild as well as poultry deaths were seen. Human deaths arising from this avian influenza through physical contact were recorded globally till early 2014, but till date no human-to-human transmission has been observed, fortunately.

The twenty-first century saw the most recent influenza pandemic in 2009. The H1N1(2009)pdm strain, known as the swine flu, started from Mexico and spread rapidly round the globe, leading to a death toll of 106,000–400,000 (Dawood et al. 2012), before it died down. In 2014–2015, the poultry industry in the USA was faced by an epidemic of H5N2 bird flu that is believed to have been a reassortment with Asian H5 and North American N2 proteins, which could mutate to a human-infecting pandemic although no human infection has been noticed to date (Nandy and Basak 2015). In 2017–2018, flu season in the USA, an H3N2 type A influenza epidemic has affected thousands of people and led to the death of around 140 children, once again underlining the high pathogenicity of the influenza virus.

9.3.2 Genetics of Influenza Virus

Influenza virus belongs to the Orthomyxoviridae family. They are of four types, A, B, C and D, referred to as IAV, IBV, ICV and IDV, respectively (CDC 2017a). Types A, B and C infect humans and a variety of animals, among which type A is the most virulent. Influenza D virus was identified as recently as 2012. Although, IDV has been found to infect cattle (Hause et al. 2014; WHO 2018a, b), its detection in swine has raised the possibility of its spread in other mammals, including humans. Wild aquatic birds are the natural reservoir of the influenza A virus. The bird flu scare of around 1997–2005 was noteworthy because of the rapidity with which the virus spread throughout the globe causing huge fatalities among birds and animals; it still continues to precipitate frequent outbreaks. Because of the worldwide spread, virulence and pandemic history, IAV is of primary importance among the four types. In this chapter, we emphasise only on influenza type A virus except where otherwise stated.

Influenza is a negative-sense strand RNA virus. It is characterised by a segmented genome (Bouvier and Palese 2008) where the IAV has 8 segmented genes coding for 11 proteins, as described in Table 9.20. The influenza A virion is roughly spherical in shape, typically around 100 nanometres wide. It consists of an envelope containing the two surface proteins, hemagglutinin (HA) and neuraminidase (NA), in the proportion of 4:1 (Bouvier and Palese 2008). The central hollow core contains the RNA genome and the other proteins essential for packaging and survival of the virion.

IAV is categorised into several subtypes based on their surface-exposed proteins HA and NA, such as H1N1, H3N2 and H5N1, among others. So far, 18 subtypes of HA and 11 subtypes of NA have been observed on the basis of their antigenicity (Tong et al. 2013), out of which only a few subtypes are found to infect humans (Nandy et al. 2014).

Table 9.20 Gene segments of influenza A virus (A/duck/Vietnam/HU5-1571/2016(H5N1)) and their encoded proteins. While most proteins are coded by the gene segments, PB2-F1, M2 and NEP are expressed from spliced RNAs or alternate reading frames. (Based on the table in Bouvier and Palese 2008)

Segment	Gene	Code	Function of the protein	Gene length (nt)	Protein length (aa)
1	Polymerase basic subunit	PB1	mRNA cap recognition	2280	760
2	Polymerase basic subunit	PB2	RNA elongation, endonuclease activity	2274	758
		PB1-F2	Pro-apoptotic activity	273	91
3	Polymerase acidic subunit	PA	Protease activity	2151	712
4	Hemagglutinin	HA	Receptor binding, fusion activities	1704	568
5	Nucleoprotein	NP	RNA-binding protein, nuclear import regulation	1497	499
6	Neuraminidase	NA	Sialidase activity, virus release	1350	450
7	Matrix protein	M1	vRNP interaction, RNA nuclear export regulation, viral budding	759	252
		M2	Virus uncoating and assembly	294	97
8	Nonstructural protein	NS1	Regulation of host gene expression	678	225
		NEP/NS2	Nuclear export of RNA	366	121

HA mediates entry of the virus in the host cell by binding with terminal sialic acid residues of the host cell glycoprotein receptors. The HA protein has two subdomains, HA1 and HA2. HA1 subdomain contains the receptor-binding region and the antigenic regions where many mutations occur. HA2 subdomain is comparatively more stable and functions to anchor the protein on the viral envelope. NA mediates exit of the virus from the host cell by cleaving the terminal sialic acid residues of the glycoprotein or glycolipid receptors (Chen and Li 2013).

The host cells primarily include epithelial cells of the nose, throat and lungs of mammals and intestines of birds. Following binding, the viral lipid membrane fuses with the host cell membrane. The terminal sialic acid residues of the human cell receptors, where the HA binds, are of two types, *α-2-3* linked and *α-2-6* linked, having differential distribution in the lower and upper respiratory tract of humans, respectively (Kumlin et al. 2008). Distribution of the *α-2-3*-linked and *α-2-6*-linked sialic acid receptors in swine respiratory tract is much more uniform. In contrast, birds have only *α-2-3*-linked sialic acid receptor in their respiratory tract. This atypical distribution of sialic-linked glycoprotein receptors in the animal kingdom plays a significant role in evolution and dissemination of the virus. In China and the Far East, where poultry and swine herds are raised in numerous villages and farms, avian influenza with *α-2-3*-linked sialic acid receptors can infect swine and develop

the ability to bind *α-2-6*-linked sialic acid receptors. This enables the modified virus to infect humans in the upper respiratory tract and make transmission between humans easy through sneezing and coughing. The scare with the bird flu, H5N1, was for this particular possibility – the virus has a high mortality ratio but to date can infect humans only through the *α-2-3*-linked sialic acid receptors. Were it to develop the ability to bind the *α-2-6* linked sialic acid receptors, humans would be susceptible to a pandemic catastrophe.

In another instance, the Spanish flu of 1918, known to be of avian origin (De 2018), is believed to have mutated to two forms (Reid et al. 2003). During the pandemic, two strains of H1N1 virus was co-circulating in nature. One strain was capable of binding *α-2-6*-linked sialic acid receptors present in the upper respiratory tract, having mutations at amino acid positions 190 (E190D) and 225 (D225G) in the HA1 segment (Reid et al. 2003). The other had the potential to bind both *α-2-6*- and *α-2-3*-linked sialic acid receptors present in the upper respiratory tract and lower respiratory epithelial cells, with mutation only at amino acid position 190 (E190D) in the HA1 segment (Reid et al. 2003).

RNA viruses are prone to very high rates of mutation (Barr and Fearns 2010; Drake 1993). The enzyme required for influenza gene replication in the host cell, RNA-dependent RNA polymerase, is devoid of any proofreading activity, which allows mutations during replication to be retained. Mutational changes in IAV are known to proceed by two well-documented mechanisms, antigenic drift and antigenic shift (CDC 2017b). Antigenic drifts arise out of random point mutations. If they occur in the viral antigenic site, the host cell may lose immunity against the virus, and the infection-immunity response cycle starts all over again. These drifts can cause seasonal outbreaks or epidemics, but they are not potent enough to give rise to pandemics. Antigenic shifts are less recurrent, occurring only when two or more virus subtypes infect a host cell. As the virus has segmented genome, the progeny virus may pack genes from different subtypes in a process known as reassortment. This may give rise to new and unique subtypes, sometimes introducing novel surface-exposed HA or NA that may predispose the viral host to high degree of pathogenesis. Antigenic shifts are often the causes of pandemics.

Interestingly, the Spanish flu virus is believed to have had greater than expected number of silent nucleotide or synonymous mutations (mutations that do not change the corresponding amino acids) in its genome in comparison with its avian and mammalian counterparts (Morens et al. 2010) and caused other pandemics through reassortments: the H2N2 pandemic of 1957, the H3N2 pandemic of 1968 and the H1N1 pandemic of 2009. The H5N1 bird flu is believed to have been a reassortment of genes from a prior H5N1 and a H9N2. Similarly, the H7N9 avian flu that struck China in 2013 (Gao et al. 2013) was a reassorted virus, picking up genes from multiple hosts like HA genes from H7N3 virus of domestic duck, NA from H7N9 virus of wild migratory birds and six other genes from the H9N2 virus of domestic poultry (CDC 2013; Liu et al. 2013). In 2015, there was an avian epidemic in the poultry market in the Midwest USA caused by H5N2 virus (Spackman et al. 2016), another product of reassortment which held possibilities of further reassortants.

There is also a third mechanism by which a virus may mutate to form a new strain. If a host cell is coinfected by two different strains of the same subtypes, the RNA

polymerase may jump during replication from one RNA segment to its counterpart on the other strain at a consensus sequence of the gene referred to as a break point, continue replication and then revert to the original strain at the next break point. This phenomenon is termed as copy-choice recombination. This can give rise to progeny virus having a novel RNA segment not present in either of the infecting strains. We have also considered the possibility of the natural division of distinct segments of a gene, such as HA1 and HA2 of HA as break points, a topic we discuss in some detail later. Recombination in viruses is a controversial subject where the incidence of recombination through copy-choice method is expected to be below 2% (Hao 2011; Boni et al. 2012). However, a mechanism where whole segments of a gene are exchanged between two strains of the same subtype by polymerase jumps has occurred in about 5% of the cases examined (De and Nandy 2015).

9.3.3 *Characterising the Flu Using Graphical Representation Methods*

Quantitative sequence descriptors, like the g_R of the 2D rectangular representation, provide a ready means to follow the progress of typical viral strains across geographical and temporal domains. To this end we have investigated different aspects of the influenza virus. Some of these are briefly recounted below to indicate to the interested reader how such studies can be actually done.

Considering the high morbidity and mortality among the avian population arising from infections with the H5N1 bird flu virus around the turn of the century and the serious concern over possible mutagenesis of the strain to a human-to-human-infecting one, we undertook detailed comparisons of the H5N1 strains' neuraminidase sequences over a 10-year span to determine what changes were taking place (Nandy et al. 2007). Assuming that mutational changes in the sequences will be quantified by estimates of the graph radii, g_R, in a 2D graphical representation model (Nandy 1994), we assembled a database of 173 neuraminidase sequences for the years 1996–2005 and computed the average g_R values separately over periods when the flu was highly pathogenic to humans (1996–1997 and 2003–2005) and when less pathogenic (1998–2002). We found from computations of the g_R that, e.g. an H5N1 neuraminidase A/Chicken/Hong Kong/220/97(H5N1) had a very close similarity to another NA sequence from the same year, A/HongKong/156/97(H5N1), a result that matched with a conventional phylogenetic study by Suarez et al. (1998), whereas comparison with A/chicken/China/1/02-(H5N1)NA 5 years later showed that the two sequences were dissimilar almost seven times as much as the dissimilarity between the previous two sequences as estimated through Δg_R. This provides a descriptor-based insight into how rapidly the virus mutates over the years. A more detailed analysis with the g_R of the individual segments showed that the g_R expanded in the years of high pathogenicity to humans, more so for the period 2003–2005, whereas the period of low pathogenicity showed reduced values of g_R. This indicates

a relative increase in the A, T component of the neuraminidase gene sequences in the pathogenic years, which could be a contributor to the high pathogenicity observed (Nandy et al. 2007).

A study of mutational changes in influenza strains such as of H5N1, H1N1, etc. is important for surveillance against the influenza subtypes building up resistance against the therapeutics and vaccines. Studying a group of 682 strains of the H5N1 virus over the years 1997–2008 (Ghosh et al. 2009), we came across an interesting phenomenon: How far a strain had travelled geographically and temporally. Considering the property that the same g_R of two sequences implies very close similarity of sequences (Nandy and Nandy 2003), we found that identical sequences have appeared over significant distances in space and time, raising questions about a virus' longevity. Based on the statistics, we found it compelling to hypothesise that virions could probably survive ex vivo (Ghosh et al. 2009, but also see Bean et al. 1982) in dried mud or dirt carried over long distances by wild birds to infect other birds and poultry in wetlands in distant areas.

The same investigations revealed an interesting facet of mutations through recombination. The role of recombination in viral sequence changes is hotly debated (Hao 2011; Boni et al. 2012), the general opinion being that if recombination does take place in RNA viruses at all through the copy-choice method, it would be below 2%. We enquired into the possibility of another style of recombination, complete segments of a gene being exchanged through polymerase jumps at the segment boundaries: the neuraminidase gene with three segments – transmembrane, stalk and body – is a good example. Taking A/chicken/Afghanistan/1573–65/2006(H5N1)NA as a test case (Ghosh et al. 2009) and computing g_R values for the three segments individually, we compared each segment with corresponding segments of a selected group of H5N1 neuraminidase genes in terms of the g_R values. We found that the g_R value for the transmembrane segment of this strain was identical, for example, with the g_R value of the transmembrane segment of A/turkey/Islamabad/NARC7873/2007(H5N1)NA, i.e. they have the exact same sequence, implying a segment exchange had taken place at some time. Similarly, the stalk segment was found identical in the test sample and in A/greatcrestedgrebe/Denmark/ 7498/06(H5N1)NA, and the body was found to have been duplicated in A/turkey/Islamabad-Pakistan/NARC-7871/02/2007(H5N1)NA, which implies that recombination-like events through exchanges of individual segments do take place (Ghosh et al. 2009).

To understand the extent of this phenomenon, we undertook a detailed survey of the hemagglutinin protein that is comparatively simpler in that it has two segments designated HA1 and HA2. Our study (De et al. 2016) involved a total of 1274 HA sequences comprised of H1N1, H3N2, H5N1 and H7N9 subtypes from Asia over the period 2010–2014. In this database we searched for sets of three strains where a HA1 from one strain and a HA2 from a different strain would combine to form a third, daughter strain (see schematic, Fig. 9.11). This was easily accomplished by computing g_R values for each segment for all the strains and comparing to find duplicates, with the proviso that the two parental strains were from the same time and place. We found a total of 73 daughter strains, but, interestingly, there were no

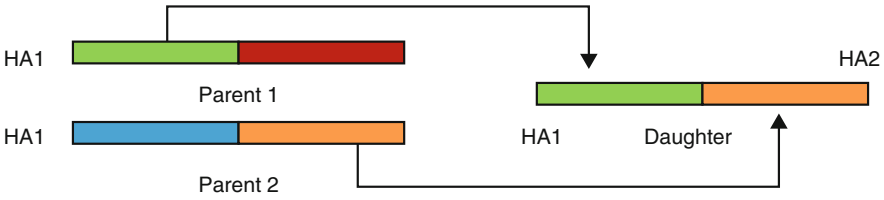

Fig. 9.11 Schematic of segment-wise recombination hypothesised for hemagglutinin. (Reproduced from Fig. 2 of De et al. (2016))

exchanges between different subtypes of the hemagglutinin. The daughter strains comprised 16 out of 427 H1 subtype, 30 of 408 H3 subtype, 20 of 350 H5 subtype and 7 of 89 H7 subtype in our database totalling 5.73% of all the strains in the database. Extending the analysis to 102 new strains from 2015 yielded another 5 instances of segment-wise recombination-like events. Thus, while the evidence for copy-choice recombination event mosaic in any gene sequence may not be readily evident, segment-wise recombination-like events do appear to occur in around 5% instances.

The two surface proteins, hemagglutinin and neuraminidase, have more in common: It appears that there is some kind of correlation between the two proteins such that suppression of the role of one influences changes in the sequence of the other and that coadaptation between the two was required for initiation of the H1N1 pandemic of 2009 (Wei 2010; Wagner et al. 2002; Xu et al. 2012; Hooper and Bloom 2013). There was also an observation that the influenza over the years had shown only a few subtypes instead of large numbers possible from the existence of 18 HA and 11 NA subtypes. Clearly, there was some interdependence between the two proteins, but there was no idea of what and how strong the interdependence could be. We undertook a study through graphical representation method (Nandy et al. 2014) and hypothesised a coupling between the HA and NA nucleotide sequences which could be measured through the g_R values. We assumed that subtypes that exist would have strong coupling and subtypes that are not found in nature would be weakly coupled, i.e. their interdependence would be weak, and hence those strains would not survive in nature. Our results (Nandy et al. 2014) amply bore this out and in fact showed that the couplings were stronger for avian flus compared to human-infecting flus, which one would have expected since wild birds seem to be the reservoir of influenza viruses of different kinds. We carried this analysis further and forecast the effects of possible reassortments of the H5N2 virus affecting poultry in the USA for therapeutic development and surveillance purposes (Nandy and Basak 2015).

As is evident from the above applications, g_R is a convenient way to scan hundreds and thousands of sequences in a short time interval to determine segments of interest. One particularly important query is identifying conserved segments in any gene sequence. By scanning the relevant sequences through a window of the appropriate size and determining the number of instances where the g_R values remain identical, the best conserved segments of a gene sequence can be identified.

Correspondingly, taking the protein sequence and computing the p_R values in the 20D model (see Sect. 9.2.3.2), the conserved domains in the protein can be identified. These techniques are an important tool in rational design of peptide vaccines where we may target those peptides that are conserved and have the capability of appropriate immune response to be the basis for vaccines that can outlive several generations of mutational changes. We had followed this procedure to identify best target sites in the NA of H5N1 (Ghosh et al. 2010) and the HA of H7N9 (Sarkar et al. 2015) influenza subtypes. Analyses for these targets are described briefly in the next section.

9.3.4 *Drugs and Vaccines Against the Flu and Vaccine Design by Graphical Method*

As in the case of most RNA viruses, drugs and vaccines against the influenza virus also face rapid obsolescence due to the high level of mutational changes. There have been several drugs that have been developed against the influenza, but only one remains effective at this time, while vaccines have to be changed every year.

Drugs used in influenza treatment globally belong to two groups having distinctive pharmacological attributes: M2 ion channel blockers and NA (neuraminidase) inhibitors (Stiver 2003). M2 protein forms a small proton channel across the viral envelope and mediates pH levels across viral membrane during cell entry and progeny maturation (Pielak and Chou 2011). It is vital for viral replication, and drugs amantadine and rimantadine were developed to block its function but has since been withdrawn because mutational changes in the prevailing influenzas have rendered them resistant to these drugs (CDC 2011). Present drugs available in the market are neuraminidase inhibitors are peramivir (trade name Rapivab), zanamivir (trade name Relenza) and oseltamivir (trade name Tamiflu). The neuraminidase inhibitors act both against IAV and IBV. However, according to the World Health Organisation (WHO), there is widespread resistance against oseltamivir globally (WHO 2009).

Vaccines that prepare the body's immune system against foreign pathogens form an alternative to drugs in the fight against viral diseases. Vaccines may be live-attenuated, inactivated or recombinant virus-like particle (VLP) type. With influenza there are so many different strains that finding a traditional vaccine against all types poses a perennial problem. Based on global surveillance data generated by WHO's in-house surveillance system, viz. the Global Influenza Surveillance and Response System (GISRS), vaccine strains are chosen for every year on the basis of preponderance of viral strains in each hemisphere. At present, for 2017–2018, WHO recommended four A and B strains, viz. A/Michigan/45/2015 (H1N1) pdm09-like virus, A/Hong Kong/4801/2014(H3N2)-like virus, B/Brisbane/60/2008-like virus (Victoria lineage) and B/Phuket/3073-like virus (Yamagata lineage), to be used as quadrivalent vaccine in the northern hemisphere (WHO 2017). The CDC also

recommends vaccines for a year based on their assessment of prevailing virus strains in the previous year. While these do reduce the scope of the infection, in some years the vaccine is only partly effective as in 2014–2015 season and again in the current year, i.e. 2017–2018. These are partly due to the continuing mutations in the viral strains and partly errors in the vaccine as is reported to have happened this year.

It has therefore been a lingering quest to develop a universal flu vaccine that could act against all strains of the virus all the time. There have been claims off and on of a universal flu vaccine, but to date no such vaccine has been marketed. We undertook a graphical and numerical analyses of 514 strains of the H5N1 and 425 strains of the H1N1 influenza virus neuraminidase gene and determined six regions which remained conserved (Ghosh et al. 2010). This was facilitated by our examination of the graphical representation of the neuraminidase gene sequence which showed a strongly conserved 50/51 base (17 aa) region near the 3′ end of the gene. When examined in the 3D crystallographic structure of the protein, this region was seen to be responsible for the bonding between the adjacent proteins in the neuraminidase quaternary structure and therefore extremely important for the structure's stability. The other regions we identified as being conserved also turned out to be surface exposed. These observations indicated a possibility, albeit subject to experimental verification, of designing inhibitors for broad-spectrum pandemic control of flu viruses with similar NA structure that could remain active for many generations of mutations of the underlying neuraminidase gene sequences (Ghosh et al. 2010). In another exercise, we explored the HA protein of H7N9 influenza virus using an improved protocol since there were concerns that the H7N9 could mutate to a human-infecting virus. We determined several targets in both the HA1 and HA2 regions of the hemagglutinin that could be used for rational design of peptide vaccines (Sarkar et al. 2015). It is to be noted that the HA1 of the hemagglutinin is highly variable, but HA2 is comparatively much more stable and is one of the targeted regions for a universal flu vaccine.

Another approach has recently been advocated for vaccine development. One of the disadvantages of current vaccines is their instability and weakness towards protease activity which mandates additional operational and transport costs; it also results in reduced biological activity and bioavailability to the immune system. Mile et al. (2018) constructed a T-cell-restricted IAV synthetic peptide containing D-amino subunits. D-amino acids are rarely found in nature and are inherently protected from protease degradation. Peptide was able to elicit an immune response in mice after oral administration. Such synthetic 'mimics' can be a cost-effective alternative to traditional vaccines.

9.4 Surveillance of the Flu

9.4.1 Background

With the high rate of mutations in influenza genome along with the high frequency and intensity of influenza epidemics and pandemics recurring every few years, surveillance of influenza genomic variations is an important issue for monitoring its changes and designing drugs and vaccines. Using influenza, hepatitis, poliomyelitis and malaria as examples, Alexander Langmuir expounded disease surveillance as the constant vigilance over the patterns and spread of disease occurrence via accumulation, integration, assessment and interpretation of data on morbidity and mortality along with other parameters and relevant information (Langmuir 1963). The Centers for Disease Control and Prevention (CDC) outlines surveillance to encompass identification of health problems for surveillance; collection, analysis and interpretation of data; dissemination of the data and its interpretations; and evaluation and improvement of surveillance (CDC 2006).

Surveillance of infectious viral diseases necessitates the analysis of antigenic constitutions of human-infecting viruses, as well as of viruses that belong to the same group which have not yet escaped their respective sylvatic cycles, in order to ascertain whether the molecular modifications have significantly altered known antigenic characteristics (Ghendon 1991). Some of the methods employed in viral discovery and analysis include tissue culture studies, immunohistochemical assays, singleplex and multiplex assays, serology and high-throughput sequencing (Lipkin and Firth 2013). One of the primary applications of disease surveillance and analysis is in the development of drugs and vaccines. The antigenic differences in the existing vaccine and the viral strains in circulation call for the synthesis of appropriate vaccine compositions to establish renewed immunity in the host population against epidemic- and pandemic-causing viruses (Smith et al. 2004). For surveillance of influenza, the Influenza Division at CDC collects specimens from over one million patients USA-wide which then follows a protocol for testing of the patient samples in various laboratories, followed by gene sequencing on approximately 6000 influenza viruses annually. Out of the pool of 6000 viruses, antigenic compositions are analysed in 2000 of them. The hemagglutination inhibition (HI) assay is used to identify the subtype of the hemagglutinin (HA) gene of a new influenza isolate on the basis of inhibition of hemagglutination by subtype-specific antibodies (Pedersen 2014). About 50 virus variants are filtered annually for potential vaccine production.

9.4.2 Application of Graphical Representation-Based Descriptors in Surveillance of Emerging Pathogens

Mathematical model analyses can extend the HI assays by quantitative estimation of antigenic differences; the results can also be reproduced visually to compare many

strains at a time and check out significant differences. Examination of a number of plots of the influenza H5N1 neuraminidase gene in a 2D graphical representation system (Nandy 1994) led to the discovery of a 50-base segment at the 3′ end that was very strongly conserved (Nandy et al. 2007) and led eventually to a proposal for design of a peptide vaccine (Ghosh et al. 2010). This was found to be applicable across multiple subtypes of influenza, including H5N1, H1N1, H7N1, H9N1, H10N1 and others.

That same work (Nandy et al. 2007) showed another facet of viral surveillance. As described in some detail in Sect. 9.3.3, for this study 173 strains of the H5N1 of the period 1996–2005 were grouped on the basis of pathogenic years in the human population, two denoting periods of human infections and another representing the absence of human cases. Significant differences in the g_R were observed between the two groups of strains. The observed differences implied that genetic drifts had occurred among the strains of the two periods with the mutations having led to a quantitative decrease in the hydrophobicity of the proteins of strains belonging to the infective period. This paper illustrated that a relatively simple mathematical metric, viz. g_R, can be related to the location and effects of mutations in RNA, sort through conserved regions in the RNA, compute similarity between multiple sequences and thus aid in the surveillance of changes in the viral genomes.

Graphical representation can also help us understand the collected data which might contain incomplete information and annotation. Partial coding sequences can be graphically analysed to determine which part of the whole gene it belongs to through an alignment-free model. The approximate location of the sequence fragment can be determined in the gene by inferences from the nucleotide distribution pattern and the quantification of the pattern in form of g_R values. In their study of the Zika envelope gene, Dey et al. (2017a) evaluated partial CDS fragments and pinpointed their nucleotide start and end positions within the whole envelope gene sequence. Moreover, by cross-checking with existing annotated complete gene sequences, they were also able to identify additional peptide fragments in their sample sequences which contributed to the genetic differences between sequences collected from two different locations, Uganda and Brazil.

Extending the feature of determining sequence similarity from g_R values, which directly correlate with the nucleotide distribution in gene sequences, the graphical method can also be specifically employed to deduce the mixing between viruses originating from different regions due to homologous recombination. This is significant in understanding the geographical dispersal and trend of viral strain and subtype circulation, a key point in surveillance study of disease spread. A study (De et al. 2016) of the hemagglutinin (HA) gene of H1N1, H5N1, H3N2 and H7N9 subtypes from Asia, spanning the years from 2010 to 2014, showed that there were homologous recombinations of whole segments between the same subtypes. Some instances indicated that the parental and progeny strains were from same geographic regions, while others showed unexpected regional disparity, for example, parental H1N1 strains were found to be from Kowloon and Guangdong in China, and the daughter strain was isolated in Singapore. These genomic trails can be used to glean movement patterns of pathogenic viruses. In another study (Ghosh et al. 2009), it

was found that H5N1 strains in geese from Qinghai, China, were also observed 5000 km away in strains in swans from Southern Russia in 2005; H5N1 strains were detected in poultry in Turkey that were also observed in chicken in Israel, via the trade route; and the same H5N1 strain were found in Egypt and Ghana, while H5N1 samples collected in 2006 from ducks in Hunan had the same sequence as found in human samples isolated from Indonesia.

Another important utility of the graphical representations is inferring the lineage of the strains. The g_R can be used to generate non-alignment-based phylogenetic trees that can indicate the evolutionary relationships between the viruses (Liao et al. 2005, 2006). Its application has also established new insights about virus families, exemplified by dengue type 2 virus (DENV2) which belongs to the flavivirus family along with Zika virus, yellow fever virus, West Nile virus and Japanese encephalitis virus, among others. The envelope gene sequences of DENV2 greatly deviate from the other aforementioned flaviviruses in their nucleotide distribution and composition. This was a new find compared to available literature which has classified dengue to be similar to the other flavivirus members. The change in their g_R values is significant (Dey et al. 2017b). This type of analytical observation will aid in surveillance of pathogens for antiviral and vaccine development.

These observations show the role of descriptors g_R in the surveillance of viruses. Earlier in the chapter, we described methods for calculation of other sequence descriptors. Different descriptors may encode different structural information on the sequences. So, a collection of sequence descriptors or orthogonal factors like principal components (PCs) derived from them may be more powerful tools for the comparison and surveillance of emerging viral pathogenesis. Such a critical analysis of different methodologies was done by us some time ago (Sen et al. 2016) and will be further developed and implemented in other areas.

9.4.3 Big Data and Social Networks in Surveillance Programmes

Dr. Tarun Weeramanthri introduced the term precision public health (Severi et al. 2014) and is now defined as the usage of computational and technological progress and big data to improve disease surveillance (Dolley 2018). The primary goal of precision public health is the accuracy involved in gathering data including emerging pathogens, reactions based on susceptibility and geographical distribution of the diseases. It makes use of real-time data generated, among other computational methods, to gather quick observable data. Traditional methods of forecasting influenza trends by CDC usually lag behind real time by 1–2 weeks, whereas information contained in cloud-based electronic health records and search queries in the Internet are typically available near real time. Yang et al. (2017) combine these cloud-based records and Internet searches with historical flu data and use dynamically selected set of variables to give the best fit in their model for the 2013–2016 flu season. Their

results correctly estimated the peak timing and magnitude of the studies of flu season. These kinds of predictive models of influenza activity help public health officials prepare and allocate resources for possible disease outbreaks.

However, as with any individual's information, the use of public data calls for discretion: consent, privacy and security must be upheld and protected. Furthermore, an application that has come into prominence is the use of millions of data available from the social media platforms on the World Wide Web. However, the risk of inaccuracy in predictive analysis based on such data is high. A wide margin of error is required to compensate for over- or underestimation. Google Flu Trends, which is the recent surveillance tracking arm of the search engine giant, came into the spotlight for over-calculating the doctor visits for influenza-like illness than the reports generated by CDC which relies on records procured from laboratories (Lazer et al. 2014). The idea was to correlate search criteria with flu incidence and establish a trend. The initial trials matched 50 million searches with 1152 data points (Ginsberg et al. 2009) and came out with erroneous results due to certain ad hoc procedures. Even after corrections to the programmes, GFT persistently overestimated flu prevalence and has been relegated to the background for now in favour of more traditional procedures as adopted, for example, by CDC. Ideally, perhaps a combination of the two approaches could auger for better predictive trend analyses.

9.4.4 Ethics in Surveillance

With any scientific research, one must take into account the ethical implications of their work and inferences. A bone of contention in the prevention of spread of avian influenza and possible epidemics in human population is the recent trend in culling of poultry birds suspected to be infected by the H5N1 and H7N9 avian flu. The ethical dilemma arises between the prices of avian life against human population safety. Furthermore, WHO sheds light on the risks which the healthcare professionals place themselves in and also illuminates on the extent of travel restrictions that can be imposed in order to mitigate the spread of disease and the individual rights to freedom of movement. In 2006, WHO published a manual to integrate ethics with the influenza pandemic response structure (WHO 2007). The guide advocates sharing of surveillance information across borders during the pandemic, as well as prior and post pandemic. Resnik (2013) talks about the dilemma faced with the morality of dissemination of scientific knowledge, exemplified by the two censored and redacted papers, authored by Kawaoka and Fouchier Enserink 2012). Their papers showed the conclusions of genetically inducing mutations in H5N1 that conferred onto the viruses the ability to transmit through air among ferrets in the form of respiratory water droplets. The authors asserted that similar conditions among human populations would produce similar results. The NSABB, National Science Advisory Board for Biosecurity, allowed revised and heavily edited versions of the papers to be published in fear of dual-use research concerns (DURC) where

scientific findings can be used for catastrophic implications. The NSABB, after much deliberation, recommended the full publication of their work.

9.5 Summary

In this chapter we have outlined powerful and novel alignment-free graphical representation and numerical characterisation (GRANCH) tools for comparative analyses of biomolecular sequences. We explained in brief some of the proposals put forward for graphical representation and numerical characterisation in various dimensionalities and mentioned some of the many applications done using these novel, alignment-free approaches. In particular, we concentrated upon several applications done through the simple and intuitive 2D graphical representation of Nandy (1994) to give an idea of how these approaches can be utilised to compare and contrast the several variants of the influenza virus, determine homologous recombinations between strains of the same influenza subtype, identify conserved segments in influenza gene sequences and design peptide vaccines. We have seen that these approaches can yield phylogenetic trees to understand the relationships between the various strains and subtypes and assist in focused surveillance to ensure advance knowledge of developments that could lead to epidemic and pandemic varieties. Graphical representation and numerical characterisation thus are very general but quantitative approaches that can be used to unravel myriad aspects of viral characteristics.

References

Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Villa T, Uriarte E (2009) Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from coffea arabica and prediction of a new sequence. J Proteome Res 8:2122–2128

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17):3389–3402

Bai F, Wang T (2006) On graphical and numerical representation of protein sequences. J Biomol Struct Dyn 23(5):537–546

Barr JN, Fearns R (2010) How RNA viruses maintain their genome integrity. J Gen Virol 91 (6):1373–1387. https://doi.org/10.1099/vir.0.020818-0

Bean B, Moore BM, Sterner B, Peterson LR, Gerding DN, Balfour HH Jr (1982) Survival of influenza viruses on environmental surfaces. J Infect Dis 146(1):47–51

Bielinska-Waz D, Clark T, Waz P, Nowak W, Nandy A (2007) 2D-dynamic representation of DNA sequences. Chem Phys Lett 442:140–144

Boni MF, Smith GJ, Holmes EC, Vijaykrishna D (2012) No evidence for intrasegment recombination of 2009 H1N1 influenza virus in swine. Gene 494(2):242–245

Bouvier NM, Palese P (2008) The biology of influenza viruses. Vaccine 26(4):D49–D53. https://doi.org/10.1016/j.vaccine.2008.07.039

CDC (2006) Principles of epidemiology in public health practice third edition. Updated on May 2012. https://www.cdc.gov/ophss/csels/dsepd/ss1978/SS1978.pdf

CDC (2011) Antiviral agents for the treatment and chemoprophylaxis of influenza: recommendations of the advisory committee on immunization practices. https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6001a1.htm

CDC (2013) Genetic evolution of H7N9 virus in China. https://www.cdc.gov/flu/pdf/avianflu/h7n9-reassortment-diagram.pdf

CDC (2017a) Types of influenza viruses. https://www.cdc.gov/flu/about/viruses/types.htm

CDC (2017b) How the flu virus can change: "Drift" and "Shift". https://www.cdc.gov/flu/about/viruses/change.htm

Chen L, Li F (2013) Structural analysis of the evolutionary origins of influenza virus hemagglutinin and other viral lectins. J Virol 87(7):4118–4120. https://doi.org/10.1128/JVI.03476-12

Chou PY, Fasman GD (1974a) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222

Chou PY, Fasman GD (1974b) Prediction of protein conformation. Biochemistry 13:222–245

Cruz-Monteagudo M, González-Díaz H, Borges F, Dominguez ER, Cordeiro MN (2008) 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. Quantitative proteome-toxicity relationships (QPTR) based on mass Spectrum spiral entropy. Chem Res Toxicol 21:619–632

Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, Bandaranayake D, Breiman RF, Brooks WA, Buchy P, Feikin DR, Fowler KB, Gordon A, Hien NT, Horby P, Huang QS, Katz MA, Krishnan A, Lal R, Montgomery JM, Mølbak K, Pebody R, Presanis AM, Razuri H, Steens A, Tinoco YO, Wallinga J, Yu H, Vong S, Bresee J, Widdowson MA (2012) Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis 12(9):687–695. https://doi.org/10.1016/S1473-3099(12)70121-4

De A, Nandy A (2015) An insight to segment based genetic exchange in influenza a virus: an in silico study. In: Proceedings of the MOL2NET, 5–15 December 2015; Sciforum Electronic Conference Series 1(015). https://doi.org/10.3390/MOL2NET-1-b015

De A (2018) Molecular evolution of hemagglutinin gene of influenza a virus. Front Biosci (Schol Ed) 1(10):101–118

De Jong J, Claas E, Osterhaus A, Webster R, Lim W (1997) A pandemic warning? Nature 389:554–554. https://doi.org/10.1038/39218

De A, Sarkar T, Nandy A (2016) Bioinformatics studies of influenza a hemagglutinin sequence data indicate recombination-like events leading to segment exchanges. BMC Res Notes 9:222. https://doi.org/10.1186/s13104-016-2017-3

Dey S, De A, Nandy A (2016) Rational design of peptide vaccines against multiple types of human papillomavirus. Cancer Informat 15(S1):1–16. https://doi.org/10.4137/CIN.S39071

Dey S, Nandy A, Basak SC, Nandy P, Das S (2017a) A bioinformatics approach to designing a Zika virus vaccine. Comput Biol Chem 68:143–152. https://doi.org/10.1016/j.compbiolchem.2017.03.002

Dey S, Roy P, Nandy A, Basak S, Das S (2017b) Comparison of base distributions in Dengue, Zika and Other Flavivirus envelope and NS5 genes. Paper presented at In: Proceedings of the MOL2NET, International conference on multidisciplinary *Sciences*, Sciforum electronic conference series 3. https://doi.org/10.3390/mol2net-03-04966

Dodin G, Vandergheynst P, Levoir P, Cordier C, Marcourt L (2000) Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J Theor Biol 206:323–326. https://doi.org/10.1006/jtbi.2000.2127

Dolley S (2018) Big Data's role in precision public health. Front Public Health 6:68. https://doi.org/10.3389/fpubh.2018.00068

Drake JW (1993) Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci USA 90:4171–4175
Enserink M (2012) Free to speak, Kawaoka reveals flu details while Fouchier stays Mum. Science. http://www.sciencemag.org/news/2012/04/free-speak-kawaoka-reveals-flu-details-while-fouchier-stays-mum
Erkoreka A (2009) Origins of the Spanish influenza pandemic (1918–1920) and its relation to the First World War. J Mol Genet Med: Int J Biomed Res 3(2):190–194
Estrada E, Uriarte E (2001) Recent advances on the role of topological indices in drug discovery research. Curr Med Chem 8:1573–1588
Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W et al (2013) Human infection with a novel avian-origin influenza A (H7N9) virus. N Engl J Med 368(20):1888–1897
Gates MA (1986) A simple way to look at DNA. J Theor Biol 11:319–328
Ghendon Y (1991) Influenza surveillance. Bull World Health Organ 69(5):509–515
Ghosh A, Nandy A, Nandy P, Gute BD, Basak SC (2009) Computational study of dispersion and extent of mutated and duplicated sequences of the H5N1 influenza neuraminidase over the period 1997− 2008. J Chem Inf Model 49(11):2627–2638. https://doi.org/10.1021/ci9001662
Ghosh A, Nandy A, Nandy P (2010) Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. BMC Struct Biol 10(6)
Ghosh A, Chattopadhyay S, Chawla-Sarkar M, Nandy P, Nandy A (2012) In silico study of rotavirus VP7 surface accessible conserved regions for antiviral drug/vaccine design. PLoS One 7(7):e40749. https://doi.org/10.1371/journal.pone.0040749
Gibbs AJ, McIntyre GA (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur J Biochem 16:1–11
Ginsberg J et al (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014. https://doi.org/10.1038/nature07634
González-Díaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1025–1039
Gonzalez-Diaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008a) Proteomics, networks and connectivity indices. Proteomics 8:750–778
Gonzalez-Diaz H, Prado-Prado F, Ubeira FM (2008b) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 8:1676–1690
Hanson, RW (2003) Fast fourier transform analysis of DNA sequences. BA thesis, Reed College, Portland
Hao W (2011) Evidence of intra-segmental homologous recombination in influenza A virus. Gene 481(2):57–64
Harary F (1969) Graph theory. Addison-Wesley, Boston
Hause BM, Collin EA, Liu R, Huang B, Sheng Z, Lu W, Wang D, Nelson EA, Li F (2014) Characterization of a novel influenza virus in cattle and swine: proposal for a new genus in the Orthomyxoviridae family. MBio 5(2):e00031–e00014. https://doi.org/10.1128/mBio.00031-14
Hooper KA, Bloom JD (2013) A mutant influenza virus that uses an N1 neuraminidase as the receptor-binding protein. J Virol 87(23):12531–12540
Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) Graph theoretical matrices in chemistry. CRC Press, Boca Raton
Ji M, Li C (2006) TB curve, a new 2D graphical representation of DNA sequence. J Math Chem 40 (2). https://doi.org/10.1007/s10910-006-9063-3
Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14):3059–3066. https://doi.org/10.1093/nar/gkf436
Knobler SL, Mack A, Mahmoud A et al (eds) (2005) The story of influenza. The threat of pandemic influenza: are we ready? Institute of Medicine (US) Forum on Microbial Threats Workshop Summary. National Academies Press (US), Washington (DC)

Kobori Y, Mizuta S (2015) Similarity estimation between DNA sequences based on local pattern histograms of binary images. Biorxiv 1–17. https://doi.org/10.1101/016089
Kumlin U, Olofson S, Dimock K, Arnberg N (2008) Sialic acid tissue distribution and influenza virus tropism. Influenza Other Respir Viruses 2(5):147–154
Langmuir AD (1963) The surveillance of communicable diseases of national importance. N Engl J Med 268:182–192
Larionov S, Loskutov A, Ryadchenko E (2008) Chromosome evolution with naked eye: palindromic context of the life origin. Chaos 18:013105
Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203–1205. https://doi.org/10.1126/SCIENCE.1248506
Leong PM, Morgenthaler S (1995) Random walk and gap plots of DNA sequences. Comput Appl Biosci 11:503–507
Li C, Fei W, Zhao Y, Yu X (2016) Novel graphical representation and numerical characterization of DNA sequences. Appl Sci 6:63. https://doi.org/10.3390/app6030063
Liao B, Wang T (2004) Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. J Chem Inf Comput Sci 44:1666–1670
Liao B, Ding K (2006) A 3D graphical representation of DNA sequences and its application. Theo Comput Sc 358:56–64
Liao B, Tan M, Ding K (2005) Application of 2-D graphical representation of DNA sequence. Chem Phys Lett 414:296–300
Liao B, Xiang X, Zhu W (2006) Coronavirus phylogeny based on 2D graphical representation of DNA sequence. J Comput Chem 27:1196–1202
Lipkin WI, Firth C (2013) Viral surveillance and discovery. Curr Opin Virol 3(2):199–204. https://doi.org/10.1016/j.coviro.2013.03.010
Liu D, Shi W, Shi Y, Wang D, Xiao H, Li W, Bi Y et al (2013) Origin and diversity of novel avian influenza a H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. Lancet 381(9881):1926–1932. https://doi.org/10.1016/S0140-6736(13)60938-1
Martin PM, Martin-Granel E (2006) 2,500-year evolution of the term epidemic. Emerg Infect Dis 12 (6):976–980
Mamelund SE (2011) Geography may explain adult mortality from the 1918-20 influenza pandemic. Epidemics 3(1):46–60. https://doi.org/10.1016/j.epidem.2011.02.001
Mile JJ, Tan MP, Dolton G, Edwards ESJ, Sae G, Laugel B et al (2018) Peptide mimic for influenza vaccination using nonnatural combinatorial chemistry. J Clin Invest 128:1–12. https://doi.org/10.1172/JCI91512
Morens DM, Taubenberger JK, Harvey HA, Memoli MJ (2010) The 1918 influenza pandemic: lessons for 2009 and the future. Crit Care Med 38(4):e10–e20
Nandy A (1994) A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr Sci 66(4):309–314
Nandy A (1996a) Graphical analysis of DNA sequence structure. III. Indications of evolutionary distinctions and characteristics of introns and exons. Curr Sci 70:661–668
Nandy A (1996b) Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. CABIOS 12(1):55–62
Nandy A (2009) Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. PLoS One 4(8):e6829. https://doi.org/10.1371/journal.pone.0006829
Nandy A, Basak SC (2015) Prognosis of possible Reassortments in recent H5N2 epidemic influenza in USA: implications for computer-assisted surveillance as well as drug/vaccine design. Curr Comp-Aided Drug Des 11:110–116
Nandy A, Nandy P (2003) On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models. Chem Phys Lett 368:102–107
Nandy A, Harle M, Basak SC (2006) Mathematical descriptors of DNA sequences: development 1276 and applications. ARKIVOC 9:211–238

Nandy A, Basak SC, Gute BD (2007) Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. J Chem Inf Model 47(3):945–951

Nandy A, Ghosh A, Nandy P (2009) Numerical characterization of protein sequences and application to voltage-gated sodium channel a subunit phylogeny. In Silico Biol 9:77–87

Nandy A, Sarkar T, Basak SC, Nandy P, Das S (2014) Characteristics of influenza HA-NA interdependence determined through a graphical technique. Curr Comp-Aided Drug Des 10:285–302

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Patterson KD, Pyle GF (1991) The geography and mortality of the 1918 influenza pandemic. Bull Hist Med 65:4–21

Pedersen JC (2014) Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus. Methods Mol Biol 1161:11–25. https://doi.org/10.1007/978-1-4939-0758-8_2

Pielak RM, Chou JJ (2011) Influenza M2 proton channels. Biochim Biophys Acta 1808 (2):522–529

Pyle GF (1986) The diffusion of influenza: patterns and paradigms. Rowan & Littlefield, New Jersey

Qi X, Wu Q, Zhang Y, Fuller E, Zhang C-Q (2011) A novel model for DNA sequence similarity analysis based on graph theory. Evol Bioinforma 7:149–158. https://doi.org/10.4137/EBO.S7364

Randic M, Vracko M, Nandy A, Basak SC (2000a) On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inf Comput Sci 40:1235–1244

Randic M, Vracko M, Nandy A, Basak SC (2000b) On 3-D representation of DNA primary sequences. J Chem Inf Comput Sci 40:1235–1244

Randic M, Vracko M, Lers N, Plavsic D (2003) Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett 368:1–6

Randic M, Zupan J, Balaban AT (2004) Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem Phys Lett 397(1–3):247–252 https://doi.org/10.1016/j.cplett.2004.08.118

Randic M, Lers N, Plavsic D, Basak SC, Balaban AT (2005) Four-color map representation of DNA or RNA sequences and their numerical characterization. Chem Phys Lett 407:205–208

Randic M, Zupan J, Balaban AT, Drazen V-T, Plavsic D (2011) Graphical representation of proteins. Chem Rev 111(2):790–862

Raychaudhury C, Nandy A (1999) Indexing scheme and similarity measures for macromolecular sequences. J Chem Inf Comput Sci 39:243–247

Reid AH, Fanning TG, Hultin JV, Taubenberger JK (1999) Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. Proc Natl Acad Sci 96(4):1651–1656. https://doi.org/10.1073/pnas.96.4.1651

Reid AH, Janczewski TA, Elliot AJ, Daniels RS, Berry CL, Oxford JS, Taubenberge JK (2003) 1918 influenza pandemic caused by highly conserved viruses with two receptor-binding variants. Emerg Infect Dis 10:1249–1253

Reid AH, Taubenberger JK, Fanning TG (2004) Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. Nat Rev Microbiol 2(11):909–914

Resnik DB (2013) H5N1 avian flu RESEARCH and the ethics of KNOWLEDGE. Hast Cent Rep 43(2):22–33. https://doi.org/10.1002/hast.143

Sarkar T, Das S, De A, Nandy P, Chattopadhyay S, Chawla-Sarkar M, Nandy A (2015) H7N9 influenza outbreak in China 2013: in silico analyses of conserved segments of the hemagglutinin as a basis for the selection of peptide vaccine targets. Comput Biol Chem 59:8–15

Sen D, Dasgupta S, Pal I, Manna S, Basak SC, Nandy A, Grunwald G (2016) Intercorrelation of major DNA/RNA sequence descriptors–a preliminary study. Curr Comput Aided Drug Des 12 (3):216–228. https://doi.org/10.2174/1573409912666160525111918

Severi G, Southey MC, English DR, Jung CH, Lonie A, McLean C et al (2014) Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. Breast Cancer Res Treat 48(3):665–673. https://doi.org/10.1007/s10549-014-3209-y

Simonsen L, Clarke MJ, Schonberger LB, Arden NH, Cox NJ, Fukuda K (1998) Pandemic versus epidemic influenza mortality: a pattern of changing age distribution. J Infect Dis 178:53–60

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. Nucleic Acids Res 13:645–656

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305 (5682):371–376

Song J, Tang H (2005) A new 2-D graphical representation of DNA sequences and their numerical characterization. J Biochem Biophys Methods 63:228–239

Spackman E, Pantin-Jackwood MJ, Kapczynski DR, Swayne DE, Suarez DL (2016) H5N2 highly pathogenic avian influenza viruses from the US 2014–2015 outbreak have an unusually long pre-clinical period in turkeys. BMC Vet Res 12:260

Stiver G (2003) The treatment of influenza with antiviral drugs. CMAJ 168(1):49–57

Suarez DL, Perdue ML, Cox N, Rowe T, Bender C, Huang J, Swayne DE (1998) Comparisons of highly virulent H5N1 influenza A virus isolated from humans and chickens from Hong Kong. J Virol 72:6678–6688

Tang XC, Zhou PP, Qiu WY (2010) On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. Chin Sci Bull 55:701–704. https://doi.org/10.1007/s11434-010-0045-2

Taubenberger JK, Morens DM (2006) 1918 influenza: the mother of all pandemics. Emerg Infect Dis 12(1):15–22. https://doi.org/10.3201/eid1201.050979

Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M et al (2013) New world bats harbor diverse influenza A viruses. PLoS Pathog 9(10):e1003657. https://doi.org/10.1371/journal.ppat.1003657

Wagner R, Matrosovich M, Klenk HD (2002) Functional balance between haemagglutinin and neuraminidase in influenza virus infections. Rev Med Virol 12:159–166

Wang J, Zhang Y (2006) Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation. Chem Phys Lett 423:50–53

Wei H (2010) The interaction between the 2009 H1N1 influenza a hemagglutinin and neuraminidase: mutations, co-mutations and the NA stalk motif. J Biomed Sc Engg 3:1–12

WHO (2009) Influenza A(H1N1) virus resistance to oseltamivir - 2008/2009 influenza season, northern hemisphere. http://www.paho.org/hq/images/stories/ad/hsd/cd/influenza/h1n120081230.pdf

WHO (2017). Recommended composition of influenza virus vaccines for use in the 2017–2018 northern hemisphere influenza season. http://www.who.int/influenza/vaccines/virus/recommendations/2017_18_north/en/

WHO (2018a) Influenza (Seasonal) Fact sheet. who.int. Updated to January 2018. http://www.who.int/mediacentre/factsheets/fs211/en/. Retrieved 12 Mar 2018

WHO (2018b) Recommended composition of influenza virus vaccines for use in the 2018–2019 northern hemisphere influenza season. http://www.who.int/influenza/vaccines/virus/recommendations/2018_19_north/en/

Wiesner I, Wiesnerova D (2010) 2D random walk representation of Begonia x_tuberhybrida multiallelic loci used for germplasm identification. Biol Plant 54:353–356

World Health Organization (2007) Ethical considerations in developing a public health response to pandemic influenza. WHO Press. WHO/CDS/EPR/GIP/20072

Xu R, Zhu X, McBride R, Nycholat CM, Yu W, Paulson JC, Wilson IA (2012) Functional balance of the hemagglutinin and neuraminidase activities accompanies the emergence of the 2009 H1N1 influenza pandemic. J Virol 86:9221–9232

Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC (2017) Using electronic health records and internet search information for accurate influenza forecasting. BMC Infect Dis 17 (1):332. https://doi.org/10.1186/s12879-017-2424-7

Yao Y-h, Nan X-y, Wang T-m (2006) A new 2D graphical representation—classification curve and the analysis of similarity/dissimilarity of DNA sequences. J Mol Struct THEOCHEM 764:101–108

Yau SS-T, Wang J, Niknejad A, Lu C, Jin N, Ho Y-K (2003) DNA sequence representation without degeneracy. Nucleic Acids Res 31:3078–3080

Yin D, Chen Y, Yau S-T (2014) A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. J Theor Biol 359:18–28

Zhang Z, Liu L, Li J, Zhang Z (2010) Spectral representation of DNA sequences and its application. IEEE 1023–1027. http://search.ror.unisa.edu.au/media/researcharchive/open/9915909466801831/53108451280001831. Accessed 22 Feb 2018

Zhao Y-B, Qi Z-H, Yang A-P (2015) Characterization and similarity analysis of DNA sequences considering codon degeneracy. Int J Hybrid Inf Technol 8(1):73–84

Chapter 10
Modern Approaches in Synthetic Biology: Genome Editing, Quorum Sensing, and Microbiome Engineering

Taj Mohammad and Md. Imtaiyaz Hassan

Abstract Synthetic biology is a new, emerging discipline which aims to design, engineer, and synthesize biological parts, devices, and systems to encompass medical therapeutics. This modern discipline integrates chemistry, biology, and engineering disciplines to enhance cell functionalities by incorporating genetic manipulations into biological cells and contributing toward better solution of current biomedical challenges, such as antibiotic resistance and cancer therapy. The advances in genetic manipulation techniques in microorganisms and development of biosynthetic units have opened new avenues to in-depth exploration of biological events such as cell response, cell death, metabolism, disease mechanism, molecular signaling, etc. Synthetic cells are being produced by genome editing and microbiome engineering and are extensively used as biofactories for drug and metabolites production and cell-based screening of targeted disease and phenotype. In this chapter, we outline the mechanisms and biomedical applications of three major components of synthetic biology including genome editing, quorum sensing, and microbiome engineering in detail.

Keywords Synthetic biology · Genome editing · Quorum sensing · Antibiotic resistance · Biomedical therapeutics · Microbiome engineering

10.1 Introduction

Synthetic biology is an emerging area of medical therapeutics employed to boost innovation in creating new biological tools for genetic engineering and developing new approaches for diagnosis, resulting safe and effective treatment of numerous diseases (Cachat and Davies 2011). With the recent advancements in the molecular

T. Mohammad · M. I. Hassan (✉)
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: mihassan@jmi.ac.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_10

biology, protein engineering, and genome editing tools synthetic biology has focused to develop programmable biological devices that can produce controlled and desired phenotypes in response to a given input, such as a molecular or electronic circuit (Kis et al. 2015). One of the main objectives of synthetic biology is to use well-characterized genetic circuits for engineering biological systems with novel functionalities (Trosset and Carbonell 2015).

The development of these custom-designed genetic circuits is not possible without advanced genome editing tools such as CRISPRs, a complement for modern genetic engineering approach. The focus of synthetic biology has also started toward adding new biological tools and approaches to the current technologies of genetic engineering, including modification of disease-related signaling pathways, gene therapy, immunotherapy, and cell therapy for the treatment of various complex diseases such as cancer (Weber and Fussenegger 2012). Furthermore, synthetic biology improves our understanding of disease mechanisms enabling the design of fast and accurate diagnostic tools and provides novel strategies for human therapeutics (Cheng and Lu 2012). It incorporates various approaches and technologies for DNA synthesis and assembly of DNA fragments for gene synthesis, even construction of synthetic chromosomes (Shen et al. 2016).

Quorum sensing is a process of gene expression regulation in response to the density changes of a bacterial population. It plays a vital role in development of biofilms which are often resistant to ultraviolet irradiation, desiccation, and antibiotic treatment (Miller and Bassler 2001). Thus, disruption or inactivation of the quorum-sensing system of bacteria can be used to affect biofilm development and differentiation for the treatment of several infectious diseases. These alterations in quorum-sensing mechanism can be made using synthetic biology approach or microbial engineering to reduce virulence of various pathogenic infections. Moreover, quorum sensing can be used as a decision-making process in any decentralized system for therapeutic purpose (Davis et al. 2015).

On the other hand, microbiome engineering is used to manipulate microbial genome to produce desired chemicals such as drugs and metabolites and to modify metabolic pathways for the treatment of metabolic disorders (Sheth et al. 2016). Here, we present the current scenario of synthetic biology with its development and advances in the field of genome editing, quorum-sensing mechanism, and microbiome engineering, including their medical applications and challenging issues. This chapter will cover a variety of approaches and applications of synthetic biology in regenerative medicine, disease modeling, drug discovery and production, as well as vaccine development for the treatment of infectious diseases, cancer, and other metabolic diseases.

10.1.1 Synthetic Gene Networks and Biological Circuits

Synthetic gene networks can be used in diagnosis and for therapeutic purposes. These synthetic gene networks and engineered biological on-off switches work

analogous to the electronic circuits and have been implicated to control cell functions. Programmable logic gates are implemented to these biological circuits to perform and control logical operations in a synthetic or biological gene network of a target cell (Friedland et al. 2009). These genetic circuits are used to control cellular behavior that changes gene expression, growth rates, and cell viability for diagnosis and detect cell line lineages. An engineered tunable dual-promoter integrator can target cancer cells which expresses effector gene when input promoters shows high level of activity. In several scientific attempts, synthetic gene networks have been developed to regulate blood glucose homeostasis to control cytokine expression and to correct insulin deficiency under the control of radio waves because it can sense uric acid concentrations in the blood (Lu et al. 2009). Furthermore, by controlling regulators in MAPK pathway, a synthetic gene network may regulate the signaling pathway, developed for therapeutic applications.

10.2 Genome Editing

Genome editing is a genetic engineering approach, used to alter genetic material in a cell in an appropriate manner to get the desired phenotype. In this technique, DNA is inserted, removed, replaced, or modified at a specific site within the genome to change the characteristic features and function of a living organism or a cell. Genome editing technique is widely used in the biomedical therapy to treat life-threatening diseases (Wood et al. 2011). Previously, genetic engineering methods were restricted to gene insertion or mutation at random or at a small number of specific sites in the genome. Such gene editing techniques lack efficiency, specificity, and reliability. Recent advances in the synthetic biology genome editing tools and higher computational power have significantly led to develop engineer cells or organism in a desired fashion.

Until recently, a novel genome editing tool has been developed, namely, CRISPR system which is a more precise, flexible, efficient, and safe method (Mali et al. 2013). Other genome editing tools are rAAV, TALENs, ZFNs, and the CRISPR/Cas system, used for making a desired modification in the genome (Maeder et al. 2013). The rAAV is a viral-based genome editing tool in which a viral vector is used to carry DNA molecule(s) into the cell, while TALENs, ZFNs, and CRISPRs used recombination-based approaches (Gaj et al. 2013). These are nuclease-based genome editing tools, used to engineer nucleases to cut the genome at a desired position. The CRISPR-Cas9 genome editing in eukaryotic cells was first reported in 2013 (Cong et al. 2013). Now, this technique has emerged tremendously. The products of genome editing and modifications are isogenic cell lines which only differ by the modifications we have introduced. These tools are also used to engineer human embryonic stem cell, induce pluripotent stem cell and erythroid cell lines, and to make knockout *C. elegans*, mice, rats, zebrafish, etc. Furthermore, these methods are also used to generate knockin organisms, for instance, Sp110 knockin cattle with increased tuberculosis resistance (Gibson et al. 2008).

Engineered microbes developed through genome editing methods, are being used for biosynthesis of various small molecules such as drugs and metabolites. Many microorganisms and phages are ideal for such biological designing, modeling, genome-scale engineering, genome editing, and synthesis, to get a desirable phenotype. In near future, use of these model organisms such as *E. coli*, *S. cerevisiae*, and bacteriophages will enable more genome engineering to the target organism for medical therapeutics and other biological advancements. One of the main applications of genetic editing in medical therapy is to correct the genetic errors responsible for a disease such as sickle cell anemia, xeroderma pigmentosum, epidermolysis bullosa, and various types of cancer (Maeder and Gersbach 2016).

Moreover, the breed transformations such as wild cattle (aurochs) into dairy cattle (Holsteins), wolves into Chihuahuas and Great Danes, and teosinte into maize are beautiful examples of genome-level engineering (Belhaj et al. 2013). Genome-level engineering and synthetic biology have paved the way to engineer and develop customized organism for human welfare. Editing a single gene is a major step toward engineering the whole genome. By developing and analyzing synthetic forms of biological systems, we can understand the obstacles inflicted by the complication of developed systems. Recent development in genomics and genome editing has pledged to play a major role in rational design in biological engineering and synthetic biology which offer new openings for scientific community, who are working to develop new biological systems with the desired function. Various types of genomic alterations and major editing tools in genome editing are illustrated in Fig. 10.1.

Genome editing has many exciting advances and potential applications in the medical therapeutics to improve human health. One of the exciting applications of genome editing in medical field is engineering stem cells for repairing the damaged tissues and organs, for instance, genetically engineered stem cells are being used to treat neurodegenerative diseases such as amyotrophic lateral sclerosis (Kumar et al. 2016), joint injury, arthritis, regenerating cardiac tissues, etc. Another interesting application of this technique is to study the disease mechanisms through patient-derived stem cells. CRISPR technology is one of the wonderful examples of this, which is used to introduce genetic modifications and gene expression modulation in stem cells for any gene to provide a comprehensive understanding of the disease mechanism (Shalem et al. 2014). By the use of genome editing, scientists are working to create "universal" stem cells that can be used in any patient without immune rejection for restoring and repairing damaged tissues and organs. It would not be impractical to say that commercialized "cell drugs" can be produced using universal stem cells, which can regenerate any organ in any patient (Lawman and Lawman 2005).

In the future, we hope that the current therapeutic strategy and approaches will be purely based on genome transplantation, stem cell editing, or direct editing of patient's somatic cells including cancerous cells. The genome editing tools will further enable us to better review the genetic alterations that occur with a particular disease (Bortesi and Fischer 2015). Although genome editing is not a standalone system, it should be combined with the other technologies, such as gene delivery,

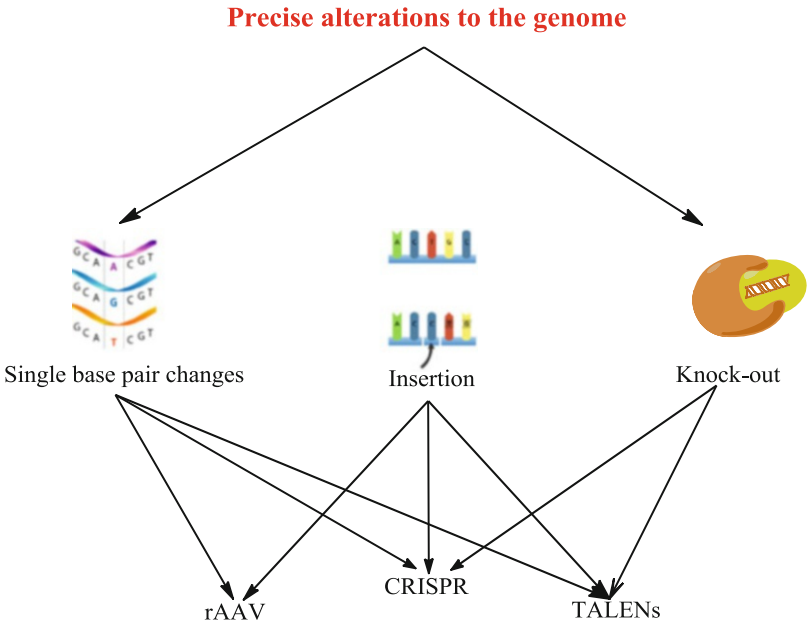

Fig. 10.1 Various types of genome editing tools in synthetic biology

cell engineering, immunotherapy, systems biology, bioinformatics, and bioprocess engineering for complete solution.

10.3 Quorum Sensing

Quorum sensing is a process in which bacteria communicate via secreting and sensing small diffusible signaling molecules called "autoinducers," to regulate the gene expression and to coordinate behaviors in response to the density of a population (Miller and Bassler 2001). It occurs within a single species as well as between different bacterial species, where the bacterial population as a whole can make a coordinated response (Fuqua et al. 1994). Various social insects such as ant, termites, and bees use quorum sensing to make decisions about destination. First quorum-sensing mechanism was discovered in *Vibrio fischeri*, a luminous marine bacterial species, where the light emission was determined only at high bacterial population density in response to the accumulation of secreted signaling molecules called as autoinducers. Diagrammatic representation of a typical *Vibrio fischeri* LuxI/LuxR quorum-sensing circuit is shown in Fig. 10.2.

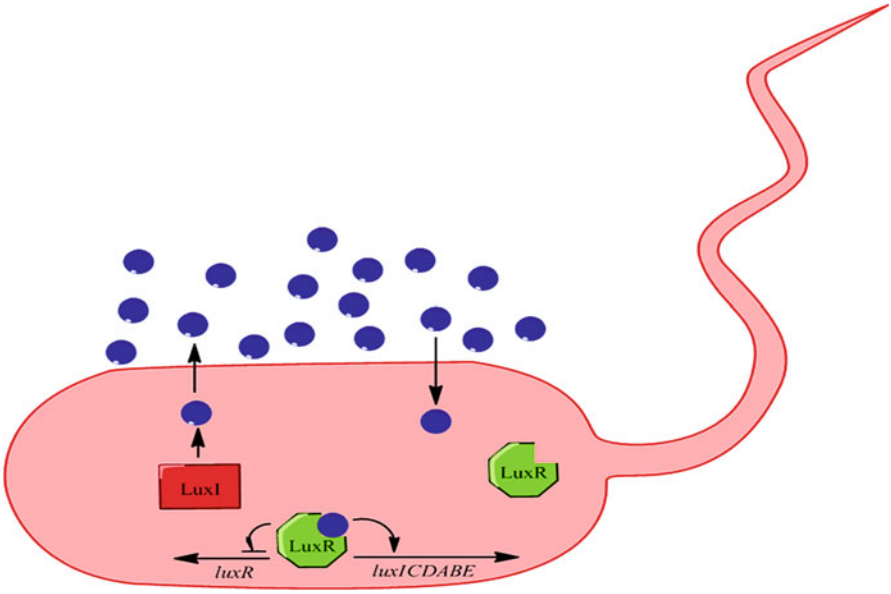

Fig. 10.2 Diagrammatic representation of *Vibrio fischeri* LuxI/LuxR quorum-sensing circuit. Five luciferase structural genes (luxCDABE) and two regulatory genes (luxR and luxI) are required for the quorum-sensing mechanism which control light emission in *Vibrio fischeri*. luxR is transcribed to the left, and the luxICDABE operon is transcribed to the right. The HSL autoinducer *N*-(3-oxohexanoyl)-homoserine lactone (hexagons) is synthesized by LuxI protein (shown in red square). As the bacterial population density increases, the concentration of the signaling molecules (autoinducer) increases. When autoinducers rapidly increase with the population and exceed a threshold level, the LuxR protein binds the autoinducer. The LuxR-autoinducer complex binds at the luxICDABE promoter and activates transcription of this operon which results in growth in autoinducer synthesis and an exponential increase in light.

10.3.1 General Mechanism

Autoinducers or pheromones and a receptor specifically detect the signaling molecules which are essential for quorum-sensing mechanism. These autoinducers or pheromones are signaling molecules secreted from bacteria and released into their surrounding environment to bind to signaling receptors present on the bacterial surface or in the cytoplasm (Waters and Bassler 2005). When these signaling molecules exceed a threshold concentration level, they stimulate quorum-sensing genes in bacteria that enable them to behave as a multicellular population rather than as an individual single-celled organism. Autoinducer/receptor complex binds to DNA promoters and activates the transcription of bacterial quorum-sensing genes. Thus, every bacterium in a group benefits from the activity of the entire group of bacterial population (Miller and Bassler 2001).

Quorum-sensing systems are different in both Gram-negative and Gram-positive bacteria. In Gram-negative bacteria, acylhomoserine lactones (AHLs) are used as the

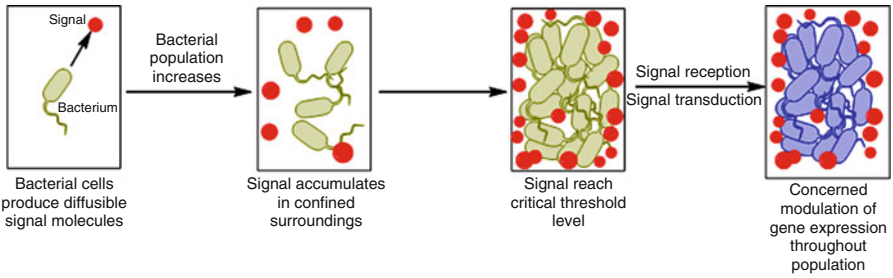

Fig. 10.3 A typical way of developing quorum-sensing mechanism in bacteria

signaling molecules or autoinducers which are secreted in bacteria and bind to AHL receptors in the cytoplasm of the bacteria (Whitehead et al. 2001). When these AHLs rapidly increase and exceed a threshold level, the autoinducer/receptor complex in the cytoplasm acts as a DNA-binding transcriptional activator. However, in Gram-positive bacteria, a peptide-mediated quorum-sensing occurs. Typically eight to ten amino acid long peptides are used as autoinducers. These oligopeptides are secreted from ATP-binding cassette transporter (ABC transporter) which cannot diffuse in and out of bacteria like AHLs and bind to autoinducer receptors on the bacterium surface. When oligopeptides reached a threshold level in the surrounding environment, the binding of the oligopeptide to its receptor activates response regulators, the DNA-binding transcriptional regulatory proteins which lead to the switch on of quorum-sensing genes and a coordinated population response (Miller et al. 2002). A typical way of developing quorum-sensing mechanism in bacteria is illustrated in Fig. 10.3.

Bacteria are highly interactive and demonstrate various social activities such as conjugal plasmid transfer, swarming motility, biofilm maturation, antibiotic resistance, and virulence and regulate various phenotypes. Major applications of quorum sensing in different fields are depicted in Fig. 10.4. Bacteria use signaling molecules which are released into the environment to communicate and measure the concentration of the molecules and cell density in a bacterial population. However, many important criteria are required to be a quorum-sensing signaling mechanism which includes the following: (i) the production of quorum-sensing signal should happen during a particular growth phase or in response to specific environmental changes, (ii) signal should be recognized by a specific bacterial receptor in the extracellular environment, and (iii) gathering of a critical threshold concentration of the quorum-sensing signal should stimulate a coordinated reaction.

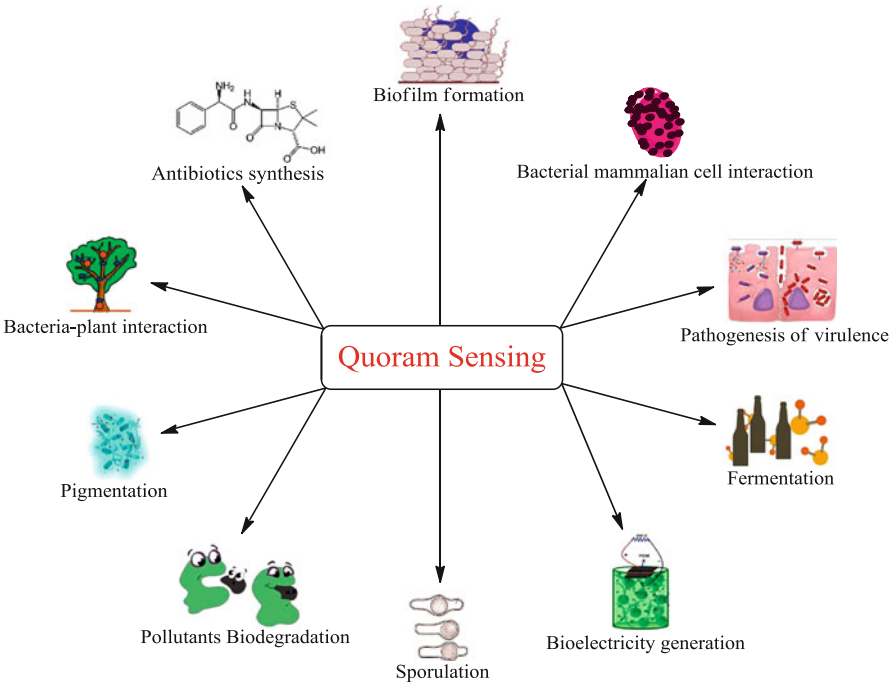

Fig. 10.4 Major applications of quorum sensing

10.3.2 Application of Quorum Sensing for Medical Therapeutics

Quorum sensing is required for the virulence of various pathogens such as *S. aureus*, *P. aeruginosa*, *Burkholderia pseudomallei*, *Burkholderia cenocepacia*, and *Vibrio cholerae*. Alteration in quorum sensing can reduce virulence in several animal hosts such as mice and nematodes. Quorum sensing also plays a vital role in the development of biofilms which are often highly resistant to ultraviolet irradiation, desiccation, and antibiotic treatment. Disruption or inactivation of the quorum-sensing system of bacteria can be used to affect biofilm development and differentiation for the treatment of several infectious diseases (Hentzer and Givskov 2003).

Quorum sensing can also be used as a decision-making process in any decentralized system for therapeutic purpose. Figure 10.5 is illustrating the formation of bacterial biofilm using quorum-sensing mechanism. Usually, antibiotics are used for the treatment of bacterial infections, based on the approach to kill or inhibit microbial growth. In this approach, bacterial resistance to antibiotics is a major concern. New pharmacological strategy is needed to overcome this issue in order to develop safe, fast, and effective therapy for microbial infection. A new approach called quorum-sensing inhibition offers a novel opportunity to develop antipathogenic drugs with strategy of inhibiting autoinducers of bacterial quorum-

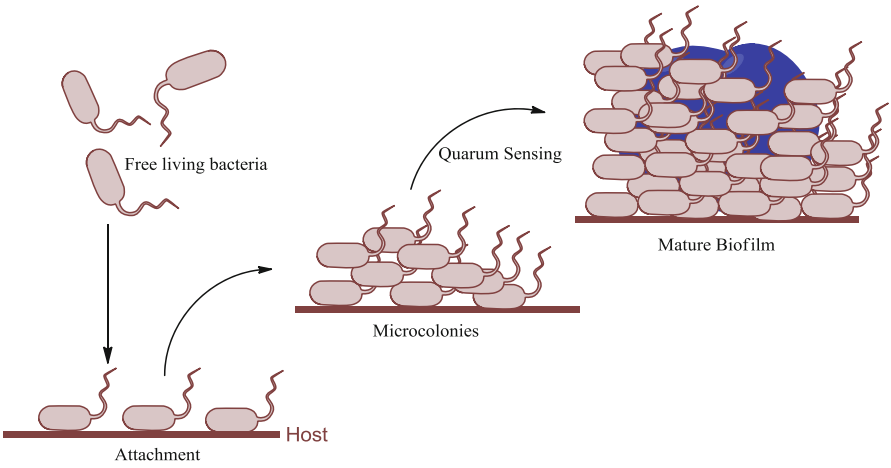

Fig. 10.5 Formation of biofilm using quorum-sensing mechanism

sensing systems. Signaling molecules called autoinducers such as autoinducer 2 (AI-2) and acylhomoserine lactones (AHLs) are being used in bacterial quorum-sensing mechanism for directed antipathogenic drug design. The inhibitors of quorum-sensing mechanism act as anti-biofilm which might act as potent antibacterial drugs (Balaban et al. 2007).

Quorum-sensing inhibition approach waves new therapeutics by interrupting the pathway of bacterial communication to regulate expression of virulence factors (Miller et al. 2002). Many plant extracts act as an inhibitor which disrupts biofilms by quenching the quorum-sensing system. These molecules inhibit quorum sensing by either disturbing AHLs activity or degrading the receptors for the AHLs. Thus, using quorum-sensing mechanism, bacteria can communicate and regulate their virulence gene expression. New therapeutic strategies targeting quorum sensing can be implicated to control virulence mechanism of bacteria. Quorum-sensing inhibitor is a better approach over antibiotics to treat several infections, because these inhibitors are able to inhibit bacterial pathogenesis and are active on both diving and replicating cells and have no side effect on host.

RNA III-inhibiting peptide (RIP), a quorum-sensing inhibitor, is an elegant example which controls *Staphylococcus aureus* virulence by inhibiting the production of autoinducers without affecting the host system. The existing quorum-sensing inhibitors such as penicillin acid, garlic extract, cyclic sulfur compounds, halogenated furanones, and 4-nitro-pyridine-*N*-oxide have been found to inhibit the expression of quorum-sensing-dependent virulence factor in *Pseudomonas aeruginosa.* Thus, synthetic inhibitors can be engineered against quorum sensing in bacteria for therapeutic approach which can be used as potential drugs for a number of infectious diseases (Ramage et al. 2002). Although the current study and research on quorum-sensing mechanism are still in its infancy, these novel antibacterial strategies could

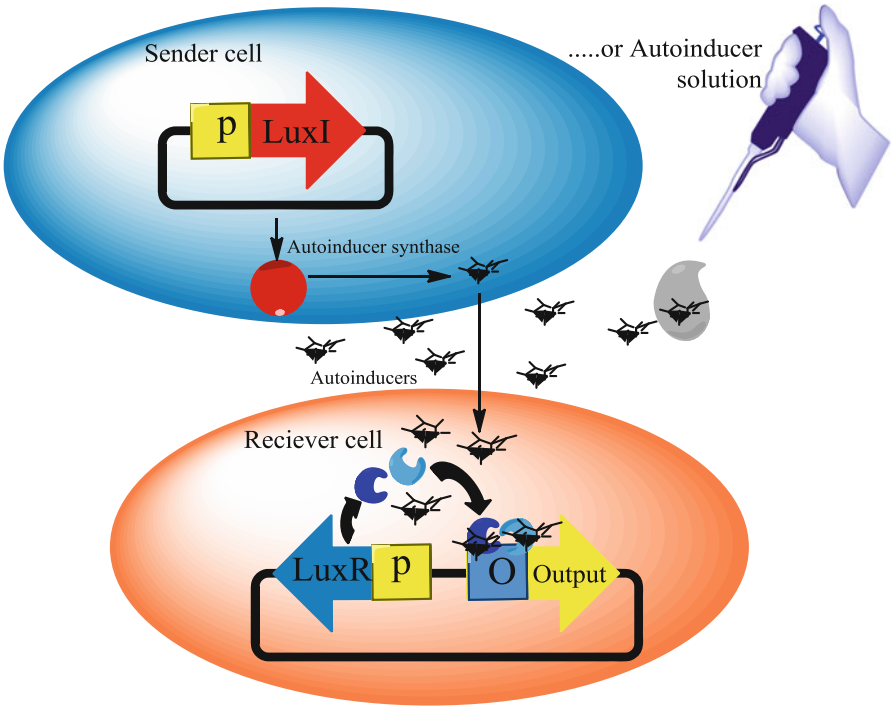

Fig. 10.6 Engineered quorum-sensing network to control the expression of any gene of interest in a synthetic system. *O* "operator"-binding site for the regulator protein, *p* constitutive promoter

be implicated for biomedical engineering with enormous therapeutic applications (Khan et al. 2009).

Another interesting outcome of biomedical engineering is the design and development of synthetic quorum sensing and genetic circuits for medical therapeutics (McAdams and Arkin 1998). The constituents of natural quorum-sensing system are used to design and develop synthetic quorum-sensing systems using Boolean network modeling that regulates gene expression in response to chemical signals (Hentzer and Givskov 2003). These synthetic circuits can be used to design various complex genetic systems for various biomaterials production and medical therapies. A basic topology of synthetic communication pathway that can generate switch-like cellular responses, to regulate various enzymatic and transcriptional functions, is shown in Fig. 10.6.

10.4 Microbiome Engineering

Currently, we are not far from that time when we will be able to grow genetically engineered "smart microbes" in our gut that can diagnose and treat diseases at the earliest stage. These smart bacteria are possible to design and synthesize using synthetic biology approaches (Sonnenburg 2015). Human microbiome is the microbial communities that colonize our body for survival. This is an essential system for human immunity or pathogenic defense and plays significant roles for proper health. It helps in providing essential vitamins, preventing pathogen growth and stimulating tissue development, and plays an essential role in metabolism, immunity, inflammation, and tumor macroenvironment (Ruder et al. 2011). Engineering and manipulating this microbiome called microbiome engineering, an approach promoting a great platform in the field of synthetic biology, seeks to engineer various microorganisms such as bacteria so that they can perform the desired function (Mueller and Sachs 2015).

Engineering human microbiome is a novel concept, where the genome modification of human microflora is carried out to achieve a desired goal. Basically, microbiome engineering is an art of engineering microorganisms to perform particular metabolic or physiological functions at the customize level to develop novel therapeutic strategies for medical applications. In this section, we attempt to describe the potential applications, challenges, and limitations of microbiome engineering particularly in relation to human health (Kali 2015). Here, we discuss how microbiome engineering is being applied to modify various microbiomes using various strategies and approaches of synthetic biology. Various microbiome activities and their medical applications are listed in Table 10.1.

Table 10.1 List of microbiome activity and their medical application

Microbiome activity/attribute	Therapeutic applications
Compete with diseased microbes for nutrients	Colonization resistance, prevention, or treatment of infectious diseases
Formation of biofilm	
Production of bacteriocin and other inhibitory molecules	
Alteration of vaginal pH	
Supplement nutrition and metabolism	Obesity, diabetes, malnutrition, and dyslipidemia
Regulation of immune response	Hypersensitivity and autoimmune diseases
Instruction to immune system	Prevention of cancers and several immune disorders
Anti-inflammatory compounds	Inflammatory bowel disease and other inflammations
Production of signaling molecules	Biomarkers for diagnosis
Modulation of cancer macroenvironment	Prevention of malignancies or cancer
Production of neuroactive microbial metabolites "psychobiotics"	Prevention or treatment of depression, anxiety, and autism

One of the main applications of microbiome engineering is modifying probiotic bacteria for the treatment of various rare metabolic diseases caused by protein deficiencies such as Gaucher's disease as well as infectious diseases, diabetes, and cancer (Turnbaugh et al. 2007). These engineered bacteria produce metabolites for the development of safe and effective drugs which are known as synthetic biotic medicines. Such drug, named SYNB1020, has been entered in the phase one trial for hyperammonemia (Quiza et al. 2015). Similarly, AZT-01, developed from recombinant probiotic bacteria *Staphylococcus epidermidis*, produces filaggrin, a protein defective in patients with eczema, or desired therapeutic proteins to deliver directly into the skin for the treatment of skin diseases ranging from eczema to staph infections. The Topgenix company has developed engineered probiotic bacteria to express a UV-protective compound for delivery to the skin. These bacteria can also be used to produce compounds of interest for therapeutic purpose.

Xycrobe, a form of live microbes, is a wonderful example of engineered bacteria which can penetrate the upper layer of dead skin to secrete and deliver biotherapeutics such as antioxidants, anti-inflammatory compounds. In this way, engineered skin microbiome can be used for the treatment of various skin diseases such as acne, dermatitis, and psoriasis. The biotherapeutics produced by these engineered bacteria can be delivered directly to the site which is used for site-specific treatment of diseases to avoid side effects, and it is especially important for vaccination for safe and effective uptake. ActoBiotics are also being developed by engineering *Lactobacillus* to orally deliver therapeutic peptides and proteins. These are in clinical phage by Intrexon Inc. in two different forms as AG013 for oral mucositis and AG014 for inflammatory bowel disease (Sonnenburg 2015).

Osel Inc. is also working in the field of microbiome engineering that modulates the natural microbiome of the human body to treat several bacterial infections. It mainly focusing on vaginal microbiome to improve woman's health to treat several infectious diseases such as urinary tract infections and bacterial vaginosis. *Lactobacillus* maintains a low pH of 4.0–4.5 and produces hydrogen peroxide to help in maintaining proper vaginal health by preventing pathogenic infections in the vaginal tract. LACTIN-V, a biotherapeutic product developed by Osel, produces various substances to protect and maintain a healthy microbiome present in vaginal tract and to adhere vaginal epithelial cells and antagonizes uropathogenic *E. coli* to prevent most of the urinary tract infections. Similarly, vaginal *Lactobacillus* has been engineered to produce potent inhibitors of HIV (Foo et al. 2017).

A different approach is microbiome deletion approach instead of addition, as most of the approaches uses addition of engineered microbes to the existing microbiome for therapeutic applications. In this approach, bacteriophages are used as an alternate to antibiotics to address bacterial infections. Bacteriophages are the specific viruses which can infect to the particular bacteria, meaning, the bacteriophage for *Salmonella* infects only the *Salmonella*, and the bacteriophage for *Mycobacterium tuberculosis* will only infect *tuberculosis*. This strategy is using natural specificity of bacteriophages to target the specific bacterial strains within the microbiome which is helpful to fight with the antibiotic-resistant strain. These engineered phages can be used to selectively eliminate pathogenic microbes. Lung

microbiome is being studied to understand the dynamics of the lung microbiome for identification of targets to treat various lung diseases such as cystic fibrosis (Krom et al. 2015).

Another remarkable strategy of microbiome engineering is programming microbes to produce signaling inhibitor such as LED209 to prevent phosphorylation of certain kinases that activates expression of virulence factors in pathogens, so that pathogenic infections can be prevented. Similarly, the oral microbiome has also been engineered to specifically eliminate *Streptococcus mutans* which is responsible for dental decay in human. Apart from that, a wonderful approach is engineering microbes by manipulating specific genetic circuits that can be used to activate or deactivate particular gene and pathways in various metabolic diseases and complex diseases such as cancer (Foo et al. 2017).

In a similar way, bacteria can be engineered to secrete anti-inflammatory molecules when inflammation is detected and automatically stop when the inflammation ends. In addition to treating inflammation, these smart bacteria can ultimately help in attacking infectious organisms, early diagnosis of cancer, diarrhea, human behavior, etc. These microorganisms also guide our immune system, metabolism, and even our mood and activities. Even abnormal behavior such as depression, anxiety, and autism in human can be treated by engineering gut flora to produce neuroactive microbial metabolites described as "psychobiotics" to influence the brain (Schmidt 2015). Hence forward, engineered microbiota can be used to express various protective antigens when desired by the immune system.

With the rapid growth and advancement in medical science and technology, it is not impractical to think to develop a bacterium that can recognize tumor antigens as receptors and specifically kill only cancerous cells without affecting normal tissues. Furthermore, these microbes are also being reprogrammed to secrete various signaling molecules that can be served as biomarkers for a particular disease (Gilbert et al. 2016). Microbiome engineering has modernized the insight into disease and its mechanism by manipulating microbiota in a plethora of ways to improve human health (Pflughoeft and Versalovic 2012). Apart from that, the development of clinically relevant biosensors and robust genetic circuits is a great achievement in the field of synthetic biology, which can be used to develop fully autonomous living therapies based on microbiota therapeutics. Various successful outcomes of microbiome engineering that have already been implemented in various scientific attempts are represented in Fig. 10.7.

All the discussed examples indicate great potential of microbiome engineering in medical therapeutics. But this field is still in its infancy and has great potential waiting to be released. Synthetic biology is a major contributor to the microbiome engineering field that plays an important role in advancement of microbiome engineering, using which various smart microbes with novel functionalities can be engineered. Installing genetic circuits into bacteria or other microbial community is an admirable application of synthetic biology in the field of microbial engineering to perform diagnostic and therapeutic operations. Microbiome engineering has gained incredible development in the previous decades by means of synthetic biology to

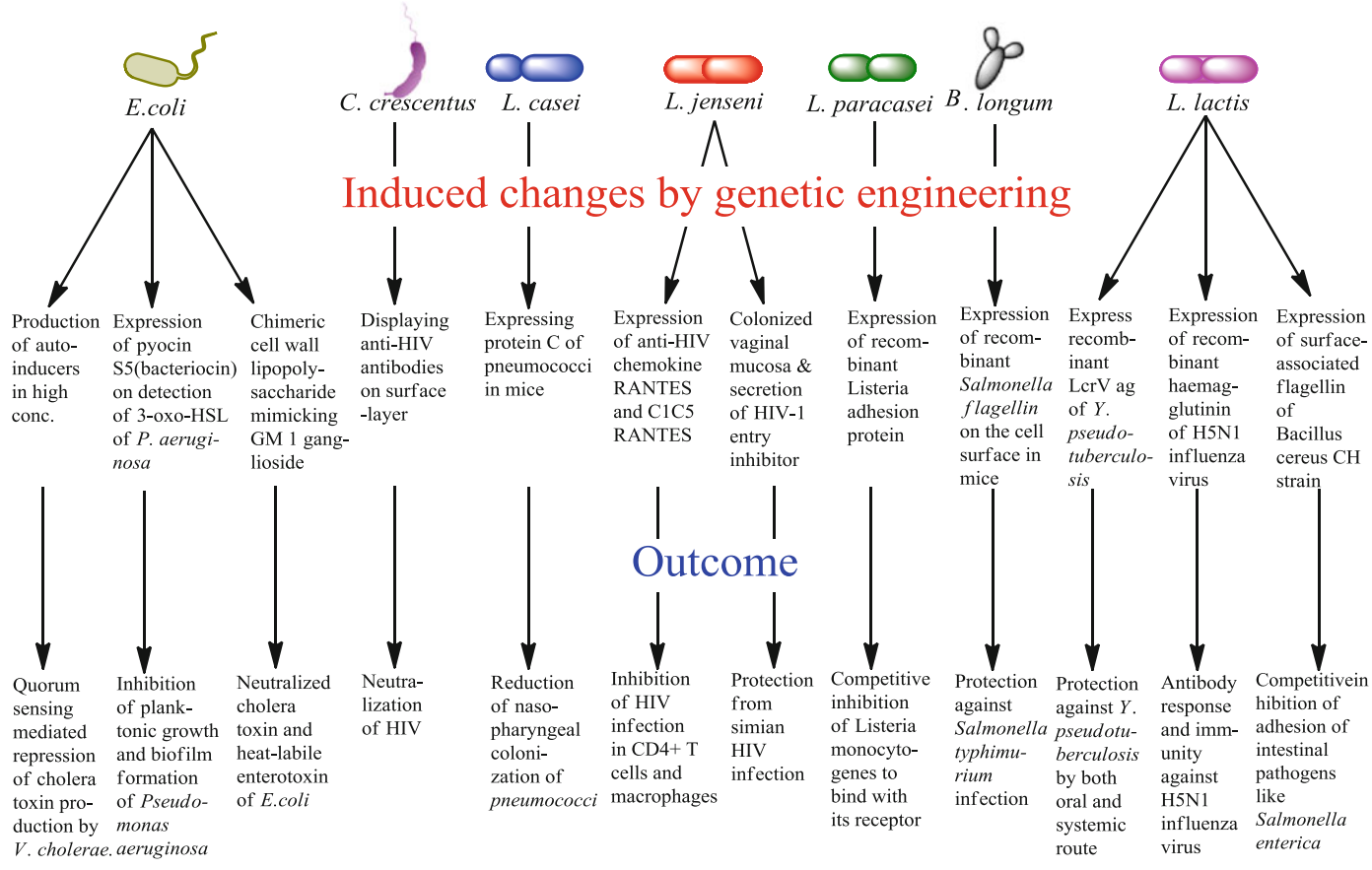

Fig. 10.7 Various successful outcomes of microbiome engineering

provide deeper insights into human microbiome and speed up microbiome engineering efforts.

10.5 Conclusion and Future Directions

Synthetic biology plays a major role in reorienting the field of modern biological research and medical therapeutics. A great success of today's synthetic biology is the successful development of synthetic cells which can act as a biofactory for production of various bioactive products such as artemisinin and several innovative compounds for pharmaceutical use. It also provides a platform for laboratory in which therapeutic target or signaling pathways can be tested. Synthetic cellular models can also be used in investigation of disease mechanisms. Modern genome engineering approaches can potentially correct various genetic diseases as well as different infectious diseases; as well as metabolic, immunological, and neurological disorders; and cancer in a safe and in an effective manner. The genome editing tools will enable us to better review the genetic alterations that occur with a particular disease. Designing and development of synthetic quorum sensing using genetic circuits is a great strategy to study cell-cell communication within bacterial population to understand drug resistance mechanisms for therapeutic purpose. It would not be impractical to expect a universal therapy-oriented gene network in the near future for clinical uses. Microbiome engineering demonstrates promising hopes for therapeutic applications to improve human health. Furthermore, novel tools and technologies of synthetic biology are being developed which provide deeper insights into microbiome and its activity and mechanisms. Manipulating microbiome by various engineering approaches will be a mainstream strategy to improve human health. In this respect, synthetic biology holds great promise for future treatment of various diseases. However, further work is needed to implement all these approaches clinically, but today, synthetic biology remains one of the most promising and exciting fields of science and technology in modern era.

References

Balaban N et al (2007) Treatment of Staphylococcus aureus biofilm infection by the quorum-sensing inhibitor RIP. Antimicrob Agents Chemother 51:2226–2229

Belhaj K, Chaparro-Garcia A, Kamoun S, Nekrasov V (2013) Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. Plant Methods 9 (39)

Bortesi L, Fischer R (2015) The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnol Adv 33:41–52

Cachat E, Davies J (2011) Application of synthetic biology to regenerative medicine. J Bioeng Biomed Sci S 2:003

Cheng AA, Lu TK (2012) Synthetic biology: an emerging engineering discipline. Annu Rev Biomed Eng 14:155–178
Cong L et al (2013) Multiplex genome engineering using CRISPR/Cas. Syst Sci 339:819–823
Davis RM, Muller RY, Haynes KA (2015) Can the natural diversity of quorum-sensing advance synthetic biology? Front Bioeng Biotechnol 3
Foo JL, Ling H, Lee YS, Chang MW (2017) Microbiome engineering: current applications and its future. Biotechnol J
Friedland AE, Lu TK, Wang X, Shi D, Church G, Collins JJ (2009) Synthetic gene networks that count. Science 324:1199–1202
Fuqua WC, Winans SC, Greenberg EP (1994) Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. J Bacteriol 176:269
Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31:397–405
Gibson DG et al (2008) Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome. Science 319:1215–1220
Gilbert JA et al (2016) Microbiome-wide association studies link dynamic microbial consortia to disease. Nature 535:94–103
Hentzer M, Givskov M (2003) Pharmacological inhibition of quorum sensing for the treatment of chronic bacterial infections. J Clin Investig 112:1300
KAli A (2015) Human microbiome engineering: the future and beyond. J Clin Diagn Res: JCDR 9: DE01
Khan MSA, Zahin M, Hasan S, Husain FM, Ahmad I (2009) Inhibition of quorum sensing regulated bacterial functions by plant essential oils with special reference to clove oil. Lett Appl Microbiol 49:354–360
Kis Z, Pereira HSA, Homma T, Pedrigi RM, Krams R (2015) Mammalian synthetic biology: emerging medical applications. J R Soc Interface 12:20141000
Krom RJ, Bhargava P, Lobritz MA, Collins JJ (2015) Engineered phagemids for nonlytic, targeted antibacterial therapies. Nano Lett 15:4808–4813
Kumar V, Islam A, Hassan MI, Ahmad F (2016) Therapeutic progress in amyotrophic lateral sclerosis-beginning to learning. Eur J Med Chem 121:903–917. https://doi.org/10.1016/j.ejmech.2016.06.017
Lawman M, Lawman P (2005) Universal stem cells. Google Patents
Lu TK, Khalil AS, Collins JJ (2009) Next-generation synthetic gene networks. Nat Biotechnol 27:1139–1150
Maeder ML, Gersbach CA (2016) Genome-editing technologies for gene and cell therapy. Mol Ther 24:430–446
Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK (2013) CRISPR RNA-guided activation of endogenous human genes. Nat Methods 10:977–979
Mali P et al (2013) RNA-guided human genome engineering via Cas9. Science 339:823–826
McAdams HH, Arkin A (1998) Simulation of prokaryotic genetic circuits. Annu Rev Biophys Biomol Struct 27:199–224
Miller MB, Bassler BL (2001) Quorum sensing in bacteria annual reviews in. Microbiology 55:165–199
Miller MB, Skorupski K, Lenz DH, Taylor RK, Bassler BL (2002) Parallel quorum sensing systems converge to regulate virulence in Vibrio cholerae. Cell 110:303–314
Mueller UG, Sachs JL (2015) Engineering microbiomes to improve plant and animal health. Trends Microbiol 23:606–617
Pflughoeft KJ, Versalovic J (2012) Human microbiome in health and disease. Annu Rev Pathol: Mech Dis 7:99–122
Quiza L, St-Arnaud M, Yergeau E (2015) Harnessing phytomicrobiome signaling for rhizosphere microbiome engineering. Front Plant Sci 6
Ramage G, Saville SP, Wickes BL, López-Ribot JL (2002) Inhibition of Candida albicans biofilm formation by farnesol, a quorum-sensing molecule. Appl Environ Microbiol 68:5459–5463

Ruder WC, Lu T, Collins JJ (2011) Synthetic biology moving into the clinic. Science 333:1248–1252
Schmidt C (2015) Mental health: thinking from the gut. Nature 518:S12–S15
Shalem O et al (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343:84–87
Shen Y et al (2016) SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. Genome Res 26:36–49
Sheth RU, Cabral V, Chen SP, Wang HH (2016) Manipulating bacterial communities by in situ microbiome engineering. Trends Genet 32:189–200
Sonnenburg JL (2015) Microbiome engineering. Nature 518:S10–S10
Trosset J-Y, Carbonell P (2015) Synthetic biology for pharmaceutical drug discovery. Drug Des Devel Ther 9:6285
Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI (2007) The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449:804
Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. Annu Rev Cell Dev Biol 21:319–346
Weber W, Fussenegger M (2012) Emerging biomedical applications of synthetic biology. Nat Rev Genet 13:21–35
Whitehead NA, Barnard AM, Slater H, Simpson NJ, Salmond GP (2001) Quorum-sensing in gram-negative bacteria FEMS microbiology reviews. FEMS Microbiol Rev 25:365–404
Wood AJ et al (2011) Targeted genome editing across species using ZFNs and TALENs. Science 333:307–307

Chapter 11
Synthetic Probes, Their Applications and Designing

Shafaque Zahra, Ajeet Singh, and Shailesh Kumar

Abstract Microbial genomics is becoming an emerging field of science that analyses and compares complete genome of microorganisms or zillions of genes, in a concomitant fashion. Cost-effective and high-throughput next-generation sequencing (NGS) has made possible to explore highly diverse microbial community for metagenomics, medical diagnostics and clinical microbiological research. The probes, which act like biosensors, are commonly exploited for therapeutics, phylogenetic analysis and common medical diagnostic techniques. Synthetic nucleic acid analogues are replacing nucleic acids because of greater stability and efficiency in in vivo applications and molecular biology research. These artificial nucleic acid analogues like LNAs (locked nucleic acids) and PNAs (peptide nucleic acids) are being exploited in research and diagnostics, and continuous efforts are being made to engineer them further for their maximal potential. A number of tools are available for probe designing for various applications. This chapter focuses on the natural and artificial nucleic acid probes, the differences in the chemistry of these probes, their advantages, limitations and synthetic applications along with some web-based tools for probe designing and quality check.

Keywords Hybridization · NGS · Microbial genomics · Oligonucleotides · Omics tools · Probes

11.1 Introduction

The plethora of microbial world has resulted in enormous diversity in the genome sequences and thus leading to functional diversification. Differences in the sequence and structure of genomes within a microbial population reflect the amalgamation of mutation, recombination and selection. With plenty of genome sequences available because of high-throughput sequencing technology, these effects have been more

S. Zahra · A. Singh · S. Kumar (✉)
National Institute of Plant Genome Research (NIPGR), New Delhi, India
e-mail: shailesh@nipgr.ac.in

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_11

thoroughly characterized and more extensively utilized for comprehensive understanding of origin, phylogeny, and evolution of microorganisms and their intricate interactions with other living organisms including humans.

Microbial genomics has become an effective tool for a wide range of applications including epidemiologic investigations, forensics, diagnostics and vaccine development (Relman 2011). The genomic sequences provide an insight into the population structure and evolutionary history for epidemiologic investigations and could also be used to develop new diagnostic tests and cultivation methods, new targets of drug development and antigens for vaccine development.

Applications of microbial genomics have expanded researchers' appreciation for the biology of microorganisms including their organismal, metabolic and environmental diversity. These are providing important insights into the 'ground rules' of pathogen evolution and will inform the development of a 'versatile platform for developing new responses to infectious disease' (Lederberg 2000). Genome sequence analysis can provide new insights for phylogenetic studies, population structure, protein expression, virulence, growth requirements and diagnostic research.

Real-time PCR assays, in situ hybridization and microarrays are being extensively used for genomic analysis and common medical diagnostic technologies. For these assays, probes are being commonly used for analysis and research.

This chapter has two portions; the first part focuses on concept of probes, chemistry of artificial nucleic acid probes and applications, and, in the latter part, some selected omics tools, viz. Picky, OligoWiz, Teolenn, mathFISH and ProbeMaker, for oligonucleotide probe designing and evaluation for quality assessment are discussed along with their workflow and limitations.

11.2 Nucleic Acid Probes

A probe is nucleic acid-based biosensor (Electrochemical nucleic acid – based biosensors. . .) comprised of labelled sequences of nucleic acids (single-stranded DNA or RNA) or their synthetic analogues which can bind to its target complementary sequences; thus they are useful for the detection of sequences from a pool of nucleic acids containing non-complementary DNA/RNA along with its complementary counterpart.

There are two types of nitrogenous bases: purines and pyrimidines. Purines include adenine (A) and guanine (G), and pyrimidines include thymine (T), cytosine (C) and uracil (U). 'A' forms hydrogen (H)-bond with 'T/U' and 'G' forms H-bond with 'C'. The H-bond is responsible for formation of stable, double-helical strand, and each strand in the helix acts as a template for complementary strand. This forms the basic principle for functional probe under optimal conditions. The nucleic acid probes have become very popular and extensively exploited in research. They are being used for diagnosis in medicine, forensic science, plant breeding and many fields including microbial genomics (Caskey 1987; Landegren et al. 1988; Gill et al. 1985).

These single-stranded nucleic acid molecules can associate with other nucleic acid molecules in the complementary regions to form double-stranded or triple-helical (Venkateswaran and Venkateswaran 1991; Singer and Tang 2015) structures. Appropriate labelling of the probe helps in detection of the hybrid (probe-target pair) (Aquino De Muro. . .). Southern successfully explained the use of probes for detection of specific sequences in 1975 (Southern 1975).

The qualities of an ideal probe are as follows:

- Should be easily hybridized to its target DNA/RNA
- Detectable at very low concentration
- Should be stable at elevated temperatures
- Should give a readable signal when the labelled probe is hybridized with its target
- Resistant to nucleases

There are two types of probes – gene probes and oligonucleotide probes.

11.2.1 Gene Probes

They are generally longer than 500 bases and contain sequences of most of the target gene. PCR is generally used for making gene probes (Southern 1975). As compared to oligonucleotide probes, gene probes are more specific and more detectable because of greater length.

11.2.2 Oligonucleotide Probes

They usually contain 18–30 bases but can be synthesized up to 100 bases and contain only a targeted portion of a gene. Thus, the probe will anneal to its target sequence under optimal hybridization conditions.

11.2.3 Labelling and Detection of Probes

Radioactive labels. Radioactive isotopes (e.g. ^{32}P, ^{35}S, ^{125}I, ^{3}H) are used for labelling of probes, and detection is done by autoradiography or Geiger-Muller counters. This is the most common type of labelling which is sensitive, but due to price and safety considerations, nowadays non-radioactive labelling is preferred.

Non-radioactive labels. These include biotin labels, chemiluminescent enzyme labels (acridinium ester, alkaline phosphatase, β-D-galactosidase, horseradish peroxidase, isoluminol, xanthine oxidase), fluorochromes, antibodies and digoxigenin system. As compared to radioactive labels, they are safer, highly stable, efficient and less time taking for signal detection. Detection method varies with the nature of the label used, using fluorescence, colour, enzymatic or chemical reactions.

11.3 Artificial Nucleic Acid Probes

For the improvement of quality and hybridization efficiency of probes, efforts have been made by altering structures of nucleic acids. A number of nucleic acid analogues have been synthesized either by incorporating artificial nucleobases (Benner et al. 1998; Geyer et al. 2003) or by replacing their ribose phosphate backbone by other sugars or linkage isomers or by short linear motifs of glycerol or glycine derivatives. GNA (glycerol-derived nucleic acid) was one of the first molecules to be synthesized (Schneider and Benner 1990). Although a number of polymeric analogues have been synthesized so far, the LNA (locked nucleic acid) and the PNA (peptide nucleic acid) are the most widely used and applied for research and diagnostics.

Wengel and Imainishi created artificial nucleotides which have inflexibility within their sugar backbone (Koshkin and Wengel 1998; Obika et al. 1997), which resulted in energetically favourable conformation which later were coined as LNAs. Similarly, a group led by Nielsen suggested designing of PNAs which were synthesized resembling polypeptide in the backbone region (Egholm et al. 1992; Nielsen et al. 1993).

11.3.1 Chemistry Behind Probes

Although the sole intention was same for both LNAs and PNAs, the basic design is different for both these probes. In LNAs, few monomeric NAs are substituted in place of natural ones. Sugar moiety is modified by putting (–O–CH2 –) between C2′ and C4′ positions (Fig. 11.1). Changing the original deoxyribose sugar forces the sugar ring into a rigid conformation, thus called ‘locked NA’ (Briones and Moreno 2012). This structure confers increased stability of probe-target hybrid, and thus energetically favourable duplex is formed.

In PNAs, whole negatively charged sugar-phosphate backbone in NAs is replaced by neutral or uncharged backbone, thus avoiding electrostatic repulsions

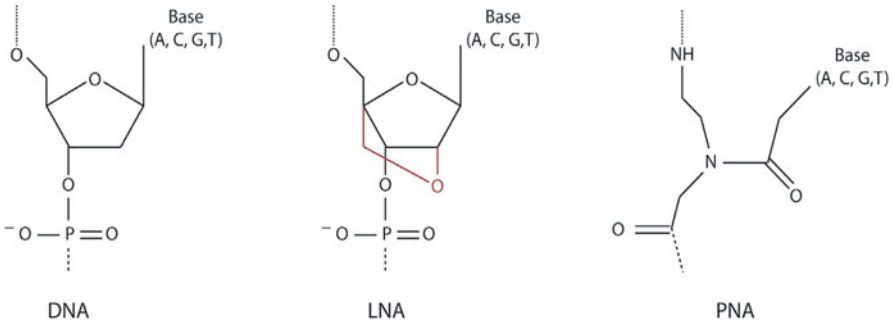

Fig. 11.1 Comparison of the structural modifications in LNA and PNA with respect to DNA. (Briones and Moreno 2012)

due to charged sugar-phosphate backbone in all natural NAs. Thus, binding affinity is highly increased and can even form triple helix, which LNA is incapable of doing. As compared to LNAs, PNAs are more advantageous as it can be used with both ss NAs and ds NAs.

11.3.2 Hybridization Conditions

Several parameters have to be considered for achieving maximal hybridization efficiency like length and composition of nucleotide sequence, melting temperature (T_m) ionic strength, optimal hybridization temperature and target format (solid support, *in solution* or in situ). Usually rate of hybridization and sensitivity increases with concentration of probe.

11.3.3 Advantages of Artificial Nucleic Acid Probes

As compared to natural NAs probes, the synthetic nucleic acid probes have several advantages:

- Increased binding affinity to their targets
- Highly specific and sensitive
- Resistant to enzymatic degradation
- Higher T_m and more stable
- Resistant to enzymatic degradation
- Faster reaction kinetics

11.3.4 Limitations of Artificial Nucleic Acid Probes

While adding LNAs to oligonucleotide probes, thermal stability, measured in T_m is fairly increased and various studies has confirmed that T_m increases by 3–10 °C for each LNA monomer replacement in LNA-modified hybrid oligomer (Campbell and Wengel 2011; Nielsen et al. 2000). Increased Tm leads to greater binding affinity which leads to increase in the levels of non-specific binding. Thus, the number of LNA monomers added to the oligonucleotide is restricted. There are more chances of dimer formation due to enhanced binding affinity.

Major limitation with PNAs is their inherent neutral nature which makes them hydrophobic and difficult to dissolve in aqueous media. This has put a length constrain of maximum up to 12 bases with low GC content (<60%) (Nielsen 1999).

11.4 Synthetic Applications of Artificial Nucleic Acid as Probes

The synthetic probes are mostly exploited in common clinical diagnostic assays, and they are replacing natural nucleic acid probes because of their improved performance.

Detection of pathogenic microorganisms and their genome and gene expression studies in biological samples using:

a) *Microarrays*

It is used for detection of specific sequences using single-stranded probes or oligomers. These probes are immobilized on a solid surface, and labelled target is added to the surface. The probe captures the target and gets fixed on the surface. Hybridization of the target to the probes results in marked increase in optical signature, and detection is done by subsequent imaging. Earlier studies have shown that LNA-modified oligomers perform better than natural NA oligomer in in situ *hybridization* (Castoldi et al. 2006), so they can be used as probe in microarrays with proper optimization (Fang et al. 2006).

PNAs being neutral work better as probes in microarray technique because of lack of any electrostatic repulsion with the target as well as greater binding affinity, and they are also very specific and will not bind in case of mismatches unlike their natural counterparts (Choi et al. 2010). The detection and genotyping of HPV (Human papilloma virus) using PNA-based microarray chip reduced Type I and Type II errors (Choi et al. 2009; Song et al. 2010).

b) *In situ hybridization*

In diagnostic assay, ISH is used to detect particular gene sequence within a cell. Fluorescent-based probes are used for locating the target sequence and detected using fluorescence microscopy. The advantage of using artificial NAs probe here is as they are hydrophobic (PNAs) as well as resistant to nuclease and protease attack (Demidov et al. 1993), so cell membrane cannot hinder their entry inside the cell and would not be degraded easily. LNA-modified FISH oligomers were used to rRNA-targeted FISH assays demonstrated by Kubota and his team in 2006, and till now majority of assays utilizes rDNA because of their relative abundance.

However, PNAs have been more exploited in FISH assays; their length ranges from 13 to 18mer, while DNA oligomers have length in the range of 20–25mer; however PNAs with length ranging from 13 to 18mer have been more exploited in FISH assays as smaller length of PNAs gives them higher specificity even at single mismatch level (Cerqueira et al. 2008). FISH-based assays utilizing PNAs have been used for detection of number of bacteria and fungi including *E. coli*, *Pseudomonas* sp., *Salmonella* sp. and *Listeria* sp., along with other species (Forrest et al. 2006; González et al. 2004).

c) *PCR*

Artificial NAs probes are also utilized in PCR-based assays. These probes cannot be used as primers because elongation process by polymerase is halted

due to structural constraints in artificial probes. Thus, in PCR-based assays, they compete with primers and help in detection of altered gene sequence.

PNAs are used commonly for primer clamping and elongation arrest strategies (Orum 2000; Orum et al. 1993).

In primer clamping, both DNA-based primer and synthetic NA probe are used; being highly specific, artificial probe will bind *only* to its off-target template or clamp (e.g. wild-type allele of a gene); no amplification will be seen, but primer will bind if the gene is mutated resulting in amplification (Fig. 11.2a). In this way, artificial probe can be used for detection of even single nucleotide polymorphism (SNP).

Elongation arrest is a less frequently used approach in PCR-based assays. During the elongation process in PCR based assays, when identical primers are

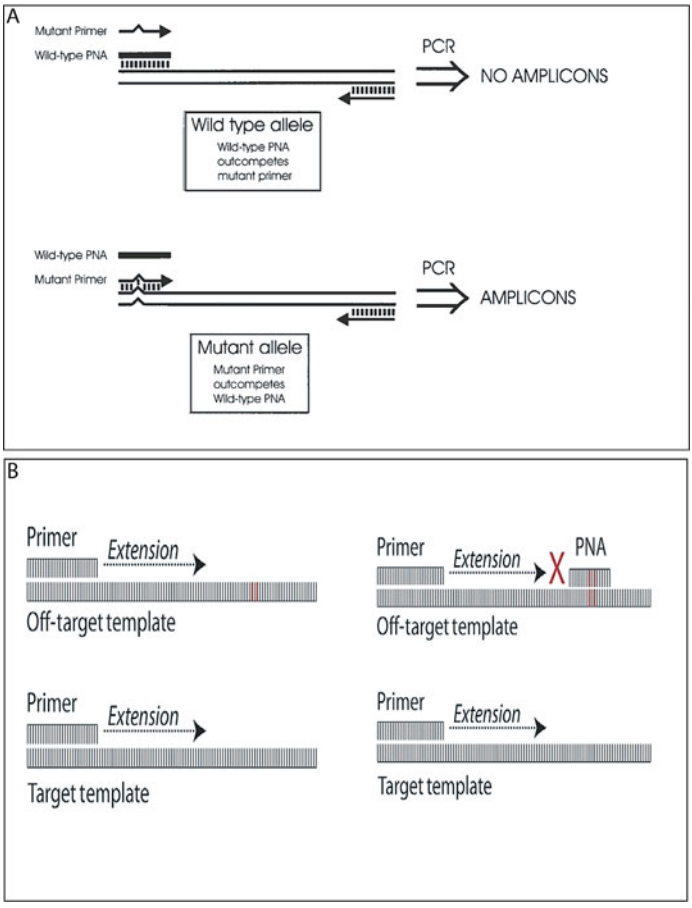

Fig. 11.2 Functional uses of PNA in polymerase chain reaction (PCR)-based assays, (**a**) PCR primer clamping process (Orum 2000), (**b**) elongation arrest using PNA. (Briones and Moreno 2012)

used for both target and off-target sequence and artificial probe is used as blocking probe for off-target sequence. Although elongation process will continue, the process will not be completed due to the presence of blocking probe (Fig. 11.2b).

Apart from the above-mentioned applications in diagnostic assays, these probes play a role in:

- Detection of miRNAs
- Therapeutics
- DNA genotyping
- Mutation detection (single-base mismatches)
- Tandem-repeats detection (forensics)

11.5 Future Prospects

Researchers are continuously trying to engineer the nucleic acids in the manner they can overdo and outperform the natural NA oligomers so that further improvement in diagnostics can be done. Burgeons of new nucleic acid analogues (e.g. bridged nucleic acids) are being made with different physicochemical and thermodynamic properties for the improvement of different assays and to be used in therapeutics for various disorders.

11.6 Tools for Probe Designing and Quality Assessment

Due to rapid development and advancement of technologies in the field of genomics and biomedical research like next-generation sequencing, microarrays and nuclear magnetic resonance, there is constant upgradation in the statistical methods of data analysis and also in the bioinformatics tools and software developed so far. More than 4000 tools are available for bioinformaticians and scientists. More and more omics data are being generated; thus there is immense need to organize and categorize all the data, databases and information about the available tools.

OMICtools (https://omictools.com) is one broad-spectrum approach, which is open access didactic directory holding information about >4400 databases and software tools (Henry et al. 2014). This meta-database contains all the classified tools that have been classified and with elaborated details. Also, the published evaluations of tool performance are incorporated for user's guidance in tool selection. Furthermore, an interface is there for anyone who is accessing the 'OMICtools' for any query related to tools and databases.

This portion gives an overview of few omics tools for probe designing and quality evaluation.

11.6.1 Picky

11.6.1.1 Introduction

It designs oligo probes that are very specific and unique to sequence regions. Picky uses rigorous whole genome-based thermodynamic screening to identify potential hydrogen binding sites of each probe (Chou et al. 2004). 32-bit and 64-bit Picky for Macintosh, Unix and PC versions are available. Along with Picky1.1, updated versions are also Picky 2.0, 2.1 and 2.2. Picky 2.0 can design shared probes for a set of genes that cannot be individually identified using unique probes. Picky 2.1 has the capability to reanalyse existing microarray probes (still valid) against updated gene sets to determine probes (Chou 2010).

11.6.1.2 Input File

Load sequence text file by simply left clicking of *Open target files*.

11.6.1.3 Oligo Design Parameters

Oligo design parameter can be specified and computed by left clicking the *Compute Oligo* (Table 11.1 Ref:http://complexcomputation.org/download/Picky/tutorials).

11.6.1.4 Analysis of Loaded Data

Blue text oligos can be used as oligo probes. Red underlined regions represent matched nontarget region (so not used as probes). The darker shades of red, orange and yellow colors, represents a high degree of similarity among other sequences, while the yellow and white show a low degree of similarity between others Fig. 11.3a.

By double clicking on any row, a new screen will be displayed for comparison of the specified sequence with others Fig. 11.3b.

According to specific matching, selected sequence appears in the first row with subsequent sequences listed below.

11.6.1.5 Limitations

1. Regarding input file – It is best to input complete base sequences, although we can assign random unknown sequences to Picky.
2. Regarding memory – Picky needs ~20 times memory of the input data for computation (Chou 2010).

Table 11.1 Oligonucleotide design parameters with description

Parameter	Description
Oligo size maximum default: 70	Picky will provide probes with the base size that lies between the minimum and maximum values
Minimum default: 50	
Ranges: 50–90	
Match length	Picky will underline any match that is identical to another sequence by this parameter's given number. These underline sequences will not be contained in any 'coloured' text probe
Default: 15	
Range: 10–25	
Sensitivity level	Picky will colour code base sequences that have high similarities with other continuous sequences. Picky is designed to show pairwise alignment that may give a staggered line of similar results. A lower parameter means a significantly longer computational period
Default: 10	
Minimum: 5	
GC content	Picky will only select probes within the selected parameter of G and C base percentages
Default max: 70	
Default min: 30	
Must have a 20-point difference	
Number of probe candidates	Picky will compare the given number based on match region to all nontarget genes for the selection of probes
Default: 100	
Range: 50–500	
Number of probes per gene	The number of distinct, nonoverlapping oligo probes to be chosen for each sequence
Default: 1	
Range: 1–5	
Minimum temp	The difference between a probe candidate's target melting temperature and its highest nontarget melting temperature that is considered to be safe
Difference default: 20.0	
Range: 5.0–30.0	
DNA concentration (nanoM)	Picky uses this value to calculate melting temperatures between target and nontarget genes and this needs to be given as an approximate value
Default: 1.0 nM	
Range: 0.001–1000.0 nM	
Salt concentration (milliM)	Picky will implement this value to calculate hybridization based on melting temperatures
Default: 75 mM	
Range: 0.1–3000 mM	

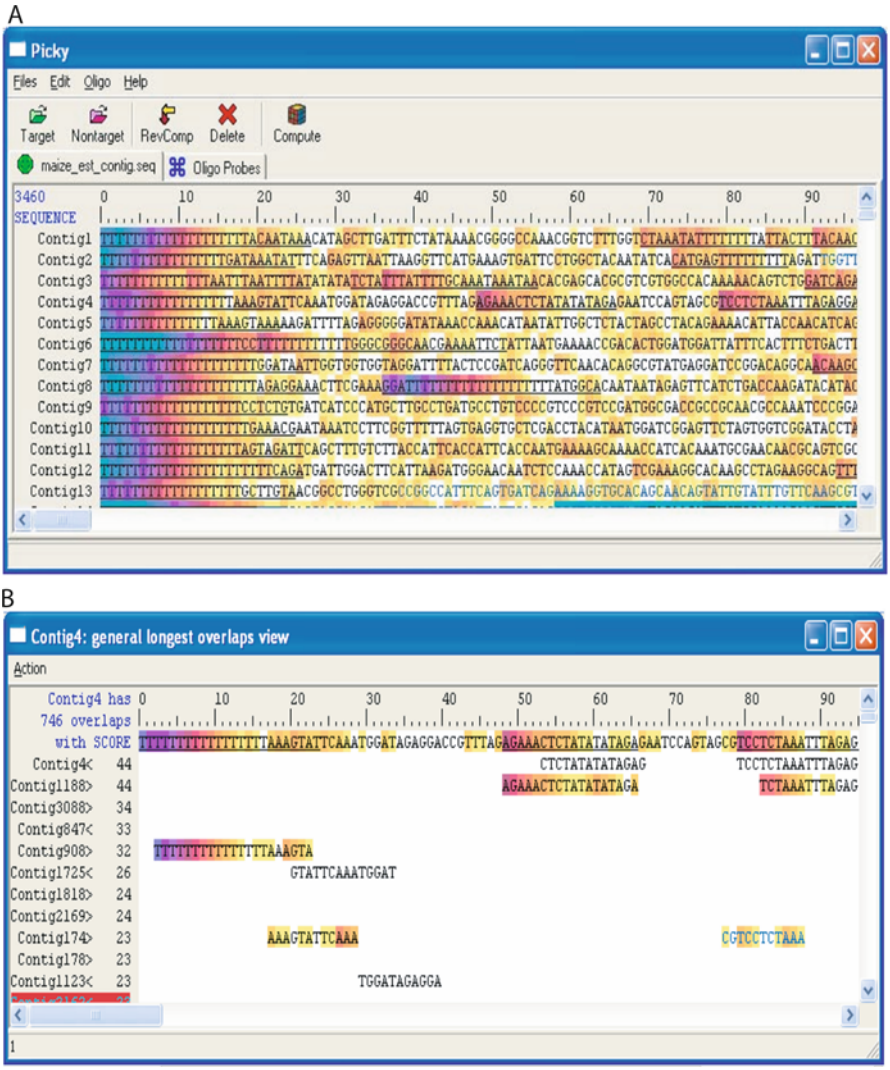

Fig. 11.3 Analysis of loaded data for probe design and comparison of specific sequences by Picky. (**a**) Schematic representation of loaded data. (**b**) Schematic representation of comparison of specific sequences

11.6.2 OligoWiz

11.6.2.1 Introduction

OligoWiz is microarray probe designing tool that allows for sequence annotation. OligoWiz 1.0 is built for selecting one single long probe (50–70 bp) per gene, while

Fig. 11.4 Schematic representation of OligoWiz 2.0

advance version, i.e. OligoWiz 2, can automatically place multiple probes according to advanced rule-based selection (Wernersson and Nielsen 2005).

It is possible to download and run both the server and client-side part of the OligoWiz suite on your system. The server needs Unix system; client runs on any system with Java (1.4 or newer). An overview of OligoWiz 2.0 is represented in Fig. 11.4.

11.6.2.2 Methods:

(a) *OligoWiz client launching*
After downloading the latest OligoWiz and Java, launch OligoWiz client by double-clicking on JAR file (Windows/Mac) or from command line (Linux).

(b) *Input file*
Prepared target sequences in FASTA format or TAB files can be used as input file. For this click the '. . .' button just next to input file.

(c) *Select species database*
Species database that will be used for calculating cross-hybridization and low complexity scores is selected.

(d) *Score parameter customization*
Score parameters can be customized by selecting the predefined parameters. Probe suitability scores are cross-hybridization, ΔT_m, folding, position and low complexity. Scoring values are in between 0.0 and 1.0. Best position for probe placement is based on total score which is also normalized in between 0 and 1.

(e) *Submit query*
Submit your query by hitting 'Submit' button. Optionally email address can be provided to get download link for result data file (.own.gz).

(f) *Place probes*
Loading the data file, launch the main interface for placing probes. For placing probes:

- Adjust score weight if needed. Disable a score by setting weight to 0.0.
- Launch the probe selection tool.
- Select criteria for probe placement. For short probes (~25 bp), user can select probes with ≥8 bp option, and for long probes (ranging from 50–70 bp), ≥2–4 bp probes are preffered. Apply criteria by pressing 'Apply to all' button.
- Inspect the placement of probes in main window. By clicking header elements notice the *Entries* and *oligos*. It makes easy to identify target sequences for which few or no probes have been selected.

(g) *Export probe sequences*
Open the Probe Export Window by pressing 'Export oligos…'. Sequences can be exported in FASTA and TAB format (Wernersson 2009).

11.6.2.3 Limitations

(a) Regarding input file: File must be a text-only file in FASTA or TAB format. File with a single large DNA sequence of an entire prokaryotic genome will not work, as OligoWiz works in a gene-oriented way.
(b) Regarding memory: More memory needed if a FASTA file with >10,000 sequences.
(c) Network problems: As OligoWiz is server based, client fails to connect if no Internet connection is there. Client communicates with server using HTTP (direct connection, without HTTP proxy).

11.6.3 Teolenn

This is a Java-based customizable workflow to design universal probe. Teolenn supplies quality scores for each designed probe for comparisons between quality scores and signal intensities (Jourdren et al. 2010). Latest version of Teolenn is v2.0.1.

11.6.3.1 System requirements

Teolenn works on Linux/Unix system with Java (≥5.0) and SOAP1.x. To use the unicity measurement module, GenomeTools is required. Multicore or multiprocessor system is recommended. Teolenn user interface is the command line. Available options and syntax can be gotten using the command: $ teolenn.sh -h

The genome, you want to use for design, must be in one file (FASTA format) with simple header. Similarly, the sequence of masked genome, used for complexity measurement, also is in one file using FASTA format with simple header. Header of both files must be the same.

11.6.3.2 The design file

This is the core element of design using Teolenn. The design file is an XML file which allows this tool to be a very flexible design tool.

11.6.3.3 Probe selection

Probe design strategy of Teolenn is represented in Fig. 11.5. Steps of probe selection are as follows:

(a) Generate all oligonucleotide sequence for both genome and masked genome, with extensions '. oligo' and '. masked', respectively. Then filter oligo sequences with '.oligo.filtered' and '.masked.filtered' extensions.

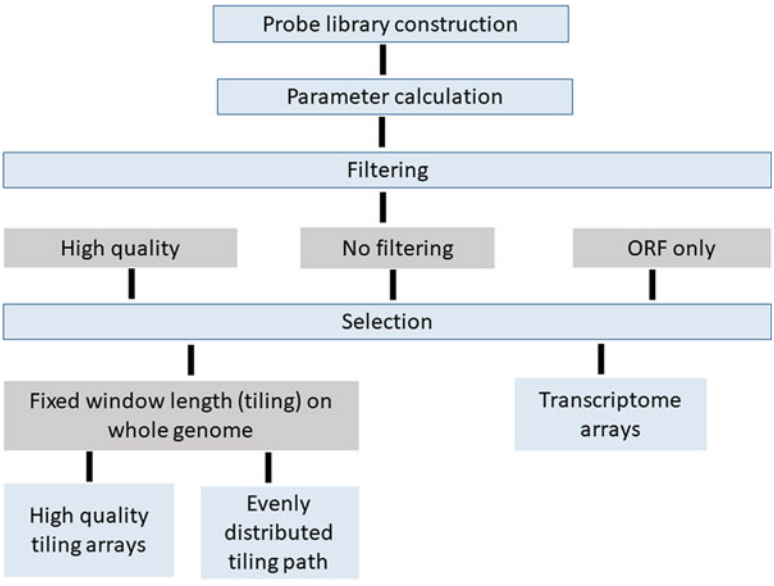

Fig. 11.5 Probe design workflow

(b) Compute and filter measurements (create oligo.mes file, filtered.mes and filtered.mes files).
(c) Compute the selection of oligonucleotide (create the select.mes file) by applying weight on each measurement score to get global score, and then choose best score oligo.

11.6.3.4 Limitations

Probes cannot be designed with genome sequence files having >500 scaffolds due to specificity calculation limitations of genome tool.

11.6.4 MathFISH

This is a web-based tool that calculates all modelled thermodynamic properties of a single probe for any target molecule provided. The output includes ΔG values, hybridization efficiency and formamide dissociation profile.

The structure of mathFISH organizes mathematical modelling applications using eight tools (Yilmaz et al. 2011):

11.6.4.1 General analysis tool

This performs all available simulations for a given probe sequence pair. The target sequence can be entered in different formats (plain text, FASTA, GenBank) Fig. 11.6a

11.6.4.2 Mismatch analysis tool

This tool accepts a probe, a perfect matching target and a nontarget sequence as input Fig. 11.6b.

11.6.4.3 Competitor analysis tool

In addition, with the above, it requests the sequence of the competitor oligonucleotide Fig. 11.6c.

Fig. 11.6 Different analytical tools of mathFISH for probe evaluation. (**a**) Schematic representation of general analysis tool. (**b**) Schematic representation of mismatch analysis tool. (**c**) Schematic representation of competitor analysis tool. (**d**) Schematic representation of $\Delta G^0_{\text{overall}}$ tool

11.6.4.4 Other tools

Other tools perform task with multiple probe/target pairs in single event. There are four tools to generate free energy series for the four different free energy values (ΔG^0_1, ΔG^0_2, ΔG^0_3, $\Delta G^0_{\text{overall}}$) by walking a perfect match oligo of constant length over a target sequence Fig. 11.6d.

11.6.4.5 Limitations

Main limitation is the lack of high-throughput analysis tools for large sets of probe pairs, for automation of probe design process. This is speed and memory problem basically. Ambiguities in probe sequences are also an issue.

11.6.5 ProbeMaker

This is Java software framework that enables constraint-based design and analysis of many different types of oligonucleotide probes. This tool assists with probe design, incorporating sequence motifs for amplification and visualization (Stenberg et al. 2005).

11.6.5.1 Workflow Setup (Example for Windows OS)

- Install the Java (least v1.4). Unpack the download file and launch AppTools workbench by double-clicking the 'start.bat' file.
- Activate the ProbeMaker plugins and choose to load GUI plugin.

a) *Target input*

- Create a new project, import target sequences and expand the targets (as we want to make two probes for each target Fig. 11.7a.

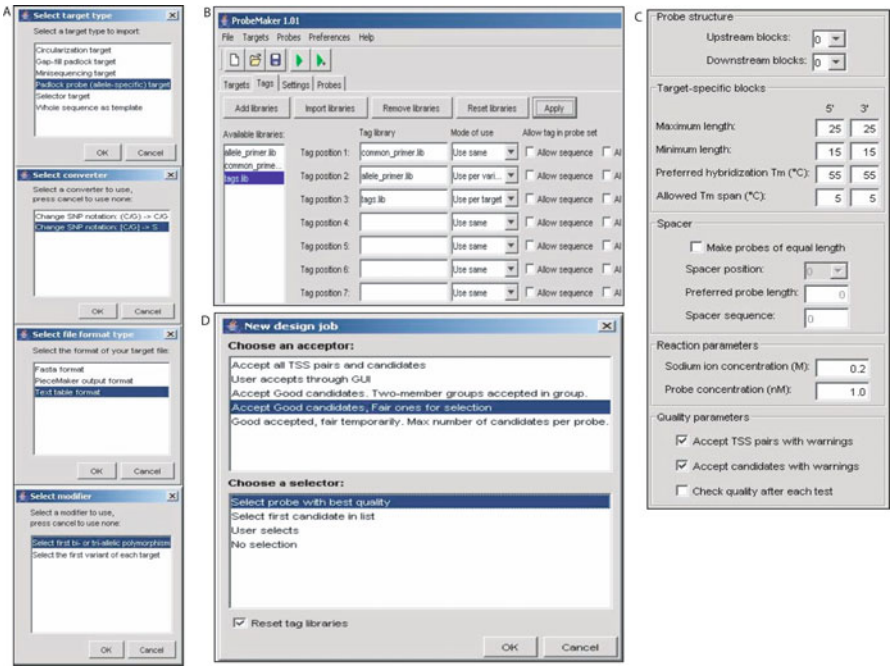

Fig. 11.7 Various steps of probe design in ProbeMaker. (**a**) Schematic representation for importing target sequences. (**b**) Schematic representation for setting tag libraries. (**c**) Schematic representation of setting design parameters. (**d**) Schematic representation of running a design job

b) *Setting tag libraries*

- For padlock probes, add two primers (one is common for all probes, and other is either of two allele-specific sequence) sequences and an address tag for each probe Fig. 11.7b.
- *Import tags* by clicking 'import libraries', select file and then select FASTA format.
- Set the 'Mode of use' and press 'Apply' to apply changes to project.

c) *Setting design parameter*

- Check and fix the parameters in settings tab and go on.
- Drag and drop module to TSS (target-specific sequence) testing box. Double-click the module name in testing box; the parameter will appear. If you want any change, change settings, and apply those by pressing 'Apply' Fig. 11.7c.

d) *Running a design job*
First save your project and then press the 'Design probes to replace' icon in the toolbar. Here select an acceptor and a selector and then click 'OK' Fig. 11.7d.

- After design job completion, newly designed probe is displayed in different colours: green, highest quality, and yellow, average quality level.

e) *Looking results*

- To view list of all probes, T_m, quality value and their 5′ and 3′, press 'Probe summary'.
- 'Message summary' displays list of problem type and their frequency.
- 'View tags by probe' displays tag allocation to which probe.
- Pressing 'Export probes' export probes to file in varied formats.

11.6.5.2 Limitations

Probe design is limited by the amount of time required and available memory. The time required to complete design job is influenced by many factors and is difficult to predict.

References

Aquino De Muro M. Probe design, production, and applications
Benner SA et al (1998) Redesigning nucleic acids. Pure Appl Chem 70:263–266
Briones C, Moreno M (2012) Applications of peptide nucleic acids (PNAs) and locked nucleic acids (LNAs) in biosensor development. Anal Bioanal Chem 402:3071–3089
Campbell MA, Wengel J (2011) Locked vs. unlocked nucleic acids (LNA vs. UNA): contrasting structures work towards common therapeutic goals. Chem Soc Rev 40:5680
Caskey CT (1987) Disease diagnosis by recombinant DNA methods. Science 236:1223–1229

Castoldi M et al (2006) A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). RNA 12:913–920
Cerqueira L et al (2008) DNA mimics for the rapid identification of microorganisms by fluorescence in situ hybridization (FISH). Int J Mol Sci 9:1944–1960
Choi J, Kim C, Park H (2009) Peptide nucleic acid-based array for detecting and genotyping human papillomaviruses. J Clin Microbiol 47:1785–1790
Choi J-J, Jang M, Kim J, Park H (2010) Highly sensitive PNA array platform technology for single nucleotide mismatch discrimination. J Microbiol Biotechnol 20:287–293
Chou H-H (2010) Shared probe design and existing microarray reanalysis using Picky. BMC Bioinforma 11:196
Chou H-H, Hsia A-P, Mooney DL, Schnable PS (2004) Picky: oligo microarray design for large genomes. Bioinformatics 20:2893–2902
Demidov V, Frank-Kamenetskii MD, Egholm M, Buchardt O, Nielsen PE (1993) Sequence selective double strand DNA cleavage by peptide nucleic acid (PNA) targeting using nuclease S1. Nucleic Acids Res 21:2103–2107
Egholm M, Buchardt O, Nielsen PE, Berg RH (1992) Peptide nucleic acids (PNA). Oligonucleotide analogs with an achiral peptide backbone. J Am Chem Soc 114:1895–1897
Electrochemical nucleic acid – based biosensors: Concepts, terms, and methodology (IUPAC Technical Report) * – Semantic Scholar. Available at: https://www.semanticscholar.org/paper/Electrochemical-nucleic-acid%2D%2D-based-biosensors-%3A-Labuda-Brett/3861ec38a788e3b20cf8a3c4c1de73284fcf74df. Accessed 23 Feb 2018
Fang S, Lee HJ, Wark AW, Corn RM (2006) Attomole microarray detection of microRNAs by nanoparticle-amplified SPR imaging measurements of surface polyadenylation reactions. J Am Chem Soc 128:14044–14046
Forrest GN et al (2006) Impact of rapid in situ hybridization testing on coagulase-negative staphylococci positive blood cultures. J Antimicrob Chemother 58:154–158
Geyer CR, Battersby TR, Benner SA (2003) Nucleobase pairing in expanded Watson-Crick-like genetic information systems. Structure 11:1485–1498
Gill P, Jeffreys AJ, Werrett DJ (1985) Forensic application of DNA 'fingerprints'. Nature 318:577–579
González V et al (2004) Rapid diagnosis of *Staphylococcus aureus* bacteremia using *S. aureus* PNA FISH. Eur J Clin Microbiol Infect Dis 23:396–398
Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ, Desfeux A (2014) OMICtools: an informative directory for multi-omic data analysis. Database (Oxford) 2014
Jourdren L et al (2010) Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments. Nucleic Acids Res 38:e117–e117
Koshkin AA, Wengel J (1998) Synthesis of novel 2′,3′-linked bicyclic thymine ribonucleosides. J Organomet Chem 63:2778–2781
Landegren U, Kaiser R, Caskey CT, Hood L (1988) DNA diagnostics – molecular techniques and automation. Science 242:229–237
Lederberg J (2000) Infectious history. Science 288:287–293
Nielsen PE (1999) Applications of peptide nucleic acids. Curr Opin Biotechnol 10:71–75
Nielsen PE, Egholm M, Berg RH, Buchardt O (1993) Sequence specific inhibition of DNA restriction enzyme cleavage by PNA. Nucleic Acids Res 21:197–200
Nielsen KE, Singh SK, Wengel J, Jacobsen JP (2000) Solution structure of an LNA hybridized to DNA: NMR study of the d(CTLGCTLTLCTLGC):d(GCAGAAGCAG) duplex containing four locked nucleotides. https://doi.org/10.1021/BC990121S
Obika S et al (1997) Synthesis of 2′-O,4′-C-methyleneuridine and -cytidine. Novel bicyclic nucleosides having a fixed C3, -endo sugar puckering. Tetrahedron Lett 38:8735–8738
Orum H (2000) PCR clamping. Curr Issues Mol Biol 2:27–30
Orum H et al (1993) Single base pair mutation analysis by PNA directed PCR clamping. Nucleic Acids Res 21:5332–5336
Relman DA (2011) Microbial genomics and infectious diseases. N Engl J Med 365:347–357

Schneider KC, Benner SA (1990) Oligonucleotides containing flexible nucleoside analogs. J Am Chem Soc 112:453–455
Singer A, Tang YW (2015) Artificial nucleic acid probes and their applications in clinical microbiology. Methods Microbiol 42:2–613 .(Elsevier Ltd.
Song HJ et al (2010) Comparison of the performance of the PANArray™ HPV test and DNA chip test for genotyping of human papillomavirus in cervical swabs. Biochip J 4:167–172
Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517
Stenberg J, Nilsson M, Landegren U (2005) ProbeMaker: an extensible framework for design of sets of oligonucleotide probes. BMC Bioinforma 6:229
Venkateswaran N, Venkateswaran KS (1991) Nucleic acid probes in microbiology. Def Sci J 41:335–356
Wernersson R (2009) Probe design for expression arrays using OligoWiz. Methods Mol Biol (Clifton, NJ) 529:23–36
Wernersson R, Nielsen HB (2005) OligoWiz 2.0 – integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res 33:W611–W615
Yilmaz LS, Parnerkar S, Noguera DR (2011) mathFISH, a web tool that uses thermodynamics-based mathematical models for in silico evaluation of oligonucleotide probes for fluorescence in situ hybridization. Appl Environ Microbiol 77:1118–1122

Important Links (for Tools)

http://complexcomputation.org/download/Picky/tutorials
http://mathfish.cee.wisc.edu
http://probemaker.sourceforge.net
http://www.cbs.dtu.dk/services/OligoWiz
http://www.outils.genomique.biologie.ens.fr/teolenn

Chapter 12
Omics-Based Nanomedicine

Chirasmita Nayak, Ishwar Chandra, Poonam Singh, and Sanjeev Kumar Singh

Abstract Traditionally "one-size-fits-all" paradigm for all patients within the disease has the limit of successful treatment. Personalized medicine aims to individualize therapeutic interventions, in combination of ex vivo omics data (i.e., genomics, proteomics, metabolomics, etc.) profiling and in vivo imaging insights on the type, the stage, and the grade of the disease called as theranostic. Personalized medicine has led to the discovery of various biomarkers that can detect the early stage of disease as well as response of bioactive molecules. Nanomedicine, the application of nanotechnology in medicine, holds great promise for diagnosing, treating, and preventing disease and traumatic injury and of relieving pain using molecular tools and molecular knowledge of the human body. Personalized nanomedicine has the power of integration of nanomedicine and molecular biomarkers to improve diagnosis, disease management, and prognosis as well as individualized drug selection by reducing side effects and cytotoxicity. In this book chapter, we have discussed about the leading technologies available for personalized nanomedicine and immense potential combination of nanomedicine with high-throughput *omics* technology.

Keywords Omics · Nanomedicine · NGS · Personalized medicine · Biomarkers

12.1 Introduction

Nanotechnology is the science and technology involving molecular imaging, measuring, modeling, and manipulating matter at dimensions of approximately 1–100 nm with unique characteristic features at the cellular, atomic, and molecular

C. Nayak · I. Chandra · S. K. Singh (✉)
Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India

P. Singh
Corrosion & Materials Protection Division, C.S.I.R – Central Electrochemical Research Institute (CECRI), Karaikudi, Tamil Nadu, India

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_12

levels (Silva 2004). It comprises various scientific fields such as chemistry, biology, physics, mathematics, engineering, and information technology which can vastly affect trade, society, human health, environment, viable growth, and security. The word "nano" is derived from the Greek word "dwarf." In the year 1959, Richard Feynman described the concept of nanotechnology in a lecture titled "There's plenty of room at the bottom" and won the Nobel Prize (Hunyadi 2017; Bhardwaj et al. 2014). The application of nanotechnology extended toward medicine, biotechnology, diagnostics, drug delivery, tissue engineering, toxicology testing, chemistry, environment, catalyst, filtration, energy, displays, foods, household, optics, textiles, cosmetics, and agriculture (Rakesh et al. 2015).

The application of nanotechnology in medicine is called as nanomedicine, which holds great potential to diagnose, treat, and prevent disease using molecular tools and molecular knowledge of the human body (Boisseau and Loubaton 2011). Nanomedicine is currently being involved in drug delivery, therapy techniques, diagnostic techniques, antimicrobial techniques, and cell repair (AkilaKesavan 2014). There are plenty examples of nanomedicines used to cure cancer and other diseases, and also recently this is applied on other diseases like Alzheimer's disease, lung disease, and AIDS treatment. Nanomedicine particularly refers to the nanoparticles, small in size with novel components which can significantly revamp the physical, chemical, and biological properties, phenomena, and processes. These nanoparticles come under three major groups such as natural, incidental, and engineered (Giese et al. 2018). The nanoparticles can provide novel tools and techniques having unique composition and functionalities which have not subsisted previously in biomedical research (Wang and Wang 2014). These nanoparticles offer various unprecedented communications with biomolecules on the cell surfaces and within the cells which can revolutionize various biochemical and physiochemical properties of these cells (Cai et al. 2008; Mody et al. 2010). Nanomedicine with the help of genomics and proteomics can revolutionize the traditional medicine to bring signature cure for individual patients.

Personalized medicine (PM) is defined in the field of healthcare as a form of medicine which utilizes each person's genes, proteins, and environmental information to thwart, diagnose, and treat the disease (Agyeman and Ofori-Asenso 2015). This approach has the ability to predict the medical treatment which will be safe and effective with less side effects for each patient and which will not. It supplies an opportunity to generate new instrument targeting a group of patients who failed in traditional medication and who do not react as intended (Vogenberg et al. 2010). PM follows the omics-based approach to make the decision for treating patients because molecular characteristics obtained from this approach can classify diseases and identify subpopulation of patients' suitability toward common treatments more precisely (Boisseau and Loubaton 2011). The application of PM has been deeply studied in the field of oncology (Al-Mozaini and Mansour 2016).

The word "omics" refers to the study of the roles, relationships, and actions of various types of molecules that make up a cell, tissue, or organism such as genes (genomics), epigenetics (epigenomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) (Horgan and Kenny 2011). The omics data

provides the diverse molecular biomarkers which has the ability to capture the whole biological pictures in an unbiased manner (Tebani et al. 2016). Each type of omics data provides a list of differences between species and individuals of a species or disease or control groups. These types of molecular profiles can vary for cell or tissue exposure to chemicals and can be useful as molecular biomarkers of the disease process and give insight as to which biological pathways are different in disease or control groups (Hasin et al. 2017).

This chapter will provide overall idea about nanotechnology, nanomedicine, and personalized medicine. This chapter has discussed the applications of nanotechnology implemented in medicine as nanomedicine and personalized medicine.

12.2 Personalized Medicine

It is common in today's scenario that the drugs are presently based on the disease and condition, i.e., signs and symptoms, rather than the individuals. It is often found during the due course of treatment that some patients with the same disease responds well to the treatment, few shows more side effects compared to others, few metabolizes the drug faster than the other hence needing higher dose than the prescribed dosage, and there could be some who might not show any effect of the treatment, whereas in worst cases, some individuals could not tolerate the drugs at all resulting in adverse effect where medical intervention becomes necessary to save or restore the health condition of patient. Presently therapeutics is favored toward "one-drug-fits-all" strategy, where the same drug is prescribed to patients for common diseases. Personalized medicine is an approach which has the capability to tailor the medical treatment of individual patients in a unique molecular or genetic mapping which will afford best response with highest benefit and lowest side effects or toxicity (Vogenberg et al. 2010). It uses the power of nanomedicines in combination with molecular biomarkers (e.g., genomics, proteomics, epigenomics, and metabolomics) to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment rather than creation of drugs or medical devices that are unique to a patient (Lee et al. 2012). PM focuses on the medicine that can prevent the diseases rather than just reactive using existing biomarkers which ultimately increase the effectiveness and reduce the inherent problems associated with traditional approach. The use of personalized medicine is briefly described as follows:

12.2.1 Small Molecule Inhibitors

The role of personalized medicine therapy with context of small molecule inhibitors is successfully implemented for treating diabetes mellitus where the patients having genotype cytochrome P450 2C9 (CYP2C9) and organic cation transporter 1 (OCT1)

were treated separately by small molecule inhibitors, namely, sulfonylurea and metformin, respectively. Mainly sulfonylureas remain in inactivated form by the action CYP2C9 enzyme in the liver, whereas in the patients having genotype of CYP2C9*2 and CYP2C9*3, they act as inhibitors. These genotypes weaken the function of CYP2C9 around 3.44 times thereby accumulating the drug concentration due to poor sulfonylurea metabolism and have been related with heightened possibility of hypoglycemia (Pearson 2016). OCT1 play important role in the absorption and transportation of metformin across the intestinal wall and involve in the activation of AMP-activated protein kinase (AMPK) to increase the remedial benefits of the medication. It has been found that the hereditary changes in OCT1 might be related with different responses of metformin. Likewise patients with various OCT1 or OCT2 genotypes react contrastingly to oxaliplatin chemotherapy (Shu et al. 2007).

In breast cancer, detection of overexpression of genes is used to develop the personalized medicine for breast cancer patients. Overexpression of human epidermal growth factor receptor 2 (HER2) is considered to be associated with breast cancer progression, and the use of trastuzumab showed promising result in overexpressed HER2 tumors. But recent study showed that the status of overexpressed HER2 could change in various stages of cancer trajectory, i.e., HER2 expression rate can be increased nearly 20% from a primary tumor state to metastasized state. In light of this, prior to planning for trastuzumab therapies, repeat biopsy could be done for assessing HER2 status in order to minimize the potential effect of any discordances in HER2 status on treatment effectiveness. Furthermore molecular imaging technologies for the detection of HER2 expression in tumors would provide a useful means in the administration of trastuzumab therapies and increase the efficacy of the treatment process (Chan et al. 2017).

12.2.2 Immunotherapy

Immunotherapy is an approach which modifies/utilizes components of the host immune system to treat the diseases. It mainly focuses on two areas to design immunotherapies such as suppressing or activating the immune system where activating immunotherapies treat the diseases like cancer and suppressing immunotherapies treat the diseases such as autoimmune diseases (Wraith 2017). Suppressing immunotherapy in diseases refers to suppress selective immune system based on the specific diseases by leaving the rest of the immune system active. This increases the specificity for diseases and decreases the risk of potential side effects which ultimately leads to provide cure of the disease. Immunotherapy grew as a new paradigm to treat cancer which is specifically targeting the antigens arising from the somatic mutations (Bethune and Joglekar 2017). The cancer immunotherapy is arbitrated by T lymphocytes though the deficiency of T lymphocytes displayed maximum incidence of cancer. T cells expressed unique type of receptors which can scan the antigens present on the major histocompatibility complexes on the surface of tumor cells. These tumor antigens are of two types: one is public antigens which can be

found in multiple cancer cells as well as on normal cells and another is neoantigens which are tumor specific and also patient specific (Gubin et al. 2015). These neoantigens were used in personalized T-cell-mediated immunotherapies as biomarker to diagnose, characterize, and monitor the reaction of T cell to checkpoint blockade and as targets by therapeutic vaccines (Gubin et al. 2014). The neoantigens were used in personalized T-cell therapy in two ways: one is personalized adaptive T-cell therapy and another is personalized T-cell receptor gene therapy. In personalized adaptive T-cell therapy, neoantigen-specific T cells were captured with various technologies (Toebes et al. 2006) and combined with multicolor encoding and fluorescence-activated cell sorting to isolate from tumor-infiltrating lymphocytes and peripheral blood (van Rooij et al. 2013). In personalized T-cell receptor gene therapy, T-cell receptors were cloned from neoantigen-reactive T cells to express differentially in a distinct population of effector cells.

Immunotherapy has been also extensively studied in Alzheimer's diseases and has two basic forms to fight, i.e., active and passive immune therapy. In Alzheimer's diseases, the accumulation of amyloid beta and hyperphosphorylated tau protein in the brain by oxidative stress and glial activation results in the synaptic dysfunction and severe neural loss that causes dementia. The active immunotherapy is used to stimulate both cellular and antibody-based immune system by administering fragments of amyloid beta or related antigens, whereas the passive immunotherapy is used to boost resistance against aggregated amyloid beta or remove already aggregated amyloid beta as plaques by injecting performed antibodies (Morgan 2011; Moreth et al. 2013; Winblad et al. 2014).

12.2.3 Chimeric Antigen Receptors

The body's immune system protects it from infection and cancer. T cells/T lymphocytes are integral to the immune system. It is a subtype of white blood cells which develops from stem cells in the bone marrow and protects the body from infection and helps fight cancer. Immunotherapy is a type of treatment which improves the body's ability to detect and kill cancer cells by utilizing the body's own immune system to fight cancer. It is based on the concept that immune cells or antibodies can recognize and kill cancer cells (Accessed from Leukemia and Lymphoma Society 2017). CAR T-cell therapy is a form of personalized cancer immunotherapy where patient's immune system T cells are modified to precisely recognize and kill patient's cancer cells. CAR symbolizes *chimeric antigen receptor* that contains genetically manipulated T cell which is engineered in laboratory which allows them in identification and activation of killing specific cancer cells. Within the body, the CAR T cells multiplies itself and remain for longer periods to check and prevent the recurrence of tumor (Pagel and West 2017).

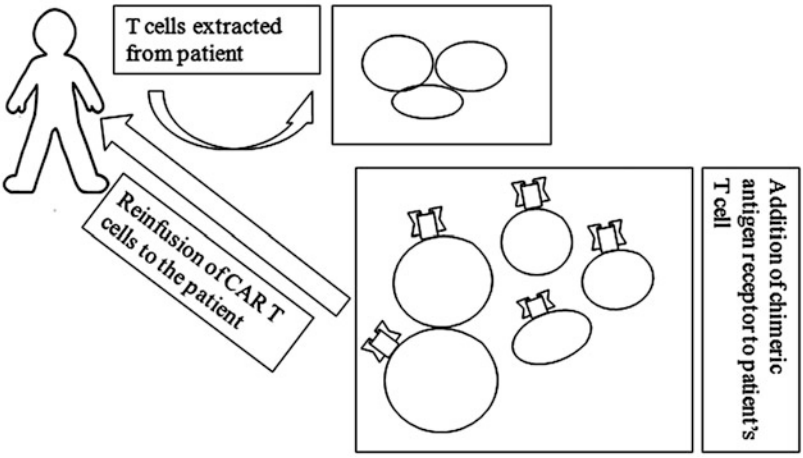

Fig. 12.1 Extraction of T cell from patients, addition of chimeric antigen receptor, and reinsertion of CAR T cells to patient

12.3 An Outline of Construction and Working of CAR T-Cell Therapy

First the T cells are collected from the patient's blood through leukapheresis. The T cells are modified in the laboratory of a drug manufacturing facility where they are genetically engineered by using disarmed virus to produce chimeric antigen receptors (CARs) on their surface. Now the T cells are known as "chimeric antigen receptor (CAR) T cells," and the CAR proteins empower the T cells to recognize an antigen on targeted tumor cells. Copies of genetically modified T cells are grown in the laboratory, frozen and considerable amount of them are returned for reinfusion into the patient where they are being treated. It takes nearly 2 weeks for this process during which the patients are kept on specific chemotherapy that prepares the immune system to support the CAR T cells on its reinfusion into the body (Fig. 12.1). Once in the bloodstream, the CAR T cells multiply and become attacker cells which recognize, and kill, cancerous cells that have targeted antigen on their surface (Accessed from Leukemia and Lymphoma Society 2017; Pagel and West 2017).

12.4 Composition Signaling and Targeting

CAR receptor is composed of three components: (i) the extracellular domain, (ii) transmembrane, and (iii) intracellular domain (Smith et al. 2016). CAR binding to the antigen falls in three general categories: (i) single-chain variable fragment (scFv) derived from antibodies, (ii) Fab (fragment antigen binding) selected from libraries, and (iii) nature ligands that engage their cognate receptor (Sadelain et al.

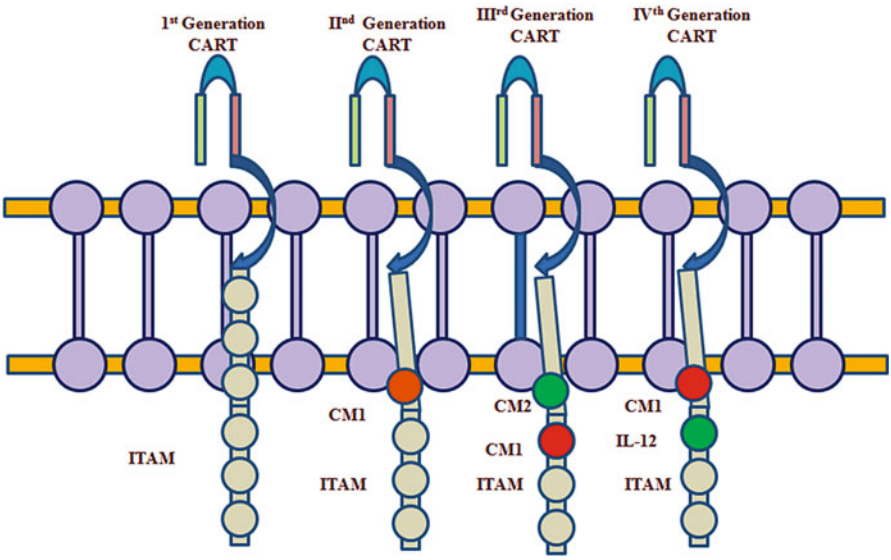

Fig. 12.2 Evaluation of the generation of CART design. (Reproduced from Pearson 2016)

2013). The scFv region and the costimulatory molecules are joined by the transmembrane domain which influences the immunogenicity based on its length (Smith et al. 2016). The intracellular domain contains CD3-zeta chain and is usually joined with costimulatory fragments like CD27, CD28, CD134, and CD137. The costimulatory fragments assist in signaling and influence the generation and the decisiveness of the T cell (Milone et al. 2009). CAR T-cell therapies are divided into four generations and are based on the absence or presence of costimulatory molecules or the ability to stimulate cytokine production (Fig. 12.2) (Abken 2015). The target and the generation of the CAR T cell decide the composition of the CAR involved in CAR T-cell therapy. Surface antigens' epitope unique to the cancer cells provides the target to the CAR T cell. Though CARs normally target highly expressed antigens, little is known including the minimum thresholds and selecting optimal epitopes for CAR targeting (Sadelain et al. 2013).

12.5 Treatment, Limitation, and Future

Different antigens are targeted in various kinds of malignancies, a list of which is given by Sadelain et al. (2013) and Hasin et al. (2017). Most common examples include CD19 antigen which is the prime target in acute lymphoblastic leukemia (ALL), chronic lymphoblastic leukemia (CLL), and B-cell malignancies and CD20 antigen which is targeted in B-cell lymphomas and B-cell malignancies (Hasin et al. 2017). Kymriah, Yescarta, and Actemra are well-known FDA-approved CAR T-cell treatment for ALL, B-cell lymphoma, and giant cell arteritis, respectively (Accessed

from US Food and Drug Administration 2017). Cytokine-release syndrome (CRS), B-cell aplasia, tumor lysis syndrome (TLS), neurologic toxicities, and life-threatening allergic reactions are some of the likely side effects of CAR T-cell therapy, and these complications could be managed by standard supportive therapy (Accessed from Leukemia and Lymphoma Society 2017). The field of immunotherapy is in its inception and has shown remarkable success in patients suffering from blood cancers. Researchers are actively involved in designing CAR T-cell therapy to target novel biomarker. Nevertheless the adverse events associated with autoimmune diseases and toxicities due to its use need to be managed for its effective use. Thus far CAR T-cell therapy has been a promising form of immunotherapy, and its significant improvements may hold a potential for treating patients with cancer.

12.5.1 Omics Era-Personalized Therapy

A precisely sure sign of cancer is the existence of genetic variation in the tumoral DNA. A comprehensive insight about these variations to turn into malignant phenotype is critical for the suitable treatment of oncologic diseases. The identification of altered genes and pathways and their specific oncogenic role immensely supports in personalized anticancer therapy (Vogelstein et al. 2013; Garraway and Lander 2013; Engin et al. 2014). The genomics era has provided splendid improvement within the understanding of cancer biology. Increase in the availability of multiple omics databases related to clinical annotation of tumor cytology, patient response, and outcome generates abundant opportunities for speedy interpretation of high-throughput omics to improve overall longevity. Cancer genomics refers to the study of the entire DNA sequence and gene expression differences between tumor cells and normal host cells by using different profiling strategies like DNA copy number, DNA methylation, and transcriptome and whole-genome sequencing technologies (Vucic et al. 2012). Cancer genomics facilitates in evaluating the omics data to analyze genes and pathways unregulated in cancer and disclose those that may support the detection and management of disease. Such discoveries will certainly enhance the knowledge of cancer biology leading to the finding of novel diagnostic, prognostic, and therapeutic markers that will eventually improve patient outcomes. The Human Genome Project has highlighted the signature genes and proteins specific to diseases, besides annotating the networks of interaction existing between them (Ideker and Sharan 2008; Han 2008). Molecular network analyses are utilized to improve disease classification (Chuang et al. 2007; Segal et al. 2005) and identify novel therapeutic targets (Albert et al. 2000; Araujo et al. 2007). For instance, in chronic myelogenous leukemia, the Philadelphia chromosome, produced by a translocation, induces substantial activation of the tyrosine kinase Abl. The recognition of this molecular modification favored the development of drugs such as imatinib mesylate that inhibit Abl kinase activity to successfully control the disease (Druker et al. 2001). The complete cure of the diseases are to be achieved yet since the occurrence of the mutations in the active ornonactive sites lead to drug resistance in

treatment (DeVita and Canellos 2011). Melanoma patients with mutation in V600E-BRAF gene can be treated with vemurafenib. BRAF mutations have been also found in different cancers like lung and numerous myelomas and hence could be utilized as a part of treating these tumors (Vucic et al. 2012). Essentially HER2 has drawn incredible consideration in cancer research due to its part in tumorigenesis and is exceedingly promulgated in breast, ovarian, and gastric cancer. Trastuzumab, a humanized monoclonal anti-HER2 antibody used to treat breast cancer, aided the investigators to create different HER2-like antibodies, inhibitors for dimerization, and kinase proteins as a cancer therapy. Concurrently, the high articulation of HER2 and the availability to its extracellular domain make HER2 a perfect epitome for the focused antitumor drugs and also imaging agents (Tai et al. 2010). Different hereditary injuries in the form of genetic lesions are required for tumor advancement in most solid tumors (Hanahan and Weinberg 2000). It has been found in the treatment of oncologic diseases that mixes of medications that follow up on various cell targets display predominance over individual medicines (Sawyers 2007). All the above given cases mean the significance of breaking down the qualities and hereditary pathways to discover novel focused treatments in different malignancy.

Next-generation sequencing (NGS) technologies and the development of the latest computational tools have advanced the cancer genomics research. Computational techniques empower the effective investigation of expansive volume of tumor genomics information to better interpret the components of malignancy. The Cancer Genome Atlas (TCGA) project and the International Cancer Genome Consortium (ICGC) are well known for cancer genomics research (Wang and Xie 2016). The cancer genomics research has created a titanic pool of data. The apiece information can be conjoined that will open potential therapeutic pathways and targets. Medicinal chemists can exploit these valuable tools and resources to rationally design novel lead compounds for newly identified targets by precisely targeting the cancer types with distinct genetic features.

12.6 Next-Generation Sequencing

The appearance of sequencing technologies has played an important role in the analysis of genomic sequences of organisms. DNA sequencing developed by Sanger in 1997 is defined as the sequencing of nucleotides within a DNA using laboratory methods. Sanger and Maxam-Gilbert sequencing technologies were the most common DNA sequencing technologies used by biologists which can be run in parallel and 115 kb lengths of sequences read by each reaction. These sequencing technologies were used until the emergence of a new era of sequencing technologies opening new perspectives for genome exploration and analysis. The new era of technologies appeared by Roche in 2005 which were commercialized as technologies capable of producing sequences with very high-throughput called as “next-generation sequencing (NGS) technologies” or “high-throughput sequencing technologies” (Kchouk et al. 2017).

NGS is the process to find the sequences of the entire genome of the organism by means of parallel sequencing and substantially minimizing the fragment of cloning which are used in Sanger sequencing of genome. NGS can be sequenced in parallel millions to billions of reads in a single run from multiple samples at much reduced cost better than Sanger sequencing. The tools of NGS consist of three popular models like 454/Roche, Illumina, and SOLiD/Life Technology (Araújo et al. 2017). In other words, NGS is revolutionizing the study of "omics" (e.g., genomics, transcriptomics, methylomics, etc.) to deviate from microarrays and traditional Sanger sequencing.

12.7 NGS Applications

The applications of NGS appear almost boundless by allowing rapid advances in many fields related to the biological sciences. The human genome sequenced to identify the genes and regulatory elements are involved in the pathological processes. It has provided a wealth of knowledge for comparative biology studies through whole-genome sequencing of a wide variety of organisms in the fields of public health and epidemiology. Additionally, RNA-Seq (NGS of RNA) can replace the use of microarray analysis to study the expression of genes, which provides the researchers and clinicians to visualize RNA expression. NGS applications are briefly described below.

(a) Whole-genome sequencing (WGS)

The determination of complete DNA sequence of an organism genome at a single time is termed as whole-genome sequencing (Van El et al. 2013). It allows the entire genome sequencing of a patient who has any disease or infected tissues to detect somatic or germ line variants (Wang et al. 2015). In addition, the NGS information, i.e., genetic/genomic profiles of patients, can be used to provide personalized medicine from the point of care to response of treatment and change treatment strategies across the condition of diseases. Personalized medicine can be used to increase the efficacy of treatment, diminish the side effects, and ultimately diminish overall cost to treat diseases for individual patients (Wang and Xu 2017).

(b) Whole-exome sequencing (WES)

Whole-exome sequencing (WES) sequenced the interested coding region of the DNA. Exome characterizes 1% of the whole-genome region, and approximately 85% of disease-causing mutations are expected to occur within the exome. Thus, WES has become the most preferred NGS method in multiple studies for clinicians (Bainbridge et al. 2010; Toledo et al. 2015; Pillai et al. 2017). Mutations in the coding or exon region can give rise to unclear clinical presentations which can be diagnosed by clinical providing many causes for a given condition or disease. With NGS, clinicians are provided a fast, affordable, and thorough way to determine the genetic cause of a disease. Several genes have been discovered in the process of

identification of genes that are relevant to inherited skin diseases using exome sequencing (Lai-Cheong and McGrath 2011). It can also facilitate the identification of disease-causing mutations in pathogenic presentations without knowing the exact genomic cause (Grada and Weinbrecht 2013). WES promises that it will accelerate to discover the genetic causes and contributors of disease in both the research and clinical settings (Singleton 2011).

(c) Targeted next-generation sequencing

Targeted sequencing is defined as the sequencing of specific genes or genomic regions which has been identified as a specific disease although whole-genome and whole-exome sequencing is available. Targeted sequencing is more affordable which can yield much higher coverage of genomic regions of interest by reducing sequencing cost and time. Researchers started developing the sequencing panels which target hotspots of disease-causing mutations (Grada and Weinbrecht 2013).

12.8 Biomarkers

Biomarker is a molecule or a set of molecules of DNA or RNA which can be measured and used for the characterization of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. It consists of genes, proteins, genetic variations, and differences in metabolic expression from different sources such as body fluids and tissues. The use of biomarkers was first described by Isaakson in 1980 when he proposed one of the most common biomarkers used and remained in urinary nitrogen as an independent measure of protein intake. Biomarkers can be classified depending upon the presence time like short-term (reflecting intake over past hours/days), medium-term (reflecting intake over weeks/months), and long-term biomarkers (reflecting intake over months/years). Biomarkers can be also classified based on the sequence of events from exposure to disease, i.e., diagnostic biomarkers, staging of disease biomarkers, disease prognosis biomarkers (cancer biomarkers), and biomarkers for monitoring the clinical response to an intervention (Mayeux 2004).

(a) Genetic Biomarkers

Genetic biomarkers are used in personalized medicine practices mainly for selecting those patients who benefit more or have less risk of adverse drug reactions (ADRs) from a particular drug. The biomarkers of drug-metabolizing gene study may lead to a better understanding of drug efficacy because each gene encoded proteins responsible for metabolizing exogenous toxicants including drugs and genetic polymorphism. The enzyme activity can be altered or abolished by gene deletions, missense, nonsense, and splice site mutations, whereas the action of enzyme distinctly modified amino acid substitutions caused by mutations. A good example of biomarkers enabling personalized medicine is the testing of mutations in the KRAS gene (Kirsten *ras*) to avoid treatment of patients with metastatic colorectal cancer from the use of panitumumab (Vectibix®) and cetuximab (Erbitux ®) who

cannot show any positive response to these medicines. KRAS is a protein encoded by the KRAS gene which plays important role in the RAS/MAPK signaling pathway involved in the regulation of cell division. Activating mutations in the KRAS gene make potent oncogenes, and the protein products of oncogenes play a role in many cancers by switching between active and inactive states, leading to cell transformation. It has been studied that the presence of mutations in the KRAS gene increased resistance to chemotherapy (panitumumab or cetuximab) and biological therapies targeting epidermal growth factor receptors. According to the US Food and Drug Administration (FDA) updated drug labels, *KRAS* mutation tests are recommended for the use of two anti-epidermal growth factor receptor (EGFR) antibody drugs in patients with metastatic colorectal cancer (Soulières et al. 2010).

(b) Epigenetic Biomarkers

The epigenetic alteration can cause the organism's genes to behave (or express themselves) differently. A biomarker which measures the epigenetic alterations that can be more directly associated with the phenotype observed and hold great promise for personalized medicine while in genomic alteration it can provide the information for developing some diseases and responses (efficacy and safety) to some drugs in precisely. An approach of epigenetic alteration is gene regulation involved in chromatin remodeling (such as DNA methylation and histone methylation), microRNAs (miRNAs), and other noncoding RNAs. The approach of epigenetic alteration includes chromatin remodeling that is achieved either by the posttranslational modification of the amino acids that make up histone proteins or by adding methyl groups to the DNA (most likely at CpG sites) to convert cytosine to 5-methylcytosine. Though epigenetic biomarkers of histone modify cation (Tai et al. 2010) and noncoding RNAs (He et al. 2005; Lu et al. 2005) have been reported, the most common ones are the biomarkers identified based on DNA methylation.

(c) Transcriptomic Biomarkers

DNA microarrays analyze the gene expression profiles by measuring expression levels of thousands of genes in a single experiment which can be applied across research and drug development. DNA microarray is a well-developed technology that measures transcriptome expression for specific cell/tissue and differential expressions between diseased and healthy populations. Therefore, microarrays are applied to discover transcriptomic biomarkers for personalized medicine. As an example, MammaPrint® is an FDA-approved DNA microarray-based diagnostic kit that could be used to measure the transcription level of 70 genes in breast cancer patients (Van't Veer et al. 2002; Van De Vijver et al. 2002; Glass et al. 2006; Buyse et al. 2006). The transcriptomic profiles were used for breast cancer patients who are less than 61 years old with stage I or stage II disease with tumor size $\leq$5.0 cm and who are lymph node negative (Hong et al. 2013).

(d) Metabolomic Biomarker

Worldwide metabolic profiling is being exercised to determine as the biological marker of healthiness. Certain small molecules are impeccably delicate markers of

being well. Metabolic profiling investigations are being utilized to decide biomarkers of medication in order to gain knowledge about the drug safety and effectiveness including the disease investigation and prediction. To know the mechanism of medication lethality and illness, a system biology approach that considers the knowledge generated from metabolic identification, genetics, transcriptomics, and genetics analysis framework is important. This can account into superior mechanism of drug interactions and diseases while recognizing characteristic-prone populations, an imperative objective in the push toward customized drug, i.e., personalized medicine (Schnackenberg and Beger 2007). Changes in the metabolome are intensified when compared to the progressions in transcriptome and proteome which are numerically more amenable. Requirement for entire genome arrangements or vast communicated succession label databases for every species is eliminated. The innovation is comprehensive as given metabolite (contrary to a transcript or protein) is the same in each creature that contains it. Metabolomic systems have thermodynamic and stoichiometric limitations that can make them easier to understand. Also metabolic profiling is substantially less expensive with higher throughput in comparison to proteomics and transcriptomics. The availability of compendia of genome-wide metabolomes and metabolic systems is accessible making it possible to look at vast number of tests from life forms that have been developed under extensive variety of conditions (Kell 2007). Flow infusion electrospray mass spectrometry was utilized to evaluate sputum for noninvasive biomarkers for lung cancer status. Statistical univariate and multivariate strategies were used to target key metabolites for hierarchical layering of lung cancer. Based on the studies by using artificial neural network system, six metabolites were segregated between non-small cell and small cell lung cancer (O'Shea et al. 2016). Similarly researchers have employed UPLC-ESI-MS-based small molecule profiling of exosome-like vesicles (ELVs) from human plasma and cell culture media. Scientists in this investigation have indicated that ELVs surely carry a rich metabolome that could not just enlarge the revelation of less abundant biomarkers yet may clarify in explaining the molecular basis of disease progression. Their approach could be effortlessly rendered to other studies seeking to develop predictive biomarkers that can be subsequently used with simplified targeted approaches (Altadill et al. 2016).

12.8.1 Omics Profiling

Omics relates to the scientific discipline which analyzes the interactions of biological information in life sciences that centers on large-scale data/information to understand life concluded in "omes" and "omics" such as genomics, proteomics, metabolomics, lipidomics, and so on (Yadav 2007). Omics techniques help in identifying, measuring, and quantifying DNA, messenger RNAs, and proteins derived from biological entities thereby enabling researchers to increase their ability to handle and measure large data sets (Morel et al. 2004). The development of big data analytics methods and data integration frameworks enables researchers to draw

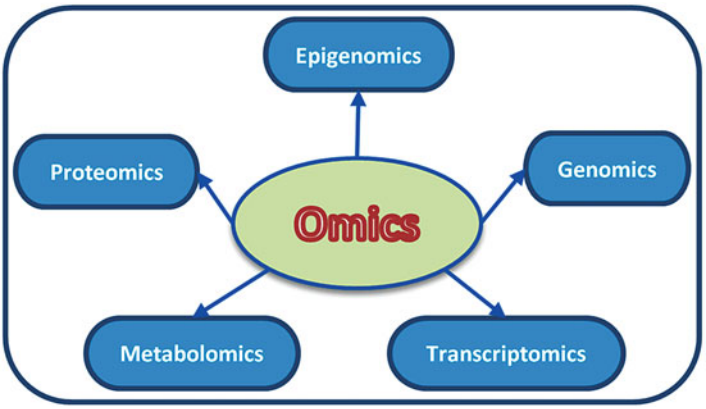

Fig. 12.3 Illustrations of the different omics technologies

inferences from diverse types of omics information to make molecular information on the diseased cells and efficient identification of concealed patterns from the data, giving further understanding of the biology of diseases and health states of individual patients. This will help in framing personalized treatment plans for each patient (Holzinger et al. 2014; Bellazzi and Zupan 2008). Thus omics profiling, exclusively the tumor subtyping, holds the promise of broadening cancer diagnosis and fostering the development of personalized cancer management. Clinical manifestation of genomics, epigenomics, transcriptomics, proteomics, and metabolomics information can delineate these findings to direct discoveries in precision medicine (Fig. 12.3).

12.8.1.1 Genomics

Genome sequencing presents the DNA arrangement changes of tumor tissues. Patient's tumor genome can be compared to the reference genome which will help in identification of genetic anomalies and their clinical implications (Vogelstein et al. 2013; Jones et al. 2015). Genomics profiling gives a methodical path for understanding the natural procedures supporting vital clinical phenotypes (Chen et al. 2012). As tumor cells harbor numerous hereditary alterations, many genes can be related to the phenotype by genomics research. Latest advancement in pathway studies deciphers about the biology of identified genes and proteins in cancer patients. Against a particular tumor, mapping a large number of altered genes or proteins into pathways magnifies the explanatory power and simplifies biological interpretations (Papin et al. 2004).

12.8.1.2 Epigenomics

Epigenomic changes, involving DNA methylation and chromatin modifications like histone acetylation, phosphorylation, ubiquitination, methylation, SUMOylation, ADP-ribosylation, deimination, and proline isomerization, may affect the expression patterns of genes. Changes in histone modification, upon alterations, can have diverse consequences on transcription. Cancer cells show many of these epigenomic changes during DNA methylation and chromatin modification (Iacobuzio-Donahue 2009; Kouzarides 2007). Profiling tools investigate the status of many types of epigenetic modification, whereas algorithms are designed to study effective pathway analysis (Hyun et al. 2015; Khatri et al. 2012).

12.8.1.3 Transcriptomics

The study of the whole set of messenger RNA (mRNA) transcripts in a cell and the amount of all transcript produced by the genome is known as transcriptomics. Microarrays were vastly used to figure the transcriptomic features of cancerous tissues. The gene expression levels mapped by microarrays can separate different types of hematologic cancer. Presently RNA-sequencing (RNA-Seq) is preferred for drafting gene expression levels (Wang et al. 2009; Golub et al. 1999).

12.8.1.4 Proteomics

The protein quantities and activities are affected in malignant cells due to distinct replication and metabolic processes. Assessing proteins and their modifications can ascertain health and diseased states. Mass spectrometry, protein arrays, and antibody-based detection methods are some of the experimental methods to analyze the proteomic profiles of cancer. Protein microarrays are another widely used analytical method for proteomics (Aebersold and Mann 2003; LaBaer and Ramachandran 2005).

12.8.1.5 Metabolomics

The study of the collection of metabolites in a biological system (e.g., cell, tissue, organ, organism) subjected to fixed condition is termed as metabolomics. Cancer cells have different metabolism and use different metabolic pathways than normal cells and can be exploited to harness insights into cancer biology and biomarker discovery through high-throughput profiling tools for metabolites. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) are routinely used technologies for metabolomics (Rochfort 2005; Shulaev 2006). Integration of different omics analyses through integration algorithms and tools can give comprehensive approach

about the disease primarily cancer biology leading to personalized tumor regulation. Readers are advised to refer review articles on omics profiling by Liu and Zhou (2014), Yu and Snyder (2016), and Bhati et al. (2012) for detailed information regarding the topic.

12.8.2 Nanomedicine

Applications of nanotechnology in medicine (nanomedicine) are involved in the design of nanoparticles to develop drugs, drug delivery techniques, diagnostics, medical devices and enhanced gene therapy, and tissue engineering procedures (Halappanavar et al. 2018). In nanomedicine, nanoscale tools are used for the diagnosis, prevention, and treatment of diseases and have the ability to enable early detection and prevention of diseases. Nanomedicines are included in the field of precision medicine and public health for treatment and therapeutic interventions in a wide variety of human diseases such as neuropsychiatric and neurodegenerative disorders, addictive disorders, inherited cancers, and cardiovascular diseases (Fornaguera and García-Celma 2017). With the help of nanoparticles, nanomedicine can provide molecular-level therapy to treat the diseases. Due to the presence of a large surface to volume ratio than traditional delivery vehicles, nanoparticles can be loaded with drugs, genes, antibodies, or radioactive materials to guide desired location of action and control its release at the therapeutic target to ensure an optimal concentration over a desired time frame. Nanomedicine includes several types of nanoparticles like liposomes, polymers, metals and metal oxides, and composites which are used in nano-drug delivery systems. These nanoparticles can be used as tools to understand the basic biological processes, drug delivery, imaging, and sensing (Wang and Wang 2014). The application of nanomedicine is categorized into three interrelated areas such as disease diagnostic tools, drug delivery, and tailoring medicine.

(a) Passive Tissue Targeting

Passive targeting, also known as physical targeting, is defined as the deposition of drug or drug nanovectors at a particular site due to physicochemical or pharmacological factors. As a result of an enhanced permeability and retention (EPR) effects, nanoparticles accumulated in the neoplastic tissues and increased vascular permeability, described by Maeda and colleagues (Matsumura and Maeda 1986). The normal tissues are not having any damage or rupture which allows the drugs to travel through blood vessels, whereas tumor tissues are highly unsystematic having leaky blood vessels and defective lymphatic drainage which allows migration of macromolecules surrounding tumor cells (Wang and Wang 2014). The optimal action of passive targeting depends on vascularization and angiogenesis, but damage and leaky blood vessels become a barrier to efficient transport of drugs. The feasibility of the passive targeting strategy to target tumor has several limitations such as all drugs cannot diffuse efficiently and unavailability of EFR effects in certain tumors. Therefore, to enhance the drug delivery and reduce cytotoxicity, an appropriate size

and surface properties can be provided with antibodies. Additionally, encapsulated drugs should be provided which must not diffuse out of the particle until the particle binds to the target. For example, the blood brain barrier can also be crossed to access the target sites for brain delivery in some inflammatory conditions (Kumar Khanna 2012). Liposomes, self-assembling colloid structure, can encapsulate a variety of drug molecules which significantly improves capability of drugs to target the tumors. Myocet and Caelyx are two first liposomes approved by the regulatory authorities. Liposomes composed of two lipid bilayers coated by aqueous cubicles, and both products contain doxorubicin and are entrapped in uncoated liposomes, but the latter is surface-modified with poly(ethylene glycol) (PEG) that diminishes the rapid recognition and prolongs the circulation time (Torchilin 2005). The encapsulated liposomal drugs have the ability to attenuate drug-related toxicity instead of decreasing the tumorigenesis. Polymorphic macromolecules are attached as linker with the therapeutic agents to passively accumulate by EPR effect in tumors and to significantly affect the therapeutic index (Duncan 2006). Abraxane, an albumin-bound paclitaxel, was the first polymer approved by the Food and Drug Administration (FDA) as passively tumor-targeted polymorphic nanomedicine (Reynolds et al. 2009).

(b) Active Targeting

Active targeting is denoted as specific modification of drug/drug-carrier nanosystems with active agent systems to act as homing devices for strapping to receptor structures expressed at the target site. This approach provides the widest opportunities and alternatives to serve as tumor-specific targets. This approach allows to import thousands of drug molecules by means of receptor-targeted ligands. Different types of moieties have been examined as targeting agents, including vitamins (Ideker and Sharan 2008), carbohydrates (Han 2008), aptamers, (Chuang et al. 2007), and peptides (e.g., Arg-Gly-Asp, allatostatin, transactivating transcriptional activator) (Segal et al. 2005; Albert et al. 2000; Araujo et al. 2007), whereas antibodies to the surface proteins and ligands for the receptors have been used in majority to target specific cells (Wakaskar 2017).

12.8.3 Molecular Imaging and Theranostics

Molecular imaging is a new medical discipline that integrates cell biology, molecular biology, and diagnostic imaging to visualize, characterize, and measure biological processes at the molecular and cellular levels in humans and other living systems. It includes two- or three-dimensional imaging as well as quantification over time by using nuclear medicine, computed tomography (CT), magnetic resonance imaging (MRI), fluorescence imaging, photoacoustic imaging, and ultrasound (US). Molecular imaging probes (endogenous molecules and exogenous probes) are used to visualize, characterize, and measure biological processes in living systems. Traditionally, molecular imaging probes consist of a targeting component and a signaling component based on small molecules. Recently new classes of molecular imaging

probes are in use including peptides, proteins, antibodies, aptamers, affibodies, and nanoparticles (Jung and Lee 2015). The nuclear medicine applications utilize devices such as single photon emission computerized tomography (SPECT) and positron emission tomography (PET) which provide target-specific information as well as function, pathway activities, and cell migration in the intact organism. Alzheimer's disease can be diagnosed by brain imaging using florbetapir as a diagnostic aid which is tagged with radioisotope fluorine-18. This complex binds with the aggregated amyloid beta and is visualized by PET (Souslova et al. 2013).

Furthermore, the assessment of disease treatment end points can yield to develop new lead molecules on the basis of personalized theranostic (image and treat) agents which can allow more accuracy in the selection of patients who may respond to treatment.

Acknowledgments SKS, CN, and IC thank the Department of Biotechnology (DBT), New Delhi for providing financial support.

Conflict of Interest The author(s) declare that there is no conflict of interest.

References

Abken H (2015) Adoptive therapy with CAR redirected T cells: the challenges in targeting solid tumors. Immunotherapy 7(5):535–544

Accessed from Leukemia and Lymphoma Society on 07/01/2017. https://www.lls.org/treatment/types-of-treatment/immunotherapy/chimeric-antigen-receptor-car-t-cell-therapy

Accessed from U.S. Food and Drug Administration on 07/01/2017. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm574058.htm

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422(6928):198–207

Agyeman AA, Ofori-Asenso R (2015) Perspective: does personalized medicine hold the future for medicine? J Pharm Bioallied Sci 7(3):239

AkilaKesavan G (2014) Nanotechnology and its applications. Scitech J 1(06):12–13

Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. Nature 406 (6794):378–382

Al-Mozaini MA, Mansour MK (2016) Personalized medicine: is it time for infectious diseases? Saudi Med J 37(12):1309

Altadill T, Campoy I, Lanau L, Gill K, Rigau M, Gil-Moreno A, Reventos J, Byers S, Colas E, Cheema AK (2016) Enabling metabolomics based biomarker discovery studies using molecular phenotyping of exosome-like vesicles. PLoS One 11(3):e0151339

Araujo RP, Liotta LA, Petricoin EF (2007) Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. Nat Rev Drug Discov 6(11):871–880

Araújo NP, Silva Kuhn GC, Svartman M (2017) Integrating next generation sequencing, bioinformatics and cytogenomics in the study of Brazilian mammals. Next Gener Seq Appl 4:147

Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA, Jeddeloh JA (2010) Whole exome capture in solution with 3 Gbp of data. Genome Biol 11(6):R62

Bellazzi R, Zupan B (2008) Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 77(2):81–97

Bethune MT, Joglekar AV (2017) Personalized T cell-mediated cancer immunotherapy: progress and challenges. Curr Opin Biotechnol 48:142–152

Bhardwaj A, Bhardwaj A, Misuriya A, Maroli S, Manjula S, Singh AK (2014) Nanotechnology in dentistry: present and future. J Int Oral Health 6(1):121

Bhati A, Garg H, Gupta A, Chhabra H, Kumari A, Patel T (2012) Omics of cancer. Asian Pac J Cancer Prev 13(9):4229–4233

Boisseau P, Loubaton B (2011) Nanomedicine, nanotechnology in medicine. Comptes Rendus Phys 12(7):620–636

Buyse M, Loi S, Van't Veer L, Viale G, Delorenzi M, Glas AM, Saghatchiand'Assignies M, Bergh J, Lidereau R, Ellis P, Harris A (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98 (17):1183–1192

Cai W, Gao T, Hong H, Sun J (2008) Applications of gold nanoparticles in cancer nanotechnology. Nanotechnol Sci Appl 1:17

Chan CW, Law BM, So WK, Chow KM, Waye MM (2017) Novel strategies on personalized medicine for breast cancer treatment: an update. Int J Mol Sci 18(11):2423

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148(6):1293–1307

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3(1):140

DeVita VT, Canellos GP (2011) Hematology in 2010: new therapies and standard of care in oncology. Nat Rev Clin Oncol 8(2):67–68

Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 2001 (344):1031–1037

Duncan R (2006) Polymer conjugates as anticancer nanomedicines. Nat Rev Cancer 6(9):688

Engin HB, Hofree M, Carter H. (2014) Identifying mutation specific cancer pathways using a structurally resolved protein interaction network. In: Pacific symposium on biocomputing co-chairs, pp 84–95

Fornaguera C, García-Celma MJ (2017) Personalized nanomedicine: a revolution at the nanoscale. J Personalized Med 7(4):12

Garraway LA, Lander ES (2013) Lessons from the cancer genome. Cell 153(1):17–37

Giese B, Klaessig F, Park B, Kaegi R, Steinfeldt M, Wigger H, Gleich A, Gottschalk F (2018) Risks, release and concentrations of engineered nanomaterial in the environment. Sci Rep 8 (1):1565

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103 (2):425–430

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Grada A, Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133(8):e11

Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, Ivanova Y, Hundal J, Arthur CD, Krebber WJ, Mulder GE (2014) Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature 515(7528):577

Gubin MM, Artyomov MN, Mardis ER, Schreiber RD (2015) Tumor neoantigens: building a framework for personalized cancer immunotherapy. J Clin Invest 125(9):3413–3421

Halappanavar S, Vogel U, Wallin H, Yauk CL (2018) Promise and peril in nanomedicine: the challenges and needs for integrated systems biology approaches to define health risk. Wiley Interdiscip Rev Nanomed Nanobiotechnol 1:10(1)

Han JD (2008) Understanding biological functions through molecular networks. Cell Res 18 (2):224–237
Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100(1):57–70
Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18(1):83
He ML, Mir PS, Beauchemin KA, Ivan M, Mir Z (2005) Effects of dietary sunflower seeds on lactation performance and conjugated linoleic acid content of milk. Can J Anim Sci 85(1):75–83
Holzinger A, Dehmer M, Jurisica I (2014) Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. BMC Bioinforma 15 (6):I1
Hong H, Zhang W, Su Z, Shen J, Ge W, Ning B, Fang H, Perkins R, Shi L, Tong W (2013) Next-generation sequencing (NGS): a revolutionary technology in pharmacogenomics and personalized medicine. In: Barh D, Dhawan D, Ganguly NK (eds) Omics for personalized medicine. Springer, New Delhi, pp 39–61, ISBN 978-81-322-1183-9
Horgan RP, Kenny LC (2011) 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. Obstet Gynaecol 13(3):189–195
Hunyadi Murph SE (2017) An Introduction to Nanotechnology. In: Hunyadi Murph S, Larsen G, Coopersmith K (eds) Anisotropic and shape-selective nanomaterials. Nanostructure science and technology. Springer, Cham, pp 3–5, Print ISBN 978-3-319-59661-7
Hyun BR, McElwee JL, Soloway PD (2015) Single molecule and single cell epigenomics. Methods 72:41–50
Iacobuzio-Donahue CA (2009) Epigenetic changes in cancer. Annu Rev Pathol Mech Dis 4:229–249
Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18(4):644–652
Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, Shukla M, Chesnick B, Kadan M, Papp E, Galens KG (2015) Personalized genomic analyses for cancermutation discovery and interpretation. Sci Transl Med 7(283):283ra53
Jung KH, Lee KH (2015) Molecular imaging in the era of personalized medicine. J Pathol Transl Med 49(1):5
Kchouk M, Gibrat JF, Elloumi M (2017) Generations of sequencing technologies: from first to next generation. Biol Med 9(3)
Kell DB (2007) Metabolomic biomarkers: search, discovery and validation. Expert Rev Mol Diagn 7(4):329–333
Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8(2):e1002375
Kouzarides T (2007) Chromatin modifications and their function. Cell 128(4):693–705
Kumar Khanna V (2012) Targeted delivery of nanomedicines. ISRN Pharmacol 10:2012
LaBaer J, Ramachandran N (2005) Protein microarrays as tools for functional proteomics. Curr Opin Chem Biol 9(1):14–19
Lai-Cheong JE, McGrath JA (2011) Next-generation diagnostics for inherited skin disorders. J Investig Dermatol 131(10):1971–1973
Lee MS, Flammer AJ, Lerman LO, Lerman A (2012) Personalized medicine in cardiovascular diseases. Kor Circ J 42(9):583–591
Liu X, Zhou L (2014) Mini review: the application of omics in targeted anticancer biopharmaceuticals development. Austin. J Biomed Eng 1(1):1–8
Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR (2005) MicroRNA expression profiles classify human cancers. Nature 435(7043):834–838
Matsumura Y, Maeda H (1986) A new concept for macromolecular therapeutics in cancer chemotherapy: mechanism of tumoritropic accumulation of proteins and the antitumor agent smancs. Cancer Res 46(12 Part 1):6387–6392
Mayeux R (2004) Biomarkers: potential uses and limitations. NeuroRx 1(2):182–188
Milone MC, Fish JD, Carpenito C, Carroll RG, Binder GK, Teachey D, Samanta M, Lakhal M, Gloss B, Danet-Desnoyers G, Campana D (2009) Chimeric receptors containing CD137 signal

transduction domains mediate enhanced survival of T cells and increased antileukemic efficacy in vivo. Mol Ther 17(8):1453–1464
Mody VV, Siwale R, Singh A, Mody HR (2010) Introduction to metallic nanoparticles. J Pharm Bioallied Sci 2(4):282
Morel NM, Holland JM, van der Greef J, Marple EW, Clish C, Loscalzo J, Naylor S (2004) Primer on medical genomics Part XIV: introduction to systems biology—a new approach to understanding disease and treatment. In: Mayo clinic proceedings May 31, Elsevier, vol. 79, no. 5, pp 651–658
Moreth J, Mavoungou C, Schindowski K (2013) Passive anti-amyloid immunotherapy in Alzheimer's disease: what are the most promising targets? Immun Ageing 10(1):18
Morgan D (2011) Immunotherapy for Alzheimer's disease. J Intern Med 269(1):54–63
O'Shea K, Cameron SJ, Lewis KE, Lu C, Mur LA (2016) Metabolomic-based biomarker discovery for non-invasive lung cancer screening: a case study. Biochim Biophys Acta (BBA)-Gen Subj 1860(11):2682–2687
Pagel JM, West HJ (2017) Chimeric antigen receptor (CAR) T-cell therapy. JAMA Oncol 3 (11):1595
Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004) Comparison of network-based pathway analysis methods. Trends Biotechnol 22(8):400–405
Pearson ER (2016) Personalized medicine in diabetes: the role of 'omics' and biomarkers. Diabet Med 33(6):712–717
Pillai S, Gopalan V, Lam AK (2017) Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. Crit Rev Oncol Hematol 116:58–67
Rakesh M, Divya TN, Vishal T, Shalini K (2015) Applications of nanotechnology. J Nanomedine Biotherapeutic Discov 5:131. https://doi.org/10.4172/2155-983X.1000131
Reynolds C, Barrera D, Jotte R, Spira AI, Weissman C, Boehm KA, Pritchard S, Asmar L (2009) Phase II trial of nanoparticle albumin-bound paclitaxel, carboplatin, and bevacizumab in first-line patients with advanced nonsquamous non-small cell lung cancer. J Thorac Oncol 4 (12):1537–1543
Rochfort S (2005) Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. J Nat Prod 68(12):1813–1820
Sadelain M, Brentjens R, Rivière I (2013) The basic principles of chimeric antigen receptor design. Cancer Discov 3(4):388–398
Sawyers CL (2007) Cancer: mixing cocktails. Nature 449(7165):993–996
Schnackenberg LK, Beger RD (2007) Metabolomic biomarkers: their role in the critical path. Drug Discov Today Technol 4(1):13–16
Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. Nat Genet 37:S38–S45
Shu Y, Sheardown SA, Brown C, Owen RP, Zhang S, Castro RA, Ianculescu AG, Yue L, Lo JC, Burchard EG, Brett CM (2007) Effect of genetic variation in the organic cation transporter 1 (OCT1) on metformin action. J Clin Investig 117(5):1422
Shulaev V (2006) Metabolomics technology and bioinformatics. Brief Bioinform 7(2):128–139
Silva GA (2004) Introduction to nanotechnology and its applications to medicine. Surg Neurol 61 (3):216–220
Singleton AB (2011) Exome sequencing: a transformative technology. Lancet Neurol 10 (10):942–946
Smith AJ, Oertle J, Warren D, Prato D (2016) Chimeric antigen receptor (CAR) T cell therapy for malignant cancers: summary and perspective. J Cell Immunother 2(2):59–68
Soulières D, Greer W, Magliocco AM, Huntsman D, Young S, Tsao MS, Kamel-Reid S (2010) KRAS mutation testing in the treatment of metastatic colorectal cancer with anti-EGFR therapies. Curr Oncol 17(Suppl 1):S31
Souslova T, Marple TC, Spiekerman AM, Mohammad AA (2013) Personalized medicine in Alzheimer's disease and depression. Contemp Clin Trials 36(2):616–623

Tai W, Mahato R, Cheng K (2010) The role of HER2 in cancer therapy and targeted drug delivery. J Control Release 146(3):264–275

Tebani A, Afonso C, Marret S, Bekri S (2016) Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations. Int J Mol Sci 17(9):1555

Toebes M, Coccoris M, Bins A, Rodenko B, Gomez R, Nieuwkoop NJ, van de Kasteele W, Rimmelzwaan GF, Haanen JB, Ovaa H, Schumacher TN (2006) Design and use of conditional MHC class I ligands. Nat Med 12(2):246

Toledo RA, Qin Y, Cheng ZM, Gao Q, Iwata S, Silva G, Prasad M, Ocal IT, Rao S, Aronin N, Barontini MB (2015) Recurrent mutations of chromatin remodeling genes and kinase receptors in pheochromocytomas and paragangliomas. Clin Cancer Res clincanres-1841

Torchilin VP (2005) Recent advances with liposomes as pharmaceutical carriers. Nat Rev Drug Discov 4(2):145

Van De Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999–2009

Van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, Howard HC, Cambon-Thomsen A, Knoppers BM, Meijers-Heijboer H, Scheffer H (2013) Whole-genome sequencing in health care. Eur J Hum Genet 21:S1–S5

van Rooij N, van Buuren MM, Philips D, Velds A, Toebes M, Heemskerk B, van Dijk LJ, Behjati S, Hilkmann H, el Atmioui D, Nieuwland M (2013) Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. J Clin Oncol Off J Am Soc Clin Oncol 31(32)

Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, Van Der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW (2013) Cancer genome landscapes. Science 339(6127):1546–1558

Vogenberg FR, Barash CI, Pursel M (2010) Personalized medicine: part 1: evolution and development into theranostics. Pharm Ther 35(10):560

Vucic EA, Thu KL, Robison K, Rybaczyk LA, Chari R, Alvarez CE, Lam WL (2012) Translating cancer 'omics' to improved outcomes. Genome Res 22(2):188–195

Wakaskar RR (2017) Passive and active targeting in tumor microenvironment. Int J Drug Dev Res 9 (2):19

Wang EC, Wang AZ (2014) Nanoparticles and their applications in cell and molecular biology. Integr Biol 6(1):9–26

Wang L, Xie XQ (2016) Cancer genomics: opportunities for medicinal chemistry? Future Med Chem 8(4):357–359

Wang K, Xu C (2017) Applications of next-generation sequencing in cancer research and molecular diagnosis. J Clin Med Genom 5:147

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63

Wang Q, Lu Q, Zhao H (2015) A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet 6:149

Winblad B, Graf A, Riviere ME, Andreasen N, Ryan JM (2014) Active immunotherapy options for Alzheimer's disease. Alzheimers Res Ther 6(1):7

Wraith DC (2017) The future of immunotherapy for cancer and autoimmune diseases: a 20 year perspective. Front Immunol 8:1668

Yadav SP (2007) The wholeness in suffix -omics, -omes, and the word om. J Biomol Tech 18 (5):277

Yu KH, Snyder M (2016) Omics profiling in precision oncology. Mol Cell Proteomics 15 (8):2525–2536

Chapter 13
Characterization of Plant Genetic Modifications Using Next-Generation Sequencing

Ana Pérez-González, Álvaro Eseverri, and Elena Caro

Abstract In the last few years, many studies have demonstrated that next-generation sequencing (NGS) technologies can facilitate the detection and molecular characterization of genetically modified organisms (GMOs). T-DNA localization and copy number determination in transgenic plants are very useful in basic research projects because of its implications in transgene expression level and stability, and are absolutely necessary for the commercialization of a GMO. The high throughput of NGS together with its continuously decreasing cost makes it a very rapid, cost-effective, and efficient tool for this task, faster and less laborious than the classical Southern blot and genome walking techniques. Moreover, the recent development of bioinformatics tools designed for users with no specific knowledge of computer science makes this approach affordable to the whole scientific community. Successful wet lab strategies and bioinformatics pipelines reported in the literature will be reviewed and discussed here.

Keywords NGS · GMO · Synthetic biology · WGR · Bioinformatics tools

Plant synthetic biology uses engineering principles to genetically modify plants creating or enhancing beneficial traits or to produce valuable products. With this purpose, genetic modules are combined in different ways to build new modules that exhibit predictable behaviors. These genetic modules are then introduced into a recipient organism's native genome via stable integration of T-DNAs into a plant genome (Fig. 13.1).

Upon *Agrobacterium*-mediated transformation, single or tandem T-DNA copies are usually integrated into one or two loci of the plant genome, sometimes with

Ana Pérez-González and Álvaro Eseverri. These authors contributed equally to the work.

A. Pérez-González · Á. Eseverri · E. Caro (✉)
Centre for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Campus Montegancedo UPM, Madrid, Spain
e-mail: elena.caro@upm.es

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_13

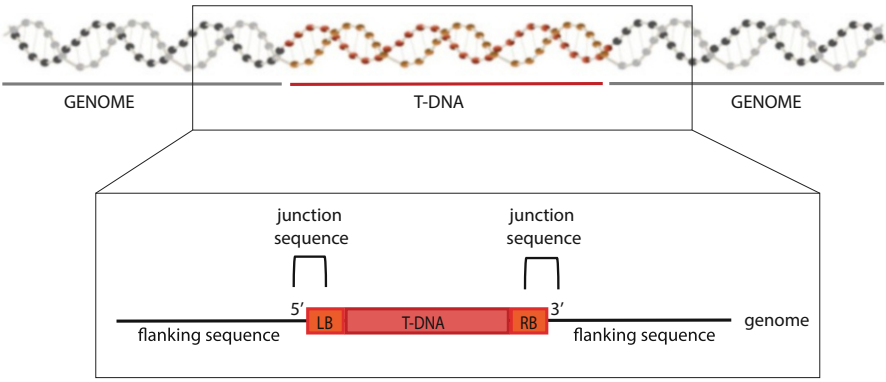

Fig. 13.1 Representation of the stable integration of a T-DNA into a plant genome. Once integrated, junction sequences are created between the T-DNA boundaries and the site of the host genome where it has inserted. Flanking sequences are the original host genome sequences that surround the inserted T-DNA. *LB* left border, *RB* right border

various rearrangements of the target site: T-DNAs are truncated at their left border at low frequency, and in some cases, vector backbone DNA is integrated too. In contrast, direct DNA transfer (like particle bombardment) renders high-copy transgenic loci and extensive rearrangements of the foreign DNA (Pérez-González and Caro 2016).

The existence of repeat-sensitive transcriptional repression mechanisms, described long ago in plants and animals, establishes that single gene copies at a defined locus are expressed much more effectively than reiterated transgenes (Wolffe 1997). Thus, the details about the insertion locus, the number of copies of the inserted T-DNA, and its structure are relevant since they can play an important role in determining if transgenes are correctly expressed or become silenced.

Also, detailed descriptions of the genetic modifications are required for risk evaluation prior to commercialization and release of GMOs into the environment. Legislation asks specifically that an applicant must provide information on the size and copy number of inserts and their organization and sequence information for both 5′ and 3′ flanking regions, with the aim of identifying interruptions of known host genes (Guidance for risk assessment of food and feed from genetically modified plants 2011).

In the past, this characterization was carried out using "classical" molecular biology techniques. The traditional way to identify foreign DNA integration was through Southern blot analysis, using a sequence-specific probe homologous to the transgene, which is very time-consuming although efficient in confirming the T-DNA copy number. For the identification of the insertion site, many PCR-based methods have been successfully used, such as thermal asymmetric interlaced PCR (TAIL-PCR) (Liu et al. 1995), adapter-ligated PCR (O'Malley and Ecker 2010), inverse PCR (IPCR) (Ochman et al. 1988), and restriction site extension PCR (RSE-PCR) (Ji and Braam 2010), but these methods present important limitations.

They all rely on having accurate sequence information of the integrated T-DNA, and sometimes huge rearrangements can take place, leading to very complex situations where the results are difficult to interpret. Other limitations of these PCR-based methods are the need of restriction enzymes that cut both the T-DNA and the genome at a convenient distance and the generation of non-specific products through PCR (Ji and Braam 2010). In any case, these approaches are always laborious and expensive and especially difficult to use in high throughput.

Whole genome re-sequencing (WGR) offers a good alternative to these conventional techniques with great effectiveness and a rapidly decreasing cost. Several publications have reported its use to detect the exact copy number, structure, and genomic location of transgenes in different species of plants, apart from detecting the presence of vector backbone and assessing the stability of T-DNAs across generations.

13.1 Workflow

In the last decades, the apparition of next-generation sequencing techniques has had a great impact on research. It has increased the throughput and speed of sequencing and allowed a decrease on the cost of the process, leading to its application in many fields. Commercial second-generation high-throughput sequencing platforms (Roche/454, Illumina HiSeq or MiSeq, ABI SOLiD, etc.), independently of the method they use for sequencing, all follow a somehow similar workflow, as described below (Kulski 2016) (Fig. 13.2).

13.1.1 Library Preparation

A library needs to be prepared by randomly breaking the DNA into small fragments and adding common adapters to their ends. Later, the template DNA will be primed by an oligonucleotide complementary to the adapter sequence.

13.1.2 Clonal Amplification

In order to achieve a detectable signal for sequencing, the DNA fragments from the library need to be clonally amplified. This can be done on a solid surface or on beads while isolated within miniature emulsion droplets or arrays. In any case, after several rounds of amplification, DNA clusters are generated, and sequencing can proceed.

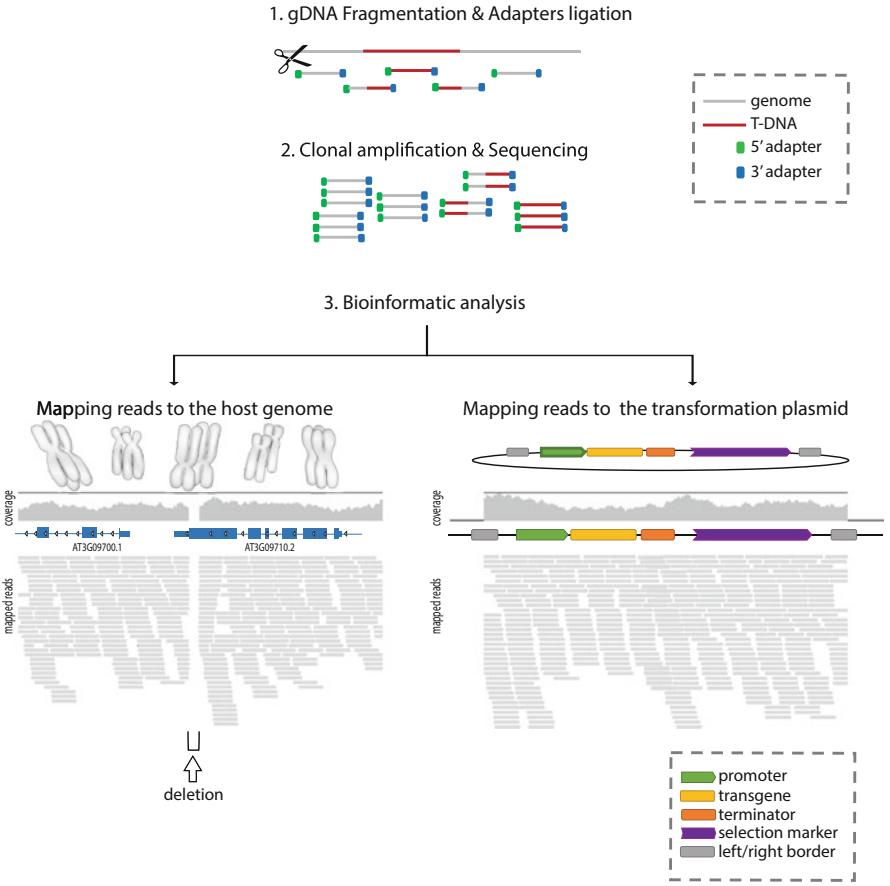

Fig. 13.2 General workflow of a whole genome re-sequencing experiment. (1) Genomic DNA (gDNA) is broken into small fragments, and adapters are added to 5′- and 3′-ends. (2) Small DNA fragments are clonally amplified. (3) Examples of sequencing reads mapping to the host genome and to the transformation plasmid. Deletion refers to a piece of the host genome missing after the insertion of the T-DNA

13.1.3 Cyclic Array Sequencing

Sequencing with wash-and-scan techniques is done following different methods. In *sequencing-by-synthesis* approaches, a complementary strand is synthesized in the presence of a polymerase enzyme. In the case of pyrosequencing, the release of pyrophosphate is detected when nucleotides are added to the DNA chain. This is the method applied on the Roche/454 platform, the first to revolutionize sequencing technology.

Another type of sequencing-by-synthesis approach is the case of cyclic reversible termination, in which during each cycle, a fragment of DNA template combined with

an adapter incorporates just one nucleotide, since the blocked 3′ group prevents additional incorporations. This is the method applied on the Illumina platform, currently the most cost-effective and most widely used sequencer.

In *sequencing by ligation*, the polymerization reaction is replaced by a ligation reaction. This is the method applied on ABI SOLiD sequencers, which render high accuracy.

A new generation of sequencing methods has been developed around the single-molecule sequencing technology. This so-called third-generation sequencing can be done without creating the time-consuming and costly amplification libraries and, thus, eliminating the errors caused by PCR. The read length of third-generation sequencing methods can reach several kilobases, thereby allowing the resolution of highly complex genomes with many long repetitive elements, copy number, and structural variations. The most popular of the third-generation sequencing platforms is the single-molecule real-time (SMRT) sequencing method used by Pacific Biosciences (PacBio).

13.1.4 Bioinformatics Analysis

Once WGR is performed, bioinformatics tools are used to identify differences between the DNA of the specific individual sequenced (transgenic organism) and that of a reference genome (wild type) (Fig. 13.3).

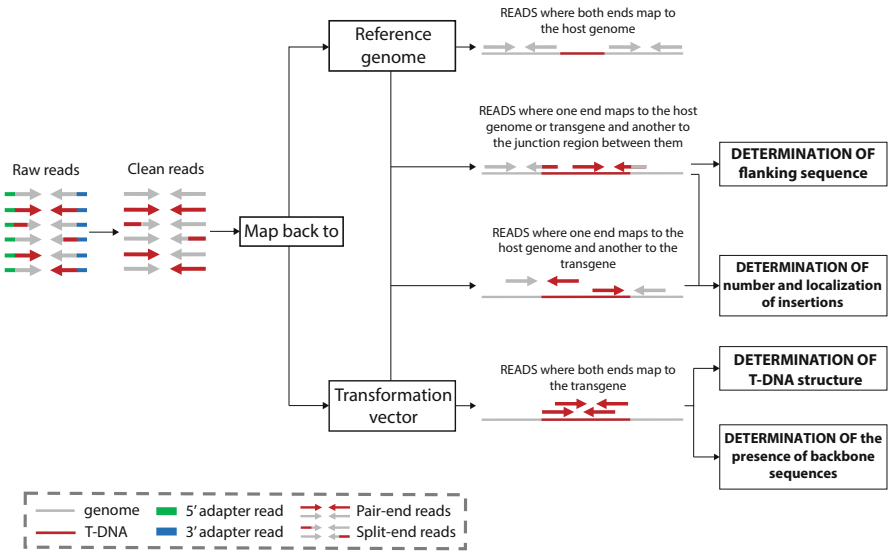

Fig. 13.3 Detailed bioinformatics pipeline for the analysis of sequencing reads in a whole genome re-sequencing experiment

T-DNA structure and rearrangements can be determined by the calculation of insert size distributions using the reads that map against the transformation plasmid sequence (Park et al. 2017).

If some of these reads map to the vector used for transformation, the presence of *vector backbone-derived sequences* in the genome can also be determined (its absence must be confirmed for safety assessment of GM crops).

And most importantly, the *determination of the T-DNA insertion site and genomic flanking sequences* can be done by analyzing the reads that match the host genome only partially and also map to the T-DNA sequence (Holst-Jensen et al. 2016).

13.2 Reported Strategies

While this is the general workflow of a WGR experiment, several specific approaches have been described to map and characterize T-DNA insertions.

The first report dates to 2012, when Polko et al. (2012) conducted a forward genetic screen to identify molecular components involved in controlling hyponastic growth in *Arabidopsis thaliana*. To identify T-DNA insertion loci in selected candidates, they first tried plasmid rescue and TAIL-PCR, but both methods repeatedly failed. Therefore, they adopted a novel approach using pooled DNA and an Illumina Genome Analyzer to produce over two million of 50-bp paired reads. The reads were mapped to the *Arabidopsis* genome, the T-DNA sequence, or both. The last ones were the ones used to detect the T-DNA insertion loci, showing for the first time an effective method to map T-DNA insertions in *Arabidopsis thaliana* without sequencing entire genomes.

On that same year, Kovalic et al. (2012) published a very thorough characterization of the genetic modification of a soybean line using WGR. They performed paired-end sequencing with an Illumina HiSeq sequencer and generated hundreds of millions of 100-bp reads that covered the genome at 75 times average coverage. Reads mapping partially to both, the transformation vector and the host plant genome, were used to map potential insertion sites that were then confirmed by conventional techniques like PCR and sequencing of the amplification products. The molecular characterization of the genetic modifications present in two herbicide-tolerant transgenic soybeans has been recently carried out following this method for regulatory submissions prior to commercialization and use in breeding programs (Guo et al. 2016).

Wahler et al. (2013) applied basically the same approach for the study of event LL62 rice, but in contrast to the situations described before, here the sequence of the transformation vector used for the genetic modification was unknown to the laboratory performing the analysis. Around 172 million of paired-end reads of 75 bp in length were obtained with an Illumina HiSeq sequencer, which represented an average 65 times coverage of the rice genome. The reads were mapped to the rice reference genome, and breakpoints could be detected. The breakpoint border

sequences were then mapped against the pCAMBIA-1300 vector sequence, and those that aligned even partially were assembled to form a putative T-DNA sequence. All reads were then mapped to this putative T-DNA sequence in a strategy named "bioinformatics read walking." Although pCAMBIA-1300 was not itself used to generate GM rice LLRice62, it contains nucleotide sequences commonly used in plant genetic engineering, like the pUC18 multiple cloning site, the aadA gene, the hptII gene, and the cauliflower mosaic virus 3′UTR and 35S promoter. The authors succeeded in developing a general approach that allows the identification of likely transgenic breakpoint border sequences, applicable to all GMOs.

Yang et al. (2013) also used WGR and bioinformatics to characterize the insertion of T-DNAs in rice when the a priori knowledge of the insertion is limited. They sequenced two transgenic rice lines with an Illumina HiSeq sequencer and obtained around 100 million of paired-end reads of 90 bp in length, an average of 25 times sequencing coverage of the rice genome. The data was analyzed considering three different scenarios, and for each case, a specific bioinformatics module was designed. Module 1 was designed for situations in which the DNA sequence of the transformation vector is completely known, an approach comparable to that of Kovalic et al. (2012). Module 2 was designed for use when the sequence of the inserted DNA is unknown, but a database of genetic elements and transgene constructs from known GMOs is available and can be used as a reference library. This approach is, thus, comparable to the one taken by Wahler et al. (2013). Module 3 was designed for use when the analyzing laboratory has absolutely no a priori knowledge of the DNA sequence of the T-DNA. In this case, first the reads were mapped to the rice genome, before de novo assembly and BLAST analysis of the retained reads. Although the results obtained were very satisfactory, the bioinformatics workload in case of the use of modules 2 and 3 was still enormous, and an automatic and simplified software needs to be developed for these strategies to be adopted by the scientific community.

A constant in all methods is that only a small fraction of the obtained reads, those mapping to the transformation plasmid, are used for the determination of the insertion site. Most recent methods that use NGS to map T-DNAs do not consider necessary to sequence the whole host genome with great depth but include strategies involving capture enrichment approaches to improve coverage of the inserted and adjacent sequences only.

Lepage et al. (2013) described "targeted genomic sequencing," a new technique that allows the simultaneous identification of multiple insertion sites in a complex DNA sample using biotinylated primers specific for the extremities of the T-DNA. The biotinylated primers were hybridized to a library consisting on a pool of equivalent amounts of genomic DNA from 64 lines, and the recovered DNA was used for Roche/454 sequencing and identification of the region flanking the T-DNA in each line. No analysis of potential rearrangements of the insertions was performed, just identification of the insertion sites.

Another related approach, referred to as "Southern-by-Sequencing," has been reported by Zastrow-Hayes et al. (2015). It also consists on the hybridization of indexed and pooled genomic DNA libraries from transgenic plants to biotinylated

probes designed against the sequence of the T-DNA. Sequencing of recovered DNA and analysis of the obtained reads revealed the sequences adjacent to the T-DNA insertions, and thus the number of insertions and their rearrangements could be inferred.

Inagaki et al. (2015) published a work exploring the application of sequence capture-based methods and NGS for high-throughput identification of T-DNA insertion site and structure. They custom-developed methods using a mixture of biotinylated hybridization probes targeted against the various T-DNA ends for target enrichment and bioinformatics tools to determine the location of T-DNA insertions in the *Arabidopsis* genome.

Guttikonda et al. (2016) showed WGR and target capture can be applied to the molecular characterization for regulatory submissions of single and stacked transgenic events. They characterized two transgenic soybean lines and their hybrid stack using paired-end sequencing and hybridization to a BAC clone including the transformation vector. When they compared their results with those obtained by traditional techniques, NGS showed a significant advantage at detecting small rearrangements of the T-DNAs.

13.3 Discussion

13.3.1 Read Length

The read length obtained after sequencing can be determinant on the feasibility of the generation of genetic maps of transgene inserts to determine insertion sites and T-DNA rearrangements. Roche 454 yields relatively long reads up to 700 bp but at a relatively high price, so most of the published studies from the last years apply Illumina technology, typically yielding short read lengths (50–300 nt). However, Illumina technology makes possible paired-end sequencing, what facilitates detection of the insertion loci of T-DNAs. With recent platforms such as PacBio and MinION, it is possible to obtain very long reads (several thousand base pairs), potentially useful to complete transgenic insert sequences (Holst-Jensen et al. 2016). Combinations of platforms and strategies have proven ideal to obtain detailed and verified information.

For example, Scouten et al. (2017) reported difficulties in assembling short Illumina reads for characterization of inserts, due to multiple inserts per plant, and PacBio sequencing appeared to be helpful to solve this complex assembly.

Another challenge faced when using short reads is the detection of a junction locus within plant-repetitive genome sequences. Illumina short reads may cause an increase in the percentage of ambiguously or incorrectly mapped reads, and this limitation can be overcome also by using sequencing platforms that generate long-read sequencing data (Park et al. 2017).

13.3.2 Coverage

A balance must be achieved between the benefits of a big amount of raw data and its high cost and time consumption. It is obvious that with a deeper sequencing and a higher coverage, the analysis leading to the molecular characterization of a transgenic line becomes easier (Park et al. 2017), but an alternative to increasing sequencing depth can be to perform a fraction enrichment step in the protocol before sequencing to enrich the target sequence (as discussed above Lepage et al. 2013; Zastrow-Hayes et al. 2015; Inagaki et al. 2015; Guttikonda et al. 2016). However, Lambirth et al. (2015) conducted paired-end sequencing of a soy sample on the Illumina HiSeq 2000 system to around five times theoretical genome-wide coverage with 100 base-pair reads, and such low coverage was reported enough to locate and identify a single-copy transgene insertion in a highly complex and repetitive genome like that of soybean. The use of lower coverage can be convenient but is dependent on the existence of a reference genome to facilitate alignments and produces information that needs posterior verification using traditional molecular biology techniques.

13.3.3 Sequenced Reference Genome

Sometimes it is necessary to sequence and de novo assemble a complete genome to study a specific genetic modification. Nowadays it might be feasible on a small computer cluster or even on a single high-quality machine, but the genomes of higher eukaryotes are large and complex and require a certain effort to sequence and assemble (Holst-Jensen et al. 2016).

The "SunUp" papaya was the first transgenic plant where the genome was fully sequenced and assembled in an effort to characterize the genetic modification that conferred it virus resistance (Ming et al. 2008). Sequencing depth was very low though (3 times coverage), and Southern blot experiments were needed to complement the NGS data.

13.3.4 Alignment Algorithm

For high-throughput NGS data processing, a read alignment tool with an algorithm is used, and the choice of that aligner tool can lead to very different efficiencies in genome-T-DNA junction detection and different computational running times. Park et al. (2017) tested different alignment programs and their combinations and compared the results obtained with them in the analysis of a transgenic rice line. In their paper, they report that the BWA-MEM combination showed the best performance

and it had longer runtimes than Bowtie or the BWA-aln algorithm, but the results obtained proved more accurate and reliable.

In any case, since many researchers have difficulties in handling large quantities of bioinformatics data and applying algorithms, there are now user-friendly methods that handle NGS data and help in the detection of genome-T-DNA junctions, that do not require almost any computational science background knowledge, and that can be used by most biologists (Park et al. 2017; Lambirth et al. 2015).

13.4 Conclusion

NGS can facilitate the molecular characterization of GMOs in light of its high throughput, continuously decreasing costs and the development of a diverse range of bioinformatics tools that allow transgene identification, location, and characterization. Compared with traditional Southern blot and PCR-based methods, whole genome re-sequencing has become a faster, cheaper, simpler, and more effective approach for transgenic modification analysis.

References

Guidance for risk assessment of food and feed from genetically modified plants (2011) EFSA J 9 (5):2150

Guo B, Guo Y, Hong H, Qiu L-J (2016) Identification of genomic insertion and flanking sequence of G2-EPSPS and GAT transgenes in soybean using whole genome sequencing method. Front Plant Sci 7:1009

Guttikonda SK, Marri P, Mammadov J, Ye L, Soe K, Richey K et al (2016) Molecular characterization of transgenic events using next generation sequencing approach ed: Jain M. PLoS One 11(2):e0149515

Holst-Jensen A, Spilsberg B, Arulandhu AJ, Kok E, Shi J, Zel J (2016) Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products. Anal Bioanal Chem 408(17):4595–4614

Inagaki S, Henry IM, Lieberman MC, Comai L (2015) High-throughput analysis of T-DNA location and structure using sequence capture ed: Candela H. PLoS One 10(10):e0139672

Ji J, Braam J (2010) Restriction site extension PCR: a novel method for high-throughput characterization of tagged DNA fragments and genome walking ed: Herrera-Estrella A, editor. PLoS One 5(5):e10577

Kovalic D, Garnaat C, Guo L, Yan Y, Groat J, Silvanovich A et al (2012) The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern. Biotechnology 5(3)

Kulski JK (2016) Next-generation sequencing — an overview of the history, tools, and "Omic" applications. In: Next generation sequencing – advances, applications and challenges. InTech, Croatia

Lambirth KC, Whaley AM, Schlueter JA, Bost KL, Piller KJ (2015) CONTRAILS: a tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing. Genomics Data 6:175–181

Lepage É, Zampini É, Boyle B, Brisson N (2013) Time- and cost-efficient identification of T-DNA insertion sites through targeted genomic sequencing. PLoS One 8(8):e70912

Liu YG, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J 8(3):457–463

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452(7190):991

O'Malley RC, Ecker JR (2010) Linking genotype to phenotype using the Arabidopsis unimutant collection. Plant J 61(6):928–940

Ochman H, Gerber AS, Hartl DL (1988) Genetic applications of an inverse polymerase chain reaction. Genetics 120(3):621–623

Park D, Park S-H, Ban YW, Kim YS, Park K-C, Kim N-S et al (2017) A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. BMC Biotechnol 17(1):67

Pérez-González A, Caro E (2016) Hindrances to the efficient and stable expression of transgenes in plant synthetic biology approaches. In: Systems biology application in synthetic biology. Springer India, New Delhi, pp 79–89

Polko JK, Temanni M-R, van Zanten M, van Workum W, Iburg S, Pierik R et al (2012) Illumina sequencing technology as a method of identifying T-DNA insertion loci in activation- tagged *Arabidopsis thaliana* plants. Mol Plant 5:948–950

Schouten HJ, van de Geest H, Papadimitriou S, Bemer M, Schaart JG, Smulders MJM et al (2017) Re-sequencing transgenic plants revealed rearrangements at T-DNA inserts, and integration of a short T-DNA fragment, but no increase of small mutations elsewhere. Plant Cell Rep 36 (3):493–504

Wahler D, Schauser L, Bendiek J, Grohmann L (2013) Next-generation sequencing as a tool for detailed molecular characterisation of genomic insertions and flanking regions in genetically modified plants: a pilot study using a rice event unauthorised in the EU. Food Anal Methods 6 (6):1718–1727

Wolffe AP (1997) Transcription control: repressed repeats express themselves. Curr Biol 7(12): R796–R798

Yang L, Wang C, Holst-Jensen A, Morisset D, Lin Y, Zhang D (2013) Characterization of GM events by insert knowledge adapted re-sequencing approaches. Sci Rep 3(1):2839

Zastrow-Hayes GM, Lin H, Sigmund AL, Hoffman JL, Alarcon CM, Hayes KR et al (2015) Southern-by-sequencing: a robust screening approach for molecular characterization of genetically modified crops. Plant Genome 8(1):0