Prashanth N. Suravajhala *Editor*

Your Passport to a Career in Bioinformatics

Second Edition

Springer

Your Passport to a Career in Bioinformatics

Prashanth N. Suravajhala
Editor

Your Passport to a Career in Bioinformatics

Second Edition

Springer

Editor
Prashanth N. Suravajhala
Department of Biotechnology
and Bioinformatics
Birla Institute of Scientific Research
Jaipur, India

Bioclues Organization
Hyderabad, India

ISBN 978-981-15-9543-1 ISBN 978-981-15-9544-8 (eBook)
https://doi.org/10.1007/978-981-15-9544-8

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

To my Mother Nirmala Sastry

Foreword

When I first heard about the field of bioinformatics, I was a university senior majoring in chemistry. It was 1995, and my intention at the time was to focus on the application of chemistry in the life sciences. In fact, in those days I was interested in any field of science or engineering that could be applied to biology. But, when it came time to select a project for my senior thesis, I was asked by my thesis adviser if I had an interest in computers. Certainly, I did. I had a year of computer science courses under my belt, but I also had an avid interest in computers as a hobby—I wrote my first BASIC program circa 1981 on a friend's Atari800. And, so my adviser proceeded to tell me that there is this nascent field called "bioinformatics," which is a hybrid of computer science and biology. I immediately fell in love with the idea that I could combine a professional interest of mine with a personal one. And, from then on, even through graduate school, all of my research projects involved programming. Not one required that I stand at a bench with a micropipette, as I knew I would be doing as a biochemist. Of course, it did not go over so well with many of the professors back then that a student would pursue a degree in either biochemistry or biology with a purely computational project. In the 1990s, there were just a handful of degree programs in bioinformatics in the whole world—one of them halfway around the world from where I lived. But I limited my own geographical options, and it seemed that my only choice was to pursue a graduate degree in "traditional" biochemistry and find an adviser and laboratory group that had an interest in performing computational analyses on their data.

Fortunately for aspiring scientists today, there are many straightforward ways to enter the field of bioinformatics. To that point, there are scores of degree programs throughout the world—many of them online degrees. And, there are other ways to further one's own career as a bioinformatics practitioner. For one, there is the Bioinformatics.Org website, of which I am the founder, with Prashanth Suravajhala among the directors. Prash also founded Bioclues.org and has been active in

mentoring students online regarding their academic projects in bioinformatics. It is because of this experience of his that I think you will be enlightened by the insight that Prash shares within these pages.

Bioinformatics.Org, Hudson, MA, USA J. W. Bizzaro

Acknowledgments

I thank Messers Springer, Bhavik Sawhney, Beracah John Martyn and Camilya Anitta for agreeing to my request to reconsider the revised edition of this book and for their consistent help in proofreading. Aninda Bose and Chandra Shekhar of Springer who have supported me all through the making of the first edition of the book have played a major role. Although the cartoons and illustrations were ideated by me, full credits to Partha Paul for bringing life to them.

My sincere obeisances to my revered Guru Maa Bijaya for her grace and blessings. My sincere gratitude goes to my mentees without whose thoughts this book would not have been here today. Likewise, I owe appreciation to my wife Renuka and my daughters Bhavya and Nirmala who always stood by me.

My peers in Bioclues.org and bioinformatics.org, ex-colleagues, and researchers in India, Denmark, the United States, and Japan, countless "e-colleagues," also contributed to my discussions. I sincerely thank Cox Murray, Jeff Bizzaro, Madhan Mohan, and Pawan Dhar who were generous enough to have responded to the questionnaire. My grandparents—Shri D. S. Sastry and D. S. R. Murthy are always remembered with fond love and affection. They have helped me in imparting clarity, coherence, and brevity to the text.

Finally, the book would not have come into good shape without the help of contributions from various authors across four continents, Springer reviewers, friends, and well-wishers, but not the least the author sincerely thanks the Springer typesetting team, Messers Nalini Gyaneshwar, Kamiya Khatter et al. for bringing the manuscripts in shape.

Prologue

Today, we define success by publicity and bank accounts. But that is not really success at all. Do not believe the hype. Success is ephemeral. You have to define it yourself.

Chris North

Most people would succeed in small things if they were not troubled with great ambitions.

Henry Wadsworth Longfellow

Any new word invites inquiry, excitement, and sometimes disdain and so was bioinformatics, at least in developing countries. Theoretical bioinformatics, although born in the 1980s, has flourished ever since, as many new academic and empirical developments with focal point on wet-lab research confirm. Bioinformatics is now regarded as a tool but fantasized as a familiar science even by few scientists who have had a track record of early career building. With research on bioinformatics mushrooming, both theoretical and wet-lab-based bioinformatics-aided works are often deemed very procedural and paraphernalia that these are not easily accessible to those who want to use the "tools for biology." Additionally, the career-driven paths using bioinformatics is tacit by the fact that one needs to attend to earn programming skills which is not always the case. This book aims to be an interface between those who aim for bioinformatics and apply research with a focus on Q and A on career growth. A great saying goes "If you want more, you have to require more from yourself." This also applies to bioinformatics. Happy reading!

Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Prashanth N. Suravajhala

Contents

About the Editor

Prashanth N. Suravajhala is currently working as a Senior Scientist at Birla Institute of Scientific Research (BISR), Jaipur. Previously, he has obtained a Ph.D. in Systems Biology. He has been involved in various projects specific to immunomodulatory/metabolic diseases and cancer biology. He has identified candidate genes using in silico approaches and provided a standard classification and scoring scheme for characterizing hypothetical proteins in vitro, specifically those that are targeted to mitochondria. More recently, his interests have expanded to transcriptomic profiling, functional proteomics, and molecular genetics of cancer/diseasome approaches linked to evolutionary aspects in understanding functional aspects of regulatory genes, epigenomic profiling, miRNAs, and long noncoding RNAs. He is a founder of Bioclues.org, a not-for-profit organization through which he mentors bioinformatics graduates

Chapter 1
Whither Bioinformatics?

Prashanth N. Suravajhala

Ever since the word "Theoretical Biology" was coined by Paulien Hogeweg in1978, bioinformatics, the current word has steadfastly come into existence with many biologists taking a leaf out of this discipline. Researchers by now know that bioinformatics is a mere tool, whereas its sister concern, computational biology, is deemed as a discipline. With bioinformatics burgeoning in the late 1990s, we relate the commencement of data deluge to the animistic knowledge that bioinformatics has brought in, lessening the scale of experimentation. Authentic bioinformatics, however, will not gain significant interest for researchers, at least until the wet laboratory biologists take a leap forward in acclimatizing the split half-term in bioinformatics. The figure of dogmas is pivotal in bringing the collaboration between biologists and cross-disciplinarians across biology as the event of dogmas in turn has introduced a plethora of new relationships between scientific studies and molecular biology. In effect, researchers have asked several questions on specialized mechanisms, if any that may be discovered in the advent of bioinformatical knowledge. This collaborative knowledge owes its impetus to the differentiation of independent eccentric science, namely, systems biology (SB). So, to ask whither bioinformatics into the enunciation and practice of the bioinformatical tools and scientific methods is a candid query.

Bioinformatics, since ages, has created a process of reasoning that was certainly not dependent on biology alone. Prior notions of intelligent algorithms clubbed with statisticians' skills, IT scientists' inclination, physicists' predictions, chemists' corner, and mathematicians' mind are a necessity to perform bioinformatics research. Not all disciplines can be made up by an individual alone but need unicentric efforts to meet the goals to derive bioinformatics knowledge. For example, the next

P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_1

generation sequencing (NGS) technologies have enabled non-sanger based sequencing technologies with unprecedented speed, thereby enabling novel biological applications. However, before bioinformatics and NGS stepped into the limelight, it must be noted that the NGS had overcome torpor in the field with the help of several cross-disciplinarians. It would never have been easy to stir up this understanding without the rapid involvement of the multifaceted scientists who have transformed biology as a whole. This obviously has the advantages of building up cross-disciplines, thereby deepening the knowledge curve between eccentric biology and information science, the latter constantly teaming up with the former to signify its discoveries with dogmas.

The greatest challenge facing the molecular biology community today is to make sense of the wealth of data that have been produced by genome sequencing projects. Conventional biology research was deemed always to be in the laboratory until the data deluge and explosion of genomic scale in the late 1990s. Thus, we are in an age of computing-to-research process. There are two different challenges one would pose: (1) sequence generation and (2) ensuring storage of the plethora of sequences generated in the laboratory with specific understanding and investigation using computers and artificial intelligence. That said, understanding the biology of an organism is a trivial issue as there are a number of focused research areas at different levels of "omics"-es, namely, genomics, proteomics, functomics, transcriptomics, need to being carried out at different levels. One of the foremost challenges today is to ensure that such data are efficiently stored, used through three forms of Es—extracting, envisaging, and elucidating this mass of data. A meaningful interpretation of such data must be done before one analyzes the complete volume for interpreting it or what we call "annotating" manually. In conclusion, discerning the function using computer tools must be the focus so as to have meaningful biological information explained.

The journey of transcriptomics starts with the discovery of ribonucleic acids in 1869 followed by their role in protein synthesis and as a catalyst in various biochemical reactions. However, the term transcriptomics first appeared in 1998 in the scientific literature (https://en.wikipedia.org/wiki/Transcriptomics_technologies) concurrent with different "omics" terminologies. Different "omes" and their respective descriptions are summarized in Table 1.1.

Why is bioinformatics interesting? All the central biological processes revolve around bioinformatics tools that need to be developed with a possible leeway in understanding the sequence–structure–function relationship (see Fig. 1.1). *DNA sequence determines the protein sequence which determines structure and function*. Why is it that we end up with protein as a determiner for every analysis? The simplest answer is that we would have less noise when we deal with protein sequences wherein, we deal with 20 odd amino acids to narrate results unlike the several compositions of four bases, namely, ATGC compendium of six reading frames translating into amino acids. This integration of information making up biological processes would allow us to understand the complete repertoire of the biology of organisms. However, the challenge faced by the biology community, especially on the inordinate data, is more from the umpteen genome sequencing projects. Traditionally, wet laboratory biologists carry experimental work even as

Table 1.1 Components defining different 'omics' technologies. The word 'ome' refers to 'many' or 'monies.' For example, genomes indicate the study of many genes

'Omes'	Description
Genome	The full complement of genetic information both coding and noncoding in an organism
Proteome	The complete set of proteins expressed by the genome in an organism
Transcriptome	The population of mRNA transcripts in the cell, weighted by their expression levels as transcripts copy number
Metabolome	The quantitative complement of all the small molecules present in a cell in a specific physiological state
Interactome	Product of interactions between all macromolecules in a cell
Phenome	Qualitative identification of the form and function derived from genes, but lacking a quantitative, integrative definition
Glycome	The population of carbohydrate molecules in the cell
Translatome	The population of mRNA transcripts in the cell, weighted by their expression levels as protein products
Regulome	Genome wide regulatory network of the cell
Operome	The characterization of proteins with unknown biological function
Synthetome	The population of the synthetic gene products
Hypothome	Interactome of hypothetical proteins

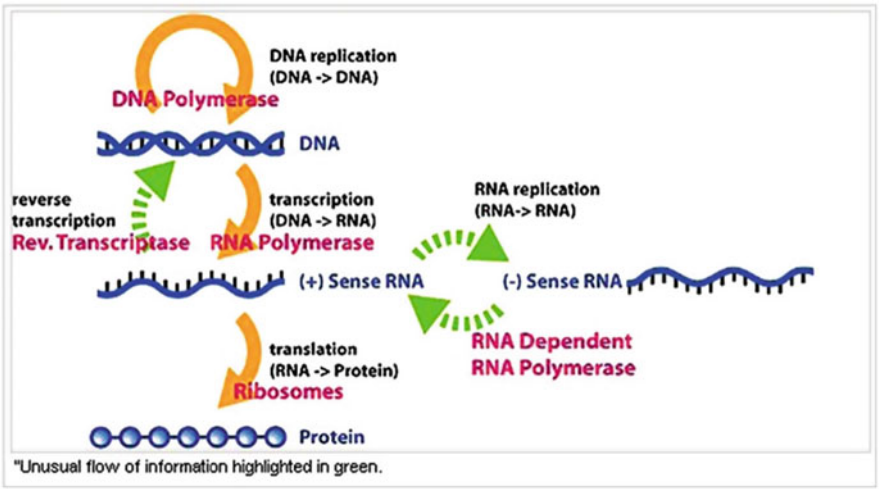


Fig. 1.1 An overview of the dogma of molecular biology with known specialized and unknown mechanisms/flows. (Image courtesy: Daniel Horspool)

the huge increase in the scale of data being produced from time to time could be better facilitated by in silico analysis. With the help of high-performance computing (HPC), sequences generated can be sporadic and further analyzed. Nevertheless, given the fact that molecular biology of a system is very complex, understanding and disseminating the information is to carried out at different levels using the "omes"

including the genome, proteome, transcriptome, and metabolome levels. There is a need for researchers, especially from the wet laboratory community, to herald bioinformatics indefatigably both in academia and industry.

What discoveries interest researchers? Looking at the dogmas, it is still not clear whether or not a protein can replicate another protein or a DNA can be obtained by direct reverse translation. It would be intriguing to understand if these could really happen in an organism. Can a genomic repertoire take shape in understanding the dogmas bottom-up? There is an aspiration but puny hope that bioinformatics can handle this. For example, protein–protein interactions (PPI) play a huge role in understanding the function of proteins. Various bioinformatics tools have been developed that allow researchers to compare proteins. Such comparative studies using algorithms such as BLAST (Altschul et al. 1990, 2005; Alstchul 1991) and other tools were carried out to distinguish unique proteins from paralogs, which later might have resulted from gene duplication events. The genomes sequenced so far were helpful in predicting not only evolutionary relationships but also identified function for the genes through functional genomics (Link et al. 1997a, b). The in silico methods such as homology search, presence of motifs, domains and signature sequences, orthology mapping, and radiation hybrid transcript mapping (Avner et al. 2001) are available for descriptive predictions of proteins with known (and sometimes unknown) function. However, these employed methods possibly might have a lot of false positives unless in vitro and/or in vivo experiments are followed to validate them. Moreover, these methods do not reveal a predicted function of hypothetical proteins (HP), thus making predictions more insignificant. Although all these methods are being employed by researchers, screening of HPs for novel translatable candidates is not often used and the researcher repeatedly performs the screening with laborious wet laboratory experiments. Furthermore, the proteins whose function remains unknown (i.e., those that remain hypothetical) and that are targeted to different organelles, especially mitochondria, could be important.

Many protein sequences contain motifs or short signature sequences called equivalogs (Haft et al. 2003), which are conserved in several organisms. These are a set of homologous proteins conserved since their last common ancestor with respect to function (Pearl et al. 2002). Some of these proteins might have a chance to be duplicated in organisms. It is therefore necessary to understand the genomic context of such proteins. An example of equivalog model is TIGR00658, identified as ornithine carbamoyltransferase. However, this enzyme is also known to act in an arginine biosynthesis pathway from ornithine (TIGR00032 and TIGR00838) in *Yersinia pestis* and arginine degradation (TIGR00746 and TIGR01078) in *Streptococcus pneumoniae*. The TIGRFAMs models, a TIGR family database, include equivalog models that have been used extensively in genome annotation. In addition, proteins with weak sequence similarity and no relevant structural homologies usually do not have known cellular function; such proteins are discarded from well-known proteins. When annotating proteins, a new molecular role for known cellular function is carefully addressed and curated. In many cases, there are numerous proteins that fall under domains whose functions are essentially known but they have no genuine role played in genomes. In addition, when annotating, if a

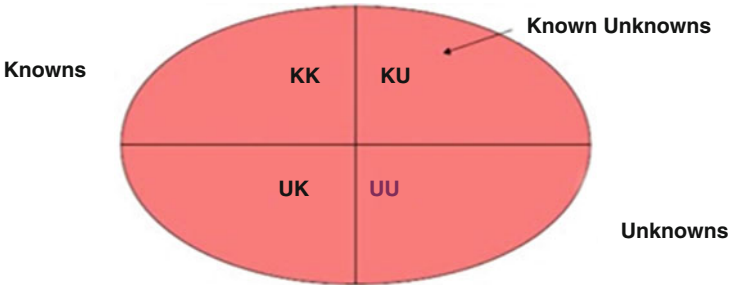


Fig. 1.2 The importance of Known Unknowns aliased "hypothetical genes" in the genome, illustrated in the form of a checkerboard. The Known is acronymed *"K"* while the Unknown "*U.*" Apparently, we seldom find "*UU*"s as it is a misnomer here. Unless the genome is sequenced, we find genes evaluating and devaluating

reference genome is considered besides comparing sequences in UniGene database (see web references) with selected protein reference sequences, the alignments would possibly suggest the function of a gene and finally, the possibility of annotating the protein as the hypothetical would be reduced. For example, many proteins in humans have been named as some repeat domains (for example, accession #CAB98209.1) maintaining homology to some known domains and all of them fall under a large category of domains. This does not necessarily mean that all these proteins makeup a function. There are also some instances of some proteins already similar to some organisms not showing up the function, although possibly studied from in silico and a few wet laboratory studies. Two such examples, one from eukaryotes and the other from bacteria, are discussed in what follows:

1. The Ankyrin repeat domain 16 (Ankrd16) has protein similarities in mice with 100% (Accession #NP_796242.1), humans with 85.7% (Accession #NP_061919.1), Xenopus with 71.1% (Accession #NP_001088685.1), and *Danio rerio* with 66.6% (Accession #NP_001017563.2). This annotation as revealed by GenBank and UniGene reference might have an update at a later point of time when new orthologs as identified from other metazoan sequences keep adding up to the annotation.
2. In bacteria, the proline proline and glutamic acid (PPE) and the proline glutamic acid (PE) gene families comprise many unique genes, some of them novel and labeled as hypothetical. Of them, many are known to be pseudogenes (Marri et al. 2006). The 10% of the coding DNA of *Mycobacterium tuberculosis* constitutes PE and PPE family genes and is involved in gene expression upon infection of macrophages, some of them as antigens mediating role in the pathogenesis or virulence. These were characterized while the expression levels and the functions of select PE/PPE family genes during various phases of infection (latent/mild/hypoxic) with *M. tuberculosis* (Kim et al. 2008) were studied (Fig. 1.2).

The aforementioned examples discussed are all resultant of the explosion of bioinformatics tools during the last three decades. Have bioinformatics technologies

Table 1.2 Pros and cons of different methods in annotating sequence

1. Sequence-based methods
Pros: Most known/reliable method
Cons: BLAST hits are electronically annotated and turn out to be false positives
2. Structure-based methods
Pros: Based on active site characterization/global fold similarity
Cons: Free energy minima always need to be set/obligation
3. Associated-based methods
Pros: Based on the domain/phylogenetic profiles
Cons: Lack of conserved proximity does not indicate a lack of functional association
4. Proteomics-based methods
Pros: Based on protein interaction domains. Gaps or holes in the known pathway can be assigned. Function awaits a protein to be characterized
Cons: Lots of false positives

revolutionized genomics and proteomics? Well, there has been a focus on molecular medicine which paved the way for establishing intervention and treatment of well-known diseases to proactive prediction and prevention of disease risk. These approaches should really require new informatics systems that will link large-scale databanks and special programs for data mining and retrieval in bioinformatics and chemoinformatics. All the wet laboratories should be able to provide a platform for powerful new molecular diagnostic tools along with multianalyte assays for expression of genes and proteins in different patterns of diseases. With researchers scaling the ladder of bioinformatics progress by leaps and bounds, there is a need for an enhanced understanding of the interactions in a system (organism). What are the components that interact with each other? What is the outcome of such interactions? Do interactions alone provide us the functional decipherment? Should we just be sufficed with the progress made on say, cures for diseases by the year 2050? Should we reach a consensus on the combination of tools, namely, rapid and inexpensive DNA sequencing technologies, HapMap project, dollar one genome (DOG), and so on? We hope that this will let us understand precisely how bioinformatics transits from research to vocation and avocation (Table 1.2).

1.1 Bioinformatics "Aging" in Systems Biology

Systems biology has gained a lot of attention over the years. Of late, biologists have been actively engaged in this discipline in different forms when molecular biology is merged with multi-context disciplines. During this process, SB ran into several definitions. To answer what is a system: We could think of multiple organelles existing in our human body as we use components to describe entities in a system.

As we integrate various vehicular components to construct a vehicle, we describe components such as organelles to makeup a living system. The biology of the system

Table 1.3 Timeline eventing important spheres in bioinformatics

• 1859 Charles Darwin's "origin of species"
• 1944 Avery, MacLeod, McCarty: DNA is the genetic material
• 1953 Structure of DNA
• 1955 Complete sequencing of insulin
• 1988 National Center for Biotechnology Information (NCBI) founded
• 1988 Sanger Centre, Hinxton, UK
• 1994 EMBL European Bioinformatics Institute, Hinxton, UK
• 1995 First bacterial genomes completely sequenced
• 1996 Yeast genome completely sequenced
• 1999 Fly genome completely sequenced
• 2000 bioinformatics.org and opensource
• 2001 Human genome and bioinformatics ****Systems Biology****
• 2002–2004 Umpteen genomes sequenced
• 2005 EVOLUTION in terms of bioinformatics as a breakthrough
• 2007 Personal genomics*person "omics"
• 2008 The hypothetical proteins and orphan genes???
• 2011 Predictive biology approaches
• 2012 Next Generation Sequencing burgeons
• 2014 Oxford Nanopore changed the pace of NGS with its first product
• 2016–2019 ScRNA-Seq and spatial genomics era

is called systems biology. Every system has an effect on its environment and so does the components in a system, even as the components entitled to SB include genes, proteins, metabolites, and enzymes as minor entities, while cells, tissues, organelles, and organs as major components. Hence, interactions among the components would be interesting to value SB. While a system could have many organelles and the components that makeup the flow of a system, they are bound to interact with one another. For example, enzymes, proteins, metabolites, genes, DNA, and functional protein domains are known to interact with each other. Integrating all the interactions of components indicates: Which survives (and competes) the best while the ultimate goal of SB is to exploit the interplay among the components. From a reductionism's point of view, researchers define SB based on whether the components in a system are interacting with each other, mutations arising and falling, proteins evaluating and devaluating, strains adapting and unfitting in the environment, and some genes if lost and found (Table 1.3).

1.2 Defining Systems Biology Through Omics: The Two Paradigms

Is systems biology (SB) all about the genes making up the proteins and how the components processing in a system interact with each other? The fields of omics in the recent past have believably revolutionized biomedicine and by far means there needs to be a focus on change in defining these upcoming omics-es. Huang S's classification of SB has yielded the loose and the apparent but broadened definitions from the dynamics and reductions approach (Huang 2004). The dynamicity of SB is based on a pure level where the system is based on models and networks: Be it quantitative or qualitative, whereas the reductionism defines SB based on the high-throughput methods involving different molecular biology techniques. Overall, the loose definition applies to projects exploring individual biological networks, while the broadened but still "derivative" definition is the outgrowth of theoretical models along with systems theory across interdisciplinary sciences such as engineering, mathematics, statistics, artificial intelligence, and so forth. However, many authors (Tracy 2008; Cornish-Bowden et al. 2007; Huang and Wikswo 2006; Strömbäck et al. 2006; Bruggeman and Westerhoff 2007) have deliberated that the concept of the gene resulting in omics has begun to outlive its usefulness while they felt that the SB could be projected into several dimensions keeping in view the multifaceted systems' complexity of living organisms (Ideker and Hood 2019). With SB maturing, researchers have started proposing an alternative means to define gene based on a richer explanation: Genetic functor, or genitor, a sweeping extension of the classical genotype/phenotype paradigm that describes the "functional" gene (Fox Keller and Harel 2007). Thus, we could understand the dynamic behaviors of molecular associations implicitly known from various methods and technologies integrating one or more of the SB data:

Overall, SB can be envisaged keeping in view the following points:

1. Systems biology is conceptualized in terms of PPI. The interplay between components in systems is exploited between protein–protein, domain–domain, DNA–DNA as a whole, or even a protein–DNA.
2. The interactions among the components are better explained in such a way that what is in theory need not fit practically implicating that a hypothesis-driven approach need not always be experimental (biological) driven.
3. With some answers to questions like if there are interactions known, we can take a measure of unknown interactions in a system, SB approaches toward understanding bona fide *PPI.*

Does SB back biologists? There are specific traits that makeup PPI networks: Everything in biology is better explained through interactions while the interactions are a priority in accordance with the organization, cooperability, and mapping the components in a system. The SB signifies if components interact with each other. This led to the birth of several disciplines such as systems molecular medicine, immunological SB, local and global metabolic profiling, systems diagnostic therapy,

and systems drug development, all budding across nascent biology disciplines. Although the PPI are outcomes of almost all cellular processes, there is diversity in protein interactions, that is, all proteins share common properties at a certainty. For example, the distortion of protein interfaces leads to the development of many diseases, and to understand its mechanism, we lead PPI experiments. When proteins recognize specific targets and bind them, it results in conservation that depends on structural and physicochemical properties. The nature and applications of SB with respect to PPI were well-reviewed elsewhere (Huang 2004; Tracy 2008; Cornish-Bowden et al. 2007; Huang and Wikswo 2006; Strömbäck et al. 2006).

1.3 Is Biology Explained Through Protein–Protein Interaction Networks Alone?

Apart from the three most common omics-es, namely, "Gen-omics," "Prote-omics," and "Transcript-omics," bioinformatics and biology researchers have been taking up omes and omics-es very rapidly as is evident from the use of the terms in PubMed (Dell et al. 1996). As a result, a variety of omics disciplines such as phenomics (Schork 1997), physiomics (Chotani et al. 2000; Gomase and Tagore 2008), metabolomics (Kuiper et al. 2001; Fiehn 2002), lipidomics (Han and Gross 2003), glycomics (Gronow and Brade 2001), interactomics (Govorun and Archakov 2002), cellomics (Taylor et al. 2001) have begun to emerge, each with their own set of instruments, techniques, reagents, and software. These have driven new areas of research consisting of DNA and protein microarrays, mass spectrometry, and a number of other instruments that enable high-throughput analyses.

While genomics forms a main hierarchy of classification, there are many other omics-es that fall under a clad of primary (gen) omics' enabled SB, for example, functional genomics, comparative genomics, computational genomics, and phylogenomics. With more than 1800 microbial genomes sequenced or being sequenced today and the number still increasing, another set of omics called metagenomics aims to access the genomic potential of an environmental sample. It would answer some of the questions we posed in the earlier sections. This environmental "omics" bridges the integration of metagenomics with complementary approaches in microbial ecology (Schloss and Handelsman 2003).

While the mapping of PPI is a key to understand biological processes through interactomics, many technologies have been reported to map interactions, widely applied in yeast. At present, the number of reported yeast protein interactions truly validated by at least one other approach is low with the amount of throughput it takes to process (Cornell et al. 2004). This is because of the false discovery rate of proteins interacting with their partners. With the advent of virtual interactions, the growth of false positives also increased, thereby allowing the researchers to keep a track of finding these false positives through statistical inference. Any dataset of interaction map is complex while tools to decipher true positives are being developed in the

form of markup languages such as system biology markup language (SBML) (Hucka et al. 2004). The mapping of human–protein interaction networks is even more complicated, suggesting that it is unreasonable to try mapping the human interactome; instead, interaction mapping in human cell lines should be focused along the lines of diseases or changes that can be associated with specific cells (Figeys 2004). This "omics revolution" would force us to re-evaluate our ability to acquire, measure, and handle large datasets. The omic platforms such as expression arrays, MS, and other high-throughput methods have enabled quantification of proteins and metabolites derived from complex tissues. Applying SB, the integrated analysis of genetic, genomic, protein, metabolite, cellular, and pathway events are in flux and interdependent. With the onset of various datasets, it necessitated the use of a variety of analytic platforms as well as biostatistics, bioinformatics, data integration, computational biology, modeling, and knowledge assembly protocols. Such sophisticated analyses would definitely provide new insight into the understanding of disease processes through phenome–genome networks and interactomics studies (Lage et al. 2007). In this regard, SB clubbed with interactomics, more appropriately considered as a process containing a series of modules, aims to provide tools and capabilities to carry out a wide range of tasks (Morel et al. 2004). Even as protein analysis is known as a field of research with a long history, several developments of a series of proteomics approaches including MS opened the door for a synergistic combination with genomic sequence analysis, focusing on aspects of genome-wide transcription control, regulomics. In analogy with all the other omics-es, a combination of MS-based proteomics with in silico regulomics analyses can produce synergistic effects in the quest to understand how cells function (Werner 2004). Carrying this further, it has been suggested that the term "translatome" could be used to describe the members of the proteome weighted by their abundance, and the "functome" to describe all the functions carried out by them (Greenbaum et al. 2001). However, there are still many difficulties resulting from the disorderliness and complexity of the information. To overcome this, removing noisy data and finding false positives could be enhanced using various tools to some degree. However, these can also be overcome by averaging broad proteomic categories such as those implicit in functional and structural classifications (Fig. 1.3).

1.4 Systems Biology in Wet Laboratory

Fundamental biological processes can now be studied by applying the full range of omics technologies (genomics, transcriptomics, proteomics, metabolomics, etc.) using the same biological sample and high-throughput methods such as MS (McGuire et al. 2008; Kim et al. 2008). A wide array of assays including high-throughput methods such as tandem mass spectrometry (MS/MS), yeast two hybrids (Y2H), and pull-down assays are preferentially used to navigate them. Clearly, it would be desirable if the concept of the sample were shared among technologies such as MS for that, until the time a biological sample is prepared for use in a specific

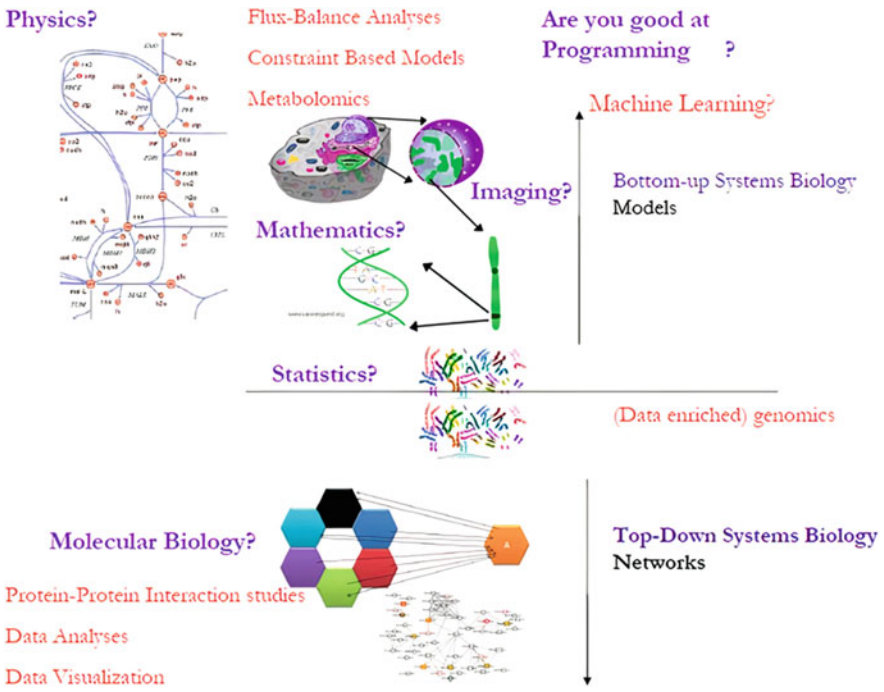


Fig. 1.3 Quantitative picture of various omics and the various fields, an enthusiast can take up

omics assay, its description is inherently technology independent. However, the compulsion for accurate analyses of all these high-throughput methods is to remove redundant and false-positive data. Redundancy of data has been the biggest threat for causing errors in data usage. Sharing a common informatics' representation would encourage data sharing, leading to a decrease in redundant data, and the potential for error. The recent introduction of *WikiProteins* has been a worthy effort that brought all annotators to come together on a common platform (Mons et al. 2008). This would result in a significant degree of harmonization across different omics data standardization activities, a task that is critical if we are to integrate data from these different data sources (Morrison et al. 2006). The bioinformatics applied to omics are varied and particularly noteworthy or characteristic of proteomics research, for example, 2DE analysis or MS. Another important task of bioinformatics is the prediction of functional properties through ontology-based functional networks from a vast number of databases.

Apart from the above-discussed issues, genome technologies are being carried out in every major model system. For example, new technologies are being developed to rapidly identify mutations or small molecules that increase the life span for aging-related research. While the DOG recently has been known to play a role as a model system for cancer, because of its similarities to human anatomy and physiology, it may prove invaluable in research and development on cancer drugs (Khanna

2006). Inversely, as dogs too naturally develop cancer they may share many characteristics with human malignancies. This probably would accelerate genome-wide, cross-comparison of organisms for finding the function of more genes ultimately using drug discovery development.

1.4.1 Metabolomics

Metabolomics has come into sight as one of the newest "omics" science with a dynamic portrait of the metabolic status of living systems. The analysis of the metabolome is particularly challenging as it has it's roots in early metabolite profiling studies but is now a rapidly expanding area of scientific research in its own right. It is a science employed toward the understanding of global SB (Rochfort 2005). The metabolomic tools aim to fill the gap between genotype and phenotype permitting simultaneous monitoring molecules in a living system. The smartness of using metabolic information could be applied in translating into diagnostic tests as they might have the potential to impact on clinical practice and might lead to the supplementation of traditional biomarkers of cellular integrity, cell and tissue homeostasis, and morphological alterations that result from cell damage or death (Claudino et al. 2007). Metabolomics has been widely applied to optimize microorganisms for white biotechnology even as it spreads to the investigation of biotransformation and cell culture. Together with the other more established omics technologies, metabolomics aims to contribute to different spheres ranging from an understanding of the in vivo function of gene products to the simulation of the whole cell in the SB approach. This will allow the construction of designer organisms and yet another science synthetic biology evolves (Oldiges et al. 2007). Although metabolomics measures the multiparametric response of living systems to genetic modification, there is a consistent debate of synonymy with metabolomics. Admittedly, there is a concurrence of the former being associated with NMR while the latter being associated with mass spectroscopy. This part of the microbial transformation has led several standards for these two meta-omics' delivering SB tools (Fiehn et al. 2006).

1.5 Mitochondriomics

Mitochondria are semiautonomous organelles, presumed to be the evolutionary product of a symbiosis between a eukaryote and a prokaryote. The organelle is present in almost all eukaryotic cells to an extent from 10^3 to 10^4 copies. The main function of mitochondria is the production of ATP by oxidative phosphorylation and its involvement in apoptosis. The organelles contain almost exclusively maternally inherited mtDNA, and they have specific systems for transcription, translation, and replication of mtDNA. Mitochondrial dysfunction has been

correlated with mitochondrial diseases where the clinical pathologies are believed to include infertility, diabetes, blindness, deafness, stroke, migraine, heart, kidney, and liver diseases (Reichert and Neupert 2004).

Recently, cancer was added to this list when investigations into human cancer cells from breast, bladder, neck, and lung revealed a high occurrence of mutations in mtDNA. With the understanding of the role of mitochondria in a vast array of pathologies, research on mitochondria and mitochondrial dysfunction has in the last decade yielded a huge amount of data in the form of publications and databases. Yet, the field of mitochondrial research is still far from exhaustion with many essentials waiting to be discovered. The recent identification of a number of proteins targeting mitochondria has enabled immense interest to understand the function of some genes unnoticed in the mitochondrion (Calvo et al. 2006). With only 13 proteins sitting inside mitochondria through oxidative phosphorylation, and more than 1500 estimated proteins targeting this tiny organelle, identifying complete protein repertoire in this machinery could decipher the biology behind mitochondria or what makes us breathe. A complete set of mitochondrial proteomes syntenic with other eukaryotes has just started and there is a promise in understanding how the organelle proteomes and interactomes could essentially be used to develop into SB (Calvo et al. 2006).

1.6 "Omic" Challenges in Systems Biology

Bioinformatics has enabled all-against-all comparison distinguishing unique proteins from proteins that are paralogs resulting from gene duplication events. The last two decades have seen an avalanche in databases while algorithms such as BLAST allowed such comparisons. In the post-genomic era, the genomes sequenced so far would essentially cover the future of omics in them as they enable predicting not only evolutionary relationships but also make use of different approaches used in identifying the function of genes. This functional genomics is the cause of understanding how proteins interact with each other and network in the living organism. The gene or protein function could be ascertained based on physiological characterization or if the two proteins are known to be physically interacting with each other or virtually interacting with each other. The SB approaches in present-day bioinformatics have brought in a special emphasis on association-based networks in the form of virtual interactions, thereby making up the possibility of phenome–genome networks grow bigger (Lage et al. 2007). Ultimately, it makes sense when such interactions bring out a function and find a candidate for disease. The increase in GenBank accessions resulted not only in the number of genes identified but also in the number of citations these accessions refer to. While various databases and terms have been defined, several omics-es are reported from time to time at http://www.omics.org (Fig. 1.4).

The last 10 years have not only seen the rise of bioinformatics producing an unprecedented amount of genome-scale data from many organisms but also the wet

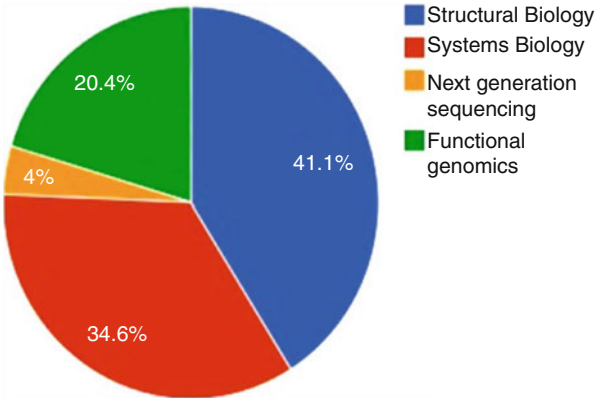


Fig. 1.4 Proportion of research articles published in different omics-es in PubMed as on September 1, 2012

laboratory research community has been successful in exploring these data on using bioinformatics many challenges still persist. One of them is the effective integration of datasets directly into approaches based on mathematical modeling of biological systems. This is where SB has bud resulting in top–down and bottom–up approaches. The advent of functional genomics has enabled the molecular biosciences to come a long way toward characterizing the molecular constituents of life. Yet, the challenge for biology overall is to understand how organisms' function. By discovering how function arises in dynamic interactions, SB is everywhere addressing the missing links between molecules and physiology. Top–down SB identifies molecular interaction networks on the basis of correlated molecular behavior observed in genome-wide "omics" studies. On the other hand, bottom–up SB examines the mechanisms through which functional properties arise in the interactions of known components. Applications in cancer are a good example to counteract these two major types of complementary strategies (Stransky et al. 2007). Several web-based repositories have been established to store protein and peptide identifications derived from MS data, and a similar number of peptide identification software pipelines and workflows have emerged to deliver identifications to these repositories. Integrated data analysis is introduced as the intermediate level of an SB approach and as a supplementary to bioinformatics to analyze different "omics" datasets, that is, genome-wide measurements of transcripts, protein levels or PPI, and metabolite levels aiming at generating a coherent understanding of biological function (Steinfath et al. 2007). Furthermore, existing and potential problems/solutions such as de facto experimental and the following bioinformatics challenges might hold prospective in the near future:

1. Challenges in high-dimensional biology (HDB): Recently, the term HDB has been proposed for investigations involving high-throughput data (Mehta et al. 2006). The HDB includes whole-genome sequences, expression levels of genes, protein abundance measurements, and other permutations. The identification of biomarkers, the effects of mutations and drug treatments, and the investigation of

diseases as multifactor phenomena can now be accomplished on an unprecedented scale.

2. Finding the function of HP: Another feature of PPI map is to find the function of unknown proteins. PPI has become a very common step in the annotation of a protein. Various tools such as iHOP (http://www.ihop-net.org), STRING (http://string.embl.de), GeneMania (http://www.genemania.org), and so on, aid the researchers to find if there are interacting partners of protein of interest. The data could be visualized through tools such as Cytoscape (www.cytoscape.org), VisANt (http://www.visant.bu.edu), and Osprey (http://biodata.mshri.on.ca/osprey/servlet/Index), and so on, for further analyses. The nearest partners would essentially mean that the hypothetical or uncharacterized protein could play a function similar to its interactor(s). In the context of PPI networks, we could consider if a model is to be developed from the network or a network is to be generated with an already established model. Precisely, the putative function of a protein could be better known from a PPI network to develop a model from it. Information on "known" or "unknown" PPI is still mostly limited but integrating tools such as these could generalize a way to find bona fide function.

1.7 Are Interactions Based on the Nature of Binding?

Does close homology between two proteins confer that they do interact in the same manner? Yes, they do and confer evolutionary constraints in lieu of structural divergence while remotely related proteins have a different interaction mode (Drummond et al. 2005). Also, conservation of protein interface indicates the average conservation of the rest of the protein. While all these forms an integral part of SB, apart from the novel interactions that arise based on the type of homology, there are interactions based on the binding entity, namely, stable and transient. The former interactions are consistent and bookmarked while the latter is temporary. There are interacting proteins that might co-express indicating that the expressed proteins, which evolve slowly are normalized wherein the normalized difference between the absolute expression data is calculated based on several tools such as microarrays (Drummond et al. 2005). However, there are other techniques such as density gradient and virtual pull-down assay methods cited as above beginning to be understood and substantiate above views.

As thousands of new genes are identified in genomics efforts, the rush is on to learn something about the functional roles of the proteins encoded by those genes. Clues to protein functions, activation states, and PPI have been revealed in focused studies of protein localization. A meta-analysis of data derived from genome-wide studies of aging in simple eukaryotes will allow the identification of conserved determinants of longevity that can be tested in other mammals (Khanna 2006; Kaeberlein 2004). Adding to the various high-throughput methods, technical breakthroughs such as GFP protein tagging and recombinase clones, large-scale screens of

protein localization are now being undertaken to understand the function of the proteins (O'Rourke et al. 2005).

1.8 Fundamental and Best Practiced Tools for Annotating Proteins and Genes

In the recent past, various bioinformatics tools have been developed that allow researchers to compare genomic and proteomic repertoire. Comparative studies using algorithms such as Blast and databases are carried out to distinguish unique proteins from paralogs, which later might have resulted from gene duplication events. The genomes sequenced so far were helpful in predicting not only evolutionary relationships but also identified function for the genes through functional genomics. Although many methods are being employed by researchers, screening of proteins for novel translatable candidates is not often used and the researcher repeatedly performs the screening with laborious wet laboratory experiments. To increase the sensitivity, further clues on tissues and development stages from the queried gene's sequences could be surveyed using tools such as gene expression omnibus (GEO) or UniGene-EST or cDNA profile database. Furthermore, protein link to the genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping, or cytogenetic mapping as available from GenBank resources would improve the understanding of protein annotation. Besides this, whether or not a protein contains a polyadenylation signal could be an added knowledge to meet the criteria of well-annotated proteins. This is because tools such as MEME reveal many 30 UTRs forming conserved motifs, which indicates these regions appear more conserved than expected. This means, higher the conservation, greater the duplications and greater is the chance of being not annotated or "hypothetical." There seem to be many unique genes that are overrepresented in the form of duplications; a simple search in GenBank gene list would reveal that there are several accessions duplicated. For example, in the case of the gene FusA2a, *bona fide* accession is mapped to CAD92986, and yet, a few of the isoforms/unique genes remain unknown (e.g., CAD93127). In summary, there could be many proteins less annotated, and yet many tools are known to describe the function. This leaves to beg a question, what would be the fate of proteins that cannot be annotated through some tools, or in contrast how many best tools are used to describe or annotate a protein?

Apart from BLAST and FASTA, the sequence-based feature annotation is applied by RefSeq using several tools, namely, BEAUTY X-Blast Enhanced Alignment Utility, and PROSITE. While many other variants of BLAST including PSI Blast and PHI Blast, sequence alignments using ClustalW, ClustalX, and Cobalt are used, not all the tools are used in tandem to eliminate false positives. Whether the protein is soluble or insoluble is known through TopPred; the topology of protein with the orientation and location of transmembrane helices attribute to the function. Additionally, orthology mapping using tools such as HomoMINT are used, which

increases the chance of the protein annotation. With the central dogma beyond the age today in bioinformatics, namely, sequence specifies structure and function; annotations have become mightier to further manually curate allowing researchers to perform experimental analyses for some proteins. The structures of proteins not only provide functions but the shapes exhibited by the proteins allow them to interact selectively with other proteins or molecules. This specificity is the key for the proteins to interact with another protein, thereby inferring the function. However, most of the bioinformatics analyses are misleading unless biochemical characterization is carried out. Furthermore, the protein annotation has gained much importance with the introduction of many metazoan genome sequencing projects in addition to the 1000 genomes project that is in progress. With 40–50% of identified genes corresponding to proteins of unknown function, the functional structural annotation screening technology using NMR (FAST-NMR) has been developed to assign a biological function which is based on the principle that a biological function can be described based on the basic dogma of biochemistry that the proteins with similar functions will have similar active sites and exhibit similar ligand-binding interactions, although there is a global difference in sequence and structure. Tools such as combinatorial extension which confer structure similarity, DALI for NMR, finally determining function, PvSOAR, and Profunc—given a 3D structure, aims at identifying a protein's function has been widely used. However, there are many other methods such as the Rosetta Stone method, phylogenetic profiling method, and conserved gene neighbors that have been widely employed and being accepted by the scientific community.

Biological function of proteins would help in the identification of novel drug targets and helps reduce the extensive cost of practical examinations on several candidates. With the enormous amount of sequence and structure information availability, innumerable automated annotation tools for proteins have also been generated. One such example is the automated protein annotation tool (APAT), which uses a markup language concept to provide wrappers for several kinds of protein annotations. While FFPred is available to predict molecular function for orphan and unannotated protein sequences, the method has been optimized for performance using a protein feature-based method through support vector machines (SVMs) that does not require prior identification of protein sequence homologs. It works on the premise of posttranslational modifications, Gene Ontology, and localization features of proteins. Yet another tool, namely, VICMpred, aids in broad functional classification of proteins of bacteria into virulence factors, information molecule, cellular process, and metabolism molecule. The VICMpred server uses an SVM-based method having patterns, amino acid, and dipeptide composition of bacterial protein sequences. ConSeq and ConSurf have been widely applied in predicting functional/structural sites in a protein using conservation and hypervariation.

The final part of annotation can be studied through interactions and associations. All interactions are associations, while not all associations are interactions. The association tools, namely, search tool for the retrieval of interacting genes/proteins (STRING), GeneCards, IntAct, MINT, biomolecular interaction network database

(BIND), which have been enhanced as biomolecular object network database (BOND). With BIND inside, BOND is a comprehensive database that helps in the annotation of proteins through unique object-based interaction studies. Although there are other variants of some of these databases such as GeneAnnot and GeneDecks of GeneCards, most of them are used for finding genes based on different queries.

Methods for predicting protein antigenic determinants from amino acid sequences were a crucial point for segment-level annotation of proteins. Since then, so many computational methods have been developed based on such basic and fundamental methods and pinpoint the importance of basic methods in the area of computational biology. Developed methods are being assorted in applications from sequence-based antigenic determinants to surface-based consensus scoring matrix approach for antigenic epitopes. Such developments significantly contributed to the refinement of existing and development of new and versatile techniques; but have roots in indispensable and conventional approaches.

Incorporating a systematic representation of fundamental and best practiced tools/servers to facilitate users for information would be useful even as additional features with their respective inputs, outputs, mode of action, and level of annotation have also been compiled (see Table 1.4), and will help experienced as well as beginners in the area of protein annotation.

1.9 Can Bioinformatics Influence Animal Experimentation?

Decades ago, legislation on the use of animals was enacted in many countries involving three R's: Reduction, refinement, and replacement of animal models. Ever since this was enacted, there was a sudden buzz about laboratory animals and their use to be reduced, refined, and replaced wherever possible, for ethical and scientific reasons. The three R's concept was put forward by W.M.S. Russell and R.L. Burch in 1959 in The Principles of Humane Experimental Technique. A great detail on the three R's was reviewed by many in the interest of good and humane science. The word "alternatives" came into use after the publication of the book "Alternatives to animal experiments" by David Smyth in 1978.

With the arrival of bioinformatics and SB, the impact on animal experiments was slowly felt. The generation of high-throughput data in the form of genomics, transcriptomics, and metabolomics, biology has essentially transformed into a computational problem. Due to this reason, we believe that the role of computation in biology leading to reducing, refining, and replacing animal experiments needs to be reviewed and discussed. Let us review this question using two approaches.

1. Reductionist approach: Today the fields of omics have revolutionized fundamental biology and biomedicine. Greater attention needs to be paid to defining upcoming omics-es based on the three Rs. We believe that the first two R's—

Table 1.4 Best practiced tools for a myriad of protein functional annotations

Tools/ Servers	Interaction/ Association	Output	Comments on methods	Annotation
Blast/ FASTA	Protein sequence database	Close and distant candidates	Heuristics	Homology
Pfam/GO	Protein sequences	Ontology based	Pattern based	Ontology
VICMPred	Annotated protein sequences	Functional information	Machine learning based	Functional annotation
Interpro/ Prodom	Protein sequence motifs and domains	Protein family and domains	Domain based	Structural and functional annotation
MEME	Protein sequence motifs	Protein motifs	Statistical	Motifs identification and analysis
TopPred		Protein solubility conditions	Machine learning based	Solubility/ insolubility of proteins
Profunc	3D structure	Various functions	Machine learning based	Functional annotation
STRING/ IntAct	PPI and database	Interactors	Pattern and mining based	Protein–protein interactions
TargetP/ PTarget	Annotated proteins	Inter- and intracellular signals	Quantitative analytical	Signal sorting
APAT	Proteins	Myriad features	Markup language based	Miscellaneous
FFPred/ RIGOR	Information based	Structural and functional element information	Structural elements based annotation	Structural and functional annotation
ConSeq/ Con surf	Protein sequence and 3D structure	MSA, phylogenetic tree, various statistical scores, conserved residues on sequence, and structure of proteins	Sequence and structural evolutionary conservation and hypervariation	Structural, functional, and evolutionary annotation
MINT/ BIND/ BOND/ GeneCards	Protein–protein interactions	Interactors and annotated pathways	Miscellaneous	Protein–protein interactions and pathways

reduction and refinement—aptly fit into the category of definition where we may not completely replace animal experimentation but at least lessen the scale and usage of laboratory animals. Thanks to high-throughput techniques through which we are able to better explore in vitro methods. However, the reductionist approach does not completely reduce or refine this process as this implies for smart experimentalists with a humane touch. Homogeneity and environmental

conditions play a major role in reducing the experimentation process. Greater the use of genetic homogeneity, greater is the chance to reduce the use of animal models. Similarly, greater the chance of maintaining and ensuring the conditions of the experiment, greater is the chance to reduce animal experimentation.

2. Dynamic or a vibrant approach: This applies to in silico models. Many computational models in biology are used nowadays. Dynamicists might not even go to that extent but think of plan B: considering a scale of sentience. A common question often asked is why not use animals that are small at the scale of taxonomy? But as computational biologists, we would not lose hope in saying that we are at our magnanimous best and not very far in bringing intelligent and sophisticated bioinformatics tools and use dynamic approaches wherein in silico models are widely exploited. Here, acceptance and use of computer-based and in vitro methods in fundamental research in testing chemicals, medicines, applying biostatistics through experimental design are inevitable thus raising questions—can animal models be replaced?

To address this key question, opinions were raised through bioinformatics.org online polls and an extensive discussion was organized through the SAB forum (http://www.scienceboard.org). Bioinformatics clubbed with SB have been practically two-fold as the practitioners understand how molecules work in silico; how chemistry works between biology and information technology, and importantly, see how genes or proteins could be predicted heuristically or non-heuristically algorithmic, thereby we could approach the wet laboratory beforehand in a more organized mode. Bioinformatics, by and large, has become an enforced tool in today's full-bodied molecular biology. So, the popularity is not for professionals from bioinformatics only. In neutrality, not every person from bioinformatics will have these types of statistics, but let us judge ourselves closer to getting hold of computational biology or bioinformatics by following the three Rs before experimenting in the laboratory!

We leave our thoughts with the following quotes by Dr. Peter Mansfield (GP, and Founder-President of 'Doctors in Britain against Animal Experiments'.) in "Animal Experiments in Medicine: The Case Against," May 1990:

> There is no comprehensive animal model for humankind. . .The truth is, and always has been, that the first clinical use of new medication in human patients provides the first reliable clues as to what can be expected of it. Premarketing research on animals is a lottery; post marketing surveillance comes too late for the first human victims of side- effects.

1.10 Addendum: Results of Poll @Bioinformatics.org

Only 22% of the people voted for yes when asked could computer models someday replace humans in clinical trials while 50% voted negative and 28% have no hopes at all.

1.10.1 Opinion of Few Scientists on Bioinformatics Influencing Animal Experimentation

Lorikelman opined that an Artificial Intelligence—Turing test could be an option that predicts human behavior. "Problem will be the unexpected interactions between pathways or organ systems that we might expect to see in a fairly large number of people that therefore could be observed in a clinical trial and so I don't think models will be replacing trials soon. Second big problem will be the occasional catastrophic individual reaction that some people have to a drug—difficult to model" was what Lorikelman had to say. Furthermore, he feels with no decent models available, the information about human metabolism and human immune reaction cannot be understood.

R. Wintle added saying that it could be a problem with regulatory agencies buying computer models as they may not seem to work. Also, uncontrollable environmental effects on drug efficacy, and potentially also stochastic effects are further hindrances to model the animals.

Jooly opines that improved computer models may be extremely helpful in terms of "reduction," but feels she cannot imagine that they will ever be good enough for complete "replacement."

R. Stevens says, "I've seen too many people say that we can someday understand "gene products" by just looking at the DNA sequence to fall for this idea. As soon as you think you know all the variables needed to understand something in silico, someone will discover that the variables were all for one gender, race, age or whatever, and the whole thing will be wrong. Not those clinical trials are perfect. Even if you do everything you can to test a new drug/treatment, there could always be something out there that wasn't predicted by the trials."

References

Altschul, S.F.: Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. **219**(3), 555–565 (1991). https://doi.org/10.1016/0022-2836(91)90193-a

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990). https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S.F., Wootton, J.C., Gertz, E.M., Agarwala, R., Morgulis, A., Schäffer, A.A., Yu, Y.K.: Protein database searches using compositionally adjusted substitution matrices. FEBS J. **272** (20), 5101–5109 (2005). https://doi.org/10.1111/j.1742-4658.2005.04945.x

Avner, P., Bruls, T., Poras, I., et al.: A radiation hybrid transcript map of the mouse genome. Nat. Genet. **29**, 194–200 (2001). https://doi.org/10.1038/ng1001-194

Bruggeman, F.J., Westerhoff, H.V.: The nature of systems biology. Trends Microbiol. **15**(1), 45–50 (2007)

Calvo, S., Jain, M., Xie, X., Sheth, S.A., et al.: Systematic identification of human mitochondrial disease genes through integrative genomics. Nat. Genet. **38**, 576–582 (2006)

Chotani, G., et al.: The commercial production of chemicals using pathway engineering. Biochim. Biophys. Acta. **1543**(2), 434–455 (2000)

Claudino, W.M., Quattrone, A., Biganzoli, L., Pestrin, M., et al.: Metabolomics: available results, current research projects in breast cancer, and future applications. J. Clin. Oncol. **25**(19), 2840–2846 (2007)

Cornell, M., Paton, N.W., Oliver, S.G.: A critical and integrated view of the yeast interactome. Comp. Funct. Genomics. **5**(5), 382–402 (2004)

Cornish-Bowden, A., Cárdenas, M.L., Letelier, J.C., Soto-Andrade, J.: Beyond reductionism: metabolic circularity as a guiding vision for a real biology of systems. Proteomics. **7**(6), 839–845 (2007)

Dell, H.G., Scott, R., et al.: A Greek-English Lexicon (1996)

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., et al.: Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U S A. **102**(40), 14338–14343 (2005)

Fiehn, O.: Metabolomics—the link between genotypes and phenotypes. Plant Mol. Biol. **48**(1–2), 155–171 (2002)

Fiehn, O., Kristal, B., Van Ommen, B., Sumner, L.W., et al.: Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. Omics. **10**(2), 158–163 (2006)

Figeys, D.: Combining different 'omics' technologies to map and validate protein-protein interactions in humans. Brief. Funct. Genomic. Proteomic. **2**(4), 357–365 (2004)

Fox Keller, E., Harel, D.: Beyond the gene. PLoS One. **2**(11), e1231 (2007)

Gomase, V.S., Tagore, S.: Physiomics. Curr. Drug Metab. **9**(3), 259–262 (2008)

Govorun, V.M., Archakov, A.I.: Proteomic technologies in modern biomedical science. Biochemistry (Mosc.). **67**(10), 1109–1123 (2002)

Greenbaum, D., Luscombe, N.M., Jansen, R., Qian, J., et al.: Interrelating different types of genomic data, from proteome to secretome: 'oming in on function'. Genome Res. **11**(9), 1463–1468 (2001)

Gronow, S., Brade, H.: Lipopolysaccharide biosynthesis: which steps do bacteria need to survive? J. Endotoxin Res. **7**(1), 3–23 (2001)

Haft, D.H., Selengut, J.D., White, O.: The TIGRFAMs database of protein families. Nucleic Acids. Res. **31**(1), 371–373 (2003). https://doi.org/10.1093/nar/gkg128

Han, X., Gross, R.W.: Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. J. Lipid Res. **4**(6), 1071–1079 (2003)

Huang, S., Wikswo, J.: Dimensions of systems biology. Rev. Physiol. Biochem. Pharmacol. **157**, 81–104 (2006)

Huang, S.: Back to the biology in systems biology: what can we learn from biomolecular networks? Brief. Funct. Genomic. Proteomic. **2**(4), 279–297 (2004)

Hucka, M., Finney, A., Bornstein, B.J., Keating, S.M., et al.: Evolving a lingua franca and associated software infrastructure for computational systems biology: the systems biology markup language (SBML) project. Syst. Biol. (Stevenage). **1**(1), 41–53 (2004)

Ideker, T., Hood, L.: A blueprint for systems biology. Clin Chem. **65**(2), 342–344 (2019). https://doi.org/10.1373/clinchem.2018.291062

Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., Brunak, S.: A human phenomeinteractome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. **25**(3), 309–316 (2007). https://doi.org/10.1038/nbt1295

Link, A.J., Phillips, D., Church, G.M.: Methods for generating precise deletions and insertions in the genome of wild-type Escherichia coli: application to open reading frame characterization. J. Bacteriol. **179**, 6228–6237 (1997a)

Link, A.J., Robison, K., Church, G.M.: Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. Electrophoresis. **18**, 1259–1313 (1997b)

Kaeberlein, M.: Aging-related research in the "-omics" age. Sci. Aging Knowledge Environ. **4**(42), pe39 (2004)

Khanna, C.: The dog as a cancer model. Nat. Biotechnol. **24**(9), 1065–1066 (2006)

Kim, T.Y., Sohn, S.B., Kim, H.U., Lee, S.Y.: Strategies for systems-level metabolic engineering. Biotechnol. J. **3**(5), 612–623 (2008)

Kuiper, H.A., Kleter, G.A., Noteborn, H.P., Kok, E.J.: Assessment of the food safety issues related to genetically modified foods. Plant J. **27**(6), 503–528 (2001)

Marri, P.R., Bannantine, J.P., Golding, G.B.: Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol. Rev. **30**, 906–925 (2006)

McGuire, J.N., Overgaard, J., Pociot, F.: Mass spectrometry is only one piece of the puzzle in clinical proteomics. Brief. Funct. Genomic Proteomic. **7**(1), 74–83 (2008)

Mehta, T.S., Zakharkin, S.O., Gadbury, G.L., Allison, D.B.: Epistemological issues in omics and high-dimensional biology: give the people what they want. Physiol. Genomics. **28**(1), 24–32 (2006)

Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., et al.: Calling on a million minds for community annotation in WikiProteins. Genome Biol. **9**(5), R89 (2008)

Morel, N.M., Holland, J.M., van der Greef, J., Marple, E.W., et al.: Primer on medical genomics. Part XIV: introduction to systems biology—a new approach to understanding disease and treatment. Clin. Proc. **79**(5), 651–658 (2004)

Morrison, N., Cochrane, G., Faruque, N., Tatusova, T., et al.: Concept of sample in OMICS technology. Omics. **10**(2), 127–137 (2006)

Oldiges, M., Lütz, S., Pflug, S., Schroer, K., et al.: Metabolomics: current state and evolving methodologies and tools. Appl. Microbiol. Biotechnol. **76**(3), 495–511 (2007)

O'Rourke, N.A., Meyer, T., Chandy, G.: Protein localization studies in the age of 'Omics'. Curr. Opin. Chem. Biol. **9**(1), 82–87 (2005)

Pearl, F.M., Lee, D., Bray, J.E., Buchan, D.W., Shepherd, A.J., Orengo, C.A.: The CATH extended protein-family database: providing structural annotations for genome sequences. Protein Sci. **11**, 233–244 (2002)

Reichert, A.S., Neupert, W.: Mitochondriomics or what makes us breathe. Trends Genet. **20**, 555–562 (2004)

Rochfort, S.: Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. J. Nat. Prod. **68**(12), 1813–1820 (2005)

Schloss, P.D., Handelsman, J.: Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. **14**(3), 303–310 (2003)

Schork, N.J.: Genetics of complex disease: approaches, problems and solutions. Am. J. Respir. Crit. Care Med. **156**(4 Pt 2), S103–S109 (1997)

Steinfath, M., Repsilber, D., Scholz, M., Walther, D., et al.: Integrated data analysis for genome-wide research. EXS. **97**, 309–329 (2007)

Stransky, B., Barrera, J., Ohno-Machado, L., De Souza, S.J.: Modeling cancer: integration of "omics" information in dynamic systems. J. Bioinform. Comput. Biol. **5**(4), 977–986 (2007)

Strömbäck, L., Jakoniene, V., Tan, H., Lambrix, P.: Representing, storing and accessing molecular interaction data: a review of models and tools. Brief. Bioinform. **7**(4), 331–338 (2006)

Taylor, D.L., Woo, E.S., Giuliano, K.A.: Real-time molecular and cellular analysis: the new frontier of drug discovery. Curr. Opin. Biotechnol. **12**(1), 75–81 (2001)

Tracy, R.P.: 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. Curr. Opin. Lipidol. **19**(2), 151–157 (2008)

Werner, T.: Proteomics and regulomics: the yin and yang of functional genomics. Mass Spectrom. Rev. **23**(1), 25–33 (2004)

Chapter 2
Ten Reasons One Should Take Bioinformatics as a Career

Prashanth N. Suravajhala

P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_2

2.1 Bioinformatics Is Challenging and One Is Free to Respect Open-Access

Bioinformatics has enabled wet laboratory biologists to respond to the demands of ensuring quick results for the research done in the wet laboratory. Making the wet-laboratory biologists introducing some methods and predictions to lessen the scale of experimentation would not only help the researchers for educating undergraduate students but also allow them to move forward. As many researchers feel bioinformatics to be not a traditional bioscience, it reflects the growing modularity of biology even as it is equally diverse and has a wide array of solving biological problems.

2.2 It Delves into Predictions but *Bona Fidelity Is* the Means for Predicting Genes

Bioinformatics tools are sometimes trivial but are based on lots of predictions. For example, a protein multiple sequence alignment would delve into the status of which sequence might have evolved first. Whether or not the sequences are related can be interpreted using BLAST against a reference dataset, the annotation associated with potential matches can therefore be used to identify the gene sequences. However, the *bona fidelity* of the sequences will be questioned, if we do not use such sequences that can be aligned with the query protein matching the original query protein (Fig. 2.1).

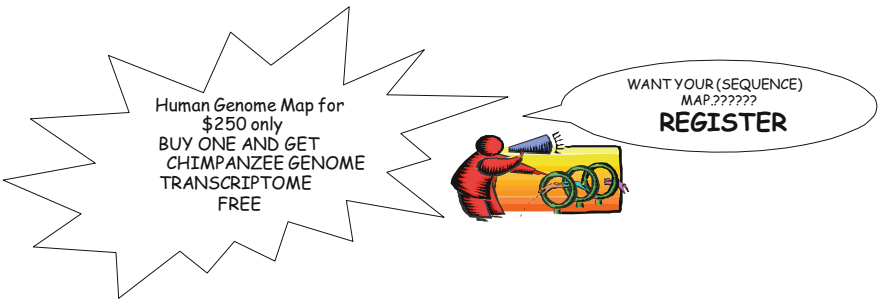


Fig. 2.1 A cartoon depicting the importance of low-cost stride of sequencing

2.3 Intelligent and Efficient Storage of Data Is the Key

To benefit from the bioinformatics opportunities while overcoming the challenges in the post-genomic era, several models are adopted, which demand efficient Information Technology (IT) approaches. These are to be integrated for efficient storage and intelligent data management. Many storage approaches have been deployed widely over the past few years that are insufficient to meet emerging storage and data management challenges, the approaches that treat data in the form of virtual computing are to be discussed.

2.4 Development of Tools and Programs Making Wet Laboratory Biologists Ease Their Experiments

To convince a biologist is like winning an idea. One of the best examples that one could pertain to is designing the primers. Although there are tools available to design primers for a sequence in an efficient way, it is far less useful as they may not get the desired amplicon every time. There are hardly any tools these days that present proof of concept by setting up experimental validation of functionality. As all possible primers are individually analyzed in terms of GC content, presence of GC clamp at 3′-end, the risks of primer–dimer formation, and intra-primer complementarities, a wet laboratory perspective for designing software would make the researchers interested to take up bioinformatics.

2.5 It Is Multifaceted and Brings Networking Among Cross Disciplinarians

Multifaceted disciplinarians and scientists should join hands for a better science. Take the example of developing a web server. The biologist would interpret the background data, a machine learner or a mathematician would think of using support vector machines. Bioinformatics is a multidisciplinary field and requires people from different working areas. It is the combination of biology and IT to discover new biological insights and there is an utmost necessity of tools that helps them to work together.

2.6 It may Partly Influence Animal Experiments

Decades ago, legislation on the use of animals was enacted in many countries involving three R's: Reduction, refinement, and replacement of animal models. Ever since this was enacted, there was a sudden buzz about laboratory animals and their use to be reduced, refined, and replaced wherever possible, for ethical and scientific reasons. The three Rs concept was put forward by W.M.S. Russell and R.L. Burch in 1959 in "The Principles of Humane Experimental Technique." A great detail on the three Rs was reviewed by many in the interest of good and humane science. The word "alternatives" came into use after the publication of the book "Alternatives to animal experiments" by David Smyth in 1978.

2.7 Bioinformatics Curation, not Annotation Is the Key for Databases

The knowledge base (KB) construction and semantic technologies (ST) have been intensely shown great importance in the growth of bioinformatics and computational biology. However, the KBs ensure manual curation is not sufficient for annotation of genomic databases.

2.8 Use of Bioinformatics Methods Propel Contract Research Organizations

A contract research organization (CRO) in bioinformatics is the need of the hour, especially, in clinical research. A CRO can provide services such as commercialization and technology licensing pharmaceutical, assay development, preclinical research, clinical research, clinical phases management, and vigilance. Many CROs specifically provide support for drugs even as software must be made available using the World Wide Web.

2.9 Bringing Core Programmers Closer

Core programmer/developers would not have any interest in biology unless one gets motivated by bioinformatics in the application of their projects. They can only be recruited as mere developers and possibly would be very good listeners as they can easily understand the biology of it. After all, a circuit diagram in the computer chipset is similar to the biology system. Isn't it?

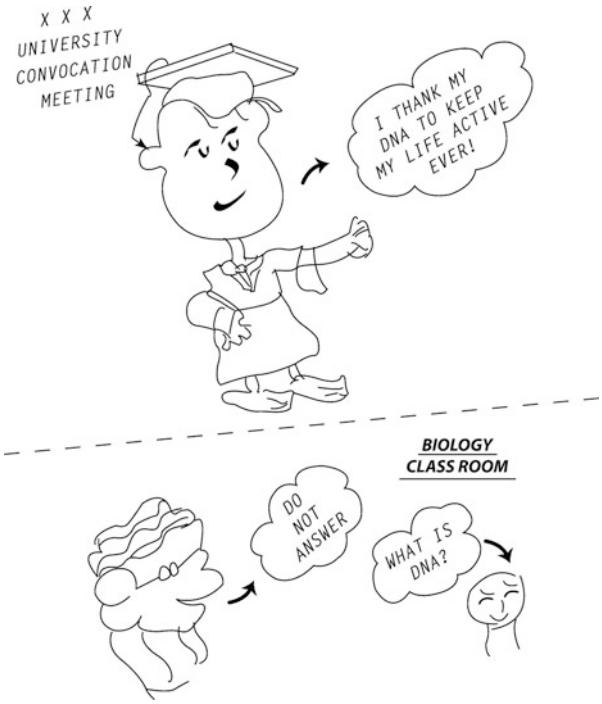
2.10 It Is Dynamic and So Is Inviting to Be Entrepreneurial

As we have documented extensively, R and D through education would have substantial returns in two forms: Privately and socially. The cross-section of researchers could fully utilize an individual's education by becoming entrepreneurs, with returns lower than in the perfect match, they are still substantial. Moreover, let there be looking back at the success rate as it seldom lasts for an entire working life. Educating entrepreneurs is at the high end of the interval and so, investments from angelists, brokers, and venture capitalists are important for offspring benefit, as more educated parents have more successful children.

Chapter 3
Developing Bioinformatics Skills

Prashanth N. Suravajhala

Cs gets degrees?
You never get a second chance to get a first impression
Oscar Wilde



P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_3

Can a mediocre student raise high above standards in bioinformatics? Did achievers succeed abruptly in a first shot? No, they have had tasted lots of failures. That said many graduate students who have taken bioinformatics as a taught program in developing countries have had problems in identifying job prospects.

So, what would be the fate of the aspirants with the third grade? In bioinformatics, we believe they stand tall. Remember as a bioinformaticist, it is your duty to remain multifaceted, think multifaceted irrespective of the stances that you take. Plan your curricula properly. Furthermore, grading standards may become even looser in the coming years, making it increasingly more difficult for graduate schools and employers to distinguish between excellent, good, and mediocre students.

The following are the two plans for the ones who do not get the job offer:

1. Plan for a management degree in bioinformatics: What is seen in a researcher is how he/she manages a laboratory. Managing a biotechnology laboratory is a very important entity for a researcher and so is whether he/she understands the intricacies of the market. For example, an intriguing question one could ask is what if not? There must always be plan B. Biotechnology or bioinformatics are just tools and one needs to be an expert in using them. If there is not a research focus, one can opt for managerial programs which can be worked on the following:

   (a) Transforming biological entities
   (b) Understand the patent regime and monopolies that lead to higher costs for drugs and treatments
   (c) Understanding Intellectual Property Right System with clear background technicalities.

What is often pointed and overlooked, however, is for those who have not at all worked hard to achieve or become successful researchers. The IPR and management courses would typically be abundant. What do you think one can make use of that? It is the student's job to find out whether or not the skills acquired from his erstwhile education interests or provide value to his successful profession. On the flipside, core competencies in bioinformatics can be increased year over year, the average GPA at universities and colleges across the nation is on the rise.

One of the other advantages is that the students may be getting a better education in bioinformatics rather "Exceptional Mediocrity."

While the third-grade students would find it crucial to discuss the integration of biology and information technology (IT) subjects, what about the rest? And you will succeed because you have leadership and communication skills. The ability to sell ice to an Eskimo does not necessarily require college credentials. Can a mediocre student get into bioinformatics? Yes, what all matters is to be disciplined, determined, dynamic, and diligent (The four D rule).

3.1 Be Devoted

Ever since evolutionary biology was developed by Ernst Mayr, many multifaceted and scientific works have been established which effectively brought bioinformatics, one among many regularities into the wider biology and extensive post-synthesis work in systems biology. Making a chief disciplined builder in bioinformatics proves to be an important step in drawing together multifaceted disciplinarians. The bioinformaticists have an increasing sense that 'new' biology-related 'IT' were emerging that would bring together the experimental methods of genetics and IT. It was not until Paulien Hogeweg and Ben Hesper introduced the term in 1978 to refer to "Theoretical aspects of Chemistry and Biology."

3.2 Be Determined

In taking up bioinformatics, one has to understand that determination succeeds in all forms of academia. Understand what is that you are good at, never set aback your temper, and please be advised that you need to face challenges from time to time. The greatest challenge for one would be to liaise between the native fields and acclimatizing it to bioinformatics. Take up a problem, formulate it, and thereon allow the 'D' to answer yourself.

The other Ds are Diligence and dynamism: One needs to identify a great profession in bioinformatics with the focus, diligence, and dynamism of all subjects involved with extremely high standards. This will also allow us to achieve a very good result in acquiring synergies between the multidimensional scientists and ultimately in becoming an even stronger force in the IT and bioinformatics.

3.3 Bioinformatics and the Three Cs of Research

3.4 Consistency

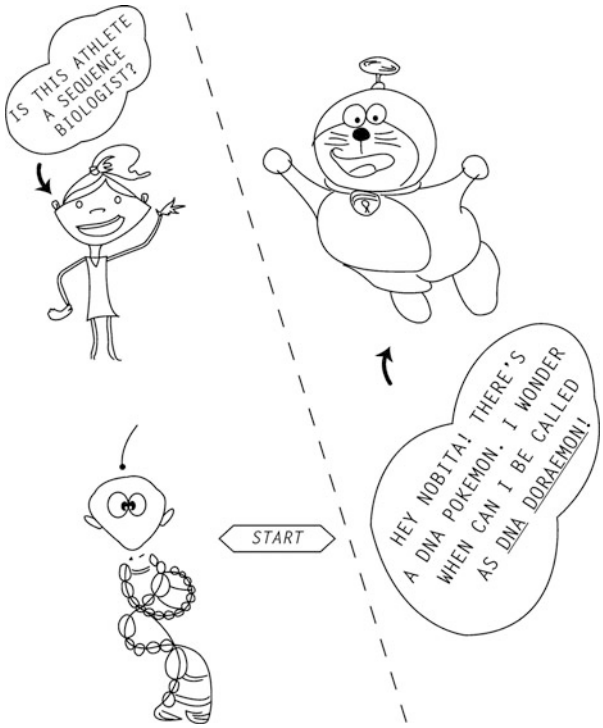
The first C, which one has to follow, is "internal consistency." This C assesses whether or not the candidate who opted for bioinformatics has chosen the same quality, skill, or characteristic. Measured by the precision, this reliable entity often helps researchers interpret data and predict the value of scores and the limits of the relationship among bioinformatics variables. Assume that a researcher designs a questionnaire to find out about bioinformatical problems with a particular focus on

say, cardiovascular diseases (CVD), analyzing the internal consistency of identifying the questions that fall not just on CVDs deals with dissatisfaction, and this we will liaise on the questionnaire focusing on CVDs. What would be the perchance of CVDs making the problem in the future? Are there any genes that makeup the problem would mean that the researcher is trying to put up a brave front in understanding the problem better?

Continuity of Research Efforts: Tracing the communication of scientific and technical information in research is characteristic of two media—the meetings and networking and the journals/way of publishing. As research is a cyclic process, researchers are the producers of scientific information. The continuity of research belies with the sources of information used by active researchers in correlation with the current research taken by them. It is nice if the authors/researchers stick to their same area and develop their specific subareas of research while involved in publications of their article wherein continuing authors could publish a subsequent similar/different article in the same area as their original article. However, there is a lack of continuity of inquiry and progress to move forward in research as authors seldom publish in the applied research but stick to reporting the results of a single study.

Credibility: Biological credibility is what one needs to describe to move forward. Whether or not, there is a logical entity proposed in research, that is, a system with which one can have a causal effect. This proposed mechanism should be consistent with the current understanding of biology. For instance, many researchers were averse to discuss synthetic products especially coming out of artificial expression, which are proteins developed in the laboratory. Scientists rejected the hypothesis because there was no known explanation for how proteins could copy themselves. This also invited discussions from statesmen on the ethics of making proteins and their artificial expression. Furthermore, the use of prions and their spread of an infection to a new individual have changed the scope of protein studies on how they copy themselves in newly infected victims. A biologically plausible mechanism is at stake while we understand how synthetic biology takes shape. So, the need of the hour is to make credible research.

3.5 Hate Wet Laboratory Work?



3.6 Coping the Pressure of Experimental Work

Coping the pressure of experimental work by judging your fairness for research. Ask how good are you able to cope with the loss of your results if you are not able to justify it substantially. You could also ask what type of nonsense experiments would have no results. Are experiments always predictable? Laboratory work is fairly time-consuming and labor-intensive, even as it involves bizarre working hours. Although time-consuming, it would be very rewarding to work in areas of very core/integrative biology, especially looking at fundamental questions on how on earth life evolved and how bioinformatics can leverage huge data? With many researchers in the laboratories renowned for doing extra and long hours with intense work ethics, one needs to ask whether or not it really works in working for long hours, thereby doing better science. Working strenuously in the laboratory is considered one of the

reasons why many people hate working. That said, coping with the wet laboratory work sometimes is seen as "slave-drivers." Thanks to bioinformatics, many researchers started using the tools thereby lessening the scale of experimentation. One would advocate the importance of spending time away from the laboratory works for keeping fresh while the others would claim the workaholism is the only way to succeed. Let there be hope and belief that the scientific community will continue to consider wet laboratory scientists toward creative endeavor, and diversity in the workplace is sure to be encouraged.

The following is the classical example of how bioinformatics has lessened the scale of experimentation. The analysis of proteins in peanut leaves has been shown to have a direct approach to define the function of their associated genes. Proteome analysis linked to genome sequence information was deciphered wherein two-dimensional gel electrophoresis in combination with in silico based sequence identification was used to determine their identity and function related to growth, development, and responses to stresses (Katam et al. 2010). Furthermore, upon verification, we transferred the protein interactors from Arabidopsis to peanut, which has enormously negated the idea of running bioassays like pull-down assays to check for protein interaction partners. In this process, we could be able to identify some potential proteins including RuBisCO, glutamine synthetase, glyoxisomal malate dehydrogenase, oxygen-evolving enhancer protein, and tubulin. Bioinformatical analyses have not only further allowed us to understand how these groups of proteins were accorded to their cellular compartmentalization and biological functionality, respectively, but also led to the development of protein markers for cultivar identification at the seedling stage of the plant.

3.7 From "Hands-on In Vitro" to "Hands-on In Silico"

It would be fugitive to say that all wet laboratory work can be formulated for predictions in silico. It is a certainty that bioinformatics predictions help the wet laboratory biologists to reduce the time frame set for experimentation. What and how good bioinformatics can help experimental scientists to overcome the stress and cope it need to be a foregone conclusion.

Many researchers consider bioinformatics a hackneyed term and do not understand the application of it. There is a lot to read about bioinformatics, more than a tool. Bioinformatics has helped wet laboratory biologists and other cross-disciplinarians tremendously in the last decade even as there has been increased automation in the generation of data from sequenom to phenomes using genotyping information. Bioinformatics has connected biological data to hypotheses by providing up-to-date descriptions in analyzing sequences. This has further allowed sequences to elucidate the literature and study the evolution of organisms. The application of high-throughput DNA sequencers has already provided an overload of sequence data from analyzing DNA and protein sequences, from motif detection to gene prediction, and annotation to curation. There has been a wide focus on gene

expression analysis from the perspective of traditional microarrays by an introduction to the evolving field of phenomics. Furthermore, there are associated mining tools that are becoming increasingly essential to interpret the vast volume of published biological information, while from a developer's point of view, one needs to describe the various data and databases toward common programming languages used for bioinformatics applications.

The following are a couple of case studies that show how one can correlate in silico approaches to a wet laboratory. While the first explains the need for statistical inference for an experiment, the second combines the bioinformatical predictions using already existing wet laboratory data.

3.7.1 Correlating and Identifying Statistically Significant Causality Data

As correlation does not imply, hence, causation is the paradigm behind understanding certain data, namely, biomolecules. From formulating a problem to separating correlated data, causality is a major difficulty in understanding complex differences of molecular biology from a systems level. So, what if we have a data correlated to different groups showing relative antigenicity data linked to them? Could we assign the likelihood of causality to these groups? For example, an antigen A is causal for another antigen B in group X, the reverse causation B being the causality of A may not hold true. Here, we consider antigenic data linked to various subpopulation of HIV infected (±patients) by developing a strategy to determine the causality of antigen-specific data in various subgroups of the diseased population. This white paper may be used as a measure to predict the antigens targeted to various groups, here more precisely "causality."

3.7.1.1 Problem

Identifying variations in the antigens susceptible to various diseased groups of subpopulation provides functional information on how genes lead to disease. Let us consider the below-mentioned data which consists of the peripheral blood mononuclear cells (PBMC) from three groups which were stimulated with different bacterial antigens. In turn, the cytokines produced by PBMCs play an important role in the immune system. Based on the production of cytokines, immunogenicity in these groups could be better understood. We propose a method to sensitize and identifying positive predictive accuracy (PPA) using *t*-tests in predicting the antigens (causal data) immunogenizing the various groups.

3.7.2 Brief Methods

3.7.2.1 Dataset

The three various groups of HIV that were analyzed are as follows:

1. HIV positive
2. HIV control
3. HIV negative and other "diseases" negative

3.7.2.2 Statistical Analyses

1. Positive Predictive Accuracy:
   The sensitivity and specificity do not always provide the probability of a correct hypothesis. At times, we must approach the data using predictive accuracy. The PPA is based on the probability of the antigen to be more immunogenic against a certain group. We used the following three calculations to perform the prediction:
   (a) Sensitivity is the proportion of true positives correctly identified in the data
   (b) Prevalence or likelihood is the number of samples/total number of samples within the group
   (c) Positive prediction accuracy or PPA is calculated using the formula.

$$\text{PPA} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

   However, since we have taken the proportion of true positives to be maximum while ignoring the true negatives, specificity remains out of question, thus making

$$\text{PPA} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} \times (1 - \text{prevalence})}$$

$$\text{PPA} = 1 - \text{prevalence}$$

2. The *t*-test of significance: Student's t-test was used based on sampling the three groups as they have unequal variances. In particular, this test is sensitive and could be used to yield a better probability, a value better than PPA, which might provide ample evidence to support the above hypothesis. While this test is the standard test for calculating the relative efficiency of other tests (in this case PPA), it also requires the most stringent assumptions.
3. Friedman test: The Friedman test is a two-way analysis based on ranks which models the ratings of a (rows) "antigens" on b (columns) "groups." The test parameter here *W* is called Kendall's coefficient of concordance.

$$W = \mathrm{Sum}\left(\mathrm{Rct}^2\right) * 12/\left(a^2 * b * \left(a^2 - 1\right)\right) - 3$$

$$* (b?1)/(b - 1)/\text{Kendall's coefficient}$$

$$Q = a * (b - 1) * W/\text{Chi square}$$

$$\text{Degrees of Freedom(DoF)} = (b - 1)/\text{Number of columns} - 1$$

Furthermore, the mean Spearman rank correlation coefficient (Rsm) between all the rows could also be identified using (a*W−1)/(a−1).

3.7.3 Interpreting the Results Based on Preliminary Analyses Using PPA

In the data provided, we have used all the three groups and scored the sensitivity, prevalence, and PPA. Interpreting the probability using prevalence is crucial and was carried out using the total number of subpopulations in the groups. If the prevalence is low, the PPA is high, which clearly indicates that the data constitute all true positives. Had specificity been inclusive, it is highly inevitable that the results will be false positives. However, a high PPA may indicate that it is statistically relevant to find antigens specific for immunogenicity, but it does not necessarily indicate the presence of immune response. Further analyses on detection and elicitation of immune response using cytokines were carried out wherein the preliminary analyses using a *t*-test of significance showed that the groups 1 and 2 are statistically more significantly compared to the third group. Considering the fact that group 2 is a control dataset, we would, however, find it to be significant; hence, a test of the hypothesis was carried out to find statistical significance.

T-Test of Significance

- For groups 1 and 2: $t = -12.01$, DF = 25, $p\backslash = 1.615\text{e-}09$
- For groups 2 and 3: t = 12.59, DF = 25, $p\backslash = 1.045\text{e-}09$
- For groups 1 and 3: t = −13.58, DF = 25, $p\backslash = 5.352\text{e-}10$.

Where N is the total numbers (population), DF is degree of freedom ($N-1$), and t, the test parameter = (Mean/SD)*sqrt(N) and p, the probability. Since t is substantially same for groups "1 and 2" and "2 and 3," group 3 is ignored for a very less probability.

Friedman Test

Kendall's coefficient $W = -3.29$
Chi-squared $Q = -164.44$*X^2
Degree of freedom DoF = 2, $p\backslash = 1.00000$

The level of significance, $p\backslash = 1.00000$, given above is based on an approximation of the chi-squared distribution. Another statistical significant concordance is if the test parameter Q is high (i.e., statistically significant), then the columns are known to be different and the rows are correlated, which in our case, the dataset holds true (Q being relatively high).

3.7.4 Predicting the Antigens Immunizing the Groups

From the above analyses, it is clear that the three methods we employed are independent of each other and are sensitive to apply statistical significance. Furthermore, sensitivity in groups 1 and 2 when averaged (0.25 and 0.58, respectively) (see data below), sets the mark for identifying the candidate antigens. Hence, the antigens whose sensitive values are par below the above-mentioned respective values for the groups are discarded (see the other data tabled) while the rest are used to identify cytokines against the antigens that play an important role in the immune system.

3.7.5 Conclusions

The antigens, namely Z, A′, and B′ are specific to the two groups 1 and 2. Based on the values obtained, the production of cytokines specific to the group as against the antigens can be identified from the data. Furthermore, these could also be used to integrate co-expression networks and genotypic data. If the data constitutes expression traits, we could establish statistical significance using the methods discussed above. The causal predictions that were made (see Table 1.2) could in turn be used wherein the data may be divided into training and testing data randomly in 8:2 ratios. To avoid the selection bias, training set cross-validation of 10-fold could be carried out producing an accuracy. While testing on the remaining 20% test dataset, the predictive accuracy could again be established using a radial basis function support vector machine (RBF-SVM) kernel. The use of an SVM-based classifier gives the best result among all other classifiers, but the limited accuracy performance might challenge the machine learning classifier.

3.7.6 Bottom Line

1. Which methods are readily implemented and able to extract biologically relevant causal connections among genes?
   - PPA, *t*-test, and Friedman tests
   - Support vector machines (SVM)

2. If a method employs data normalization, what are the strengths and weaknesses of the normalization algorithm in terms of facilitating data analysis and interpreting the data and results?
   - Strengths: Sensitive and highly accurate, and easy to store data
   - Weakness: Loss of data as it is pretty difficult to ascertain
3. What are the pros and cons of each method?
   - PPA

     Pros: Direct disease could be ascertained
     Cons: It is extrinsic, meaning it is always dependent on other factors, namely, prevalence.
   - The *t*-test

     Pros: Sampling huge datasets, correlation
     Cons: The user must be aware of how big the sample is and what for the data are to be used.
   - Friedman test

     Pros: Easy ranking and nonparametric test
     Cons: High rate of false positives and dependent on other datasets
   - SVM

     Pros: Prediction accuracy is always high
     Cons: Cannot be used for training if the data are less and mediocre.
4. Are there pitfalls to avoid with a given method or circumstances under which the method may be less reliable?
   - Using SVM, the number and inappropriate set of descriptors and multiple target classes make this method more cumbersome to act as an efficient tool especially to predict genes.
5. What criteria were used to evaluate and rank the methods?
   - Ranking is based on the existence of antigens specific to a group and whether or not a cytokine particular to the group is produced.

3.8 Case Study on Nematome: Protein Interactions Specific to Parasitism in Nematodes

The nematodes, commonly called the roundworms (belong to phylum Nematoda), are the most diverse of all animals. Of more than 28,000 described so far, approximately 16,000 are parasitic. The parasitic nematodes especially those from plants have not yet been known better. Furthermore, genome sequences of the plant-

parasitic nematodes are just beginning to yield results even as RNA-Seq (transcriptomic) analysis is being done by us and many other laboratories. With *Caenorhabditis elegans'* genome completely sequenced way back in 1998, nematode genome sequences hold a great promise to understand the umpteen nematodes' genomes waiting to be sequenced. It has been known that the genomic repertoire and gene-centered density is roughly about 1 gene/5 kb with 24% introns on an average across all nematodes. While many genes are arranged in the polycistronic series of operon models, there holds greater importance to understand mitochondrial genome as well owing to the identification of parasitic genes in nematodes. The RNA-Seq and RNAi studies have started yielding results too with biology curators appraising the set of known genes even as the predictions need to reach consensus along with the flourished datasets of ESTs, RNAseq, and genomic repertoire. So that begs a question of whether or not any commonalities of all these genes are known across all nematodes? The answer belies in how and what kind of organisms are these: Parasitic, nonparasitic, entomopathogenic parasitoids, non-entomo nematodes, and so on.

For example, *Caenorhabditis briggsae* genome and further comparative genomic analyses determined the novel gene sequences from the same genus nematodes such as *Caenorhabditis remanei, Caenorhabditis japonica* which further enthused knowledge that they might be less likely to complete and remain accurate than that of *C. elegans*. That said, the worm-based database is void of many plant-parasitic nematodes. Further understanding to nematodes has revealed that there has been an accelerated rate of evolution in the parasitic lineages while several phylogenetically ancient (Read parasitic) genes might have been lost and found elsewhere across all nematode species. In that process, RNA interference (RNAi) experiments leading to gene loss of f unction were done even as researchers were able to knock down about 86% of the*20,000 genes in the worm with an established functional role mounting to 9% of the nematode genomes on an average. However, the story does not end here with ascribing function to the genes as the aforementioned methods are mediocre and involve lots of false-positive datasets. With systems biology burgeoning, there is a need to understand the how of interactions in nematodes. The *C elegans* protein interaction network (PIN) was a masterpiece of genomic catalog of protein–protein interactions (PPI).The interactions have been established based on the small-scale experiments while there is need to complement the bona fide interaction studies with large-scale datasets. Predicting the PIN across individual nematode genomes involves lots of experiments, reactions, and importantly wastage of man-hours. Therefore, we wish to propose a uni-comparative biology approach to predict PPI across ergonomically important nematodes of parasitoids, *viz.* root knot, migratory, and most damaging. Nevertheless, the interactions can be ascertained and the function can be better ascribed across these nematodes wherein we would identify the proteins involved in parasitism. That said, we propose a word called nematome fort hose set of (most commonly occurring) proteins implicated in parasitism.

3.8.1 Methods

Wet laboratory experiments.

Plasmid Construction

1. Two-hybrid constructs: The cDNAs encoding full-length parts will be amplified by PCR using gene-specific primers containing REs with EcoRI and BamHI restriction sites. The PCR products would then be digested with REs and then ligated.
2. Plant/nematode expression constructs: The constructs used to transiently express the interactor proteins will be based on a plasmid, namely, pMON999, which will contain the proteins specific to the promotor and terminators (van Bokhoven et al. 1993). The cDNAs of the positive clones will then be excised from the vectors using EcoRI and ligated in the EcoRI site to allow expression of the interactors tagged with other proteins.

3.8.2 Interaction Analyses

- Yeast two-hybrid screening (or co-immunoprecipitation): The initial experiments will be enthused with a *C. elegans* cDNA library in the desired vector (Promega/Clontech). This library thus should be constructed with all possible independent cDNAs. Colonies are then selected on agar plates lacking histidine, tryptophan, and leucine over a 7-day period while positive yeast transformants will be picked up and replated Gal assay. A positive interaction can then be determined by the appearance of blue colonies and the plasmids can be isolated. In subsequent experiments, bait and prey constructs containing full-length proteins or domains will be analyzed and later can be transfected.
- Immunofluorescence labeling and fluorescent microscopy: This can be performed essentially based on the aforementioned results and these can be imaged using fluorescent microscopy.

3.8.3 Dry Laboratory/Bioinformatics

1. Traditional mapping of interactions for parasitic genes with respect to functional annotation.
2. Interolog mapping: While it is known that the orthologous genes are highly conserved between closely related species, we presume that the systems might utilize the same genes and share interactant information across the orthology datasets across different organisms. However, it does not necessarily mean that the amount of sequence conservation is directly proportional to interaction even

as certain studies comparing high-throughput data including expression, protein–protein, protein–DNA, and genetic interactions between close species show conservation at a much lower rate than expected. We would like to identify those parasitic genes from step (1) and show that conservation is maintained between species albeit through network modules. Furthermore, we would like to employ a confidence score for interactions based on available experimental evidence and conservation across species. (Please refer flowchart below.)

3. Filtering the datasets and reaching the consensus: We would then filter the interaction datasets and integrate them with a high-confidence interval thereby reaching consensus. This would ensure that the estimated size of the parasitic interactome of nematodes (nematome) would have an approximate number of interactions. Comparison with other types of functional genomic data would show the complementarities of distinct experimental approaches in predicting different functional relationships between genes or proteins. Finally, we would like to compare them against different tissue-related proteins with respect to co-immunoprecipitation (CoIP) assays. A further re-examination of the connectivity of essential genes in nematodes could support the presumption that the number of interaction partners can accurately predict whether a gene is essential and if essential to which organelle. This would yield organelle proteomic analyses. In conclusion, our analysis should facilitate an integrative systems biology approach to elucidating the nematode cellular networks that contribute to diseases (Fig. 3.1).

3.9 Tips and Traps in Writing a Research Article in Bioinformatics

Many consider writing the article/proposal to be the toughest and perhaps most boring part of the entire report-writing process. The best way to start a project work especially Ph.D. is to start a review. The student must have a firm grasp on the topic that they have been introduced to or the work they have been acquainted with, and therefore put in countless man-hours of the literature review, formulate a problem, and finally then submit the results in writing which is a difficult task, to say the least. The following guiding principles could provide the reader with tips on taking up the problem, defining it, and the logistics of the write-up, and so on.

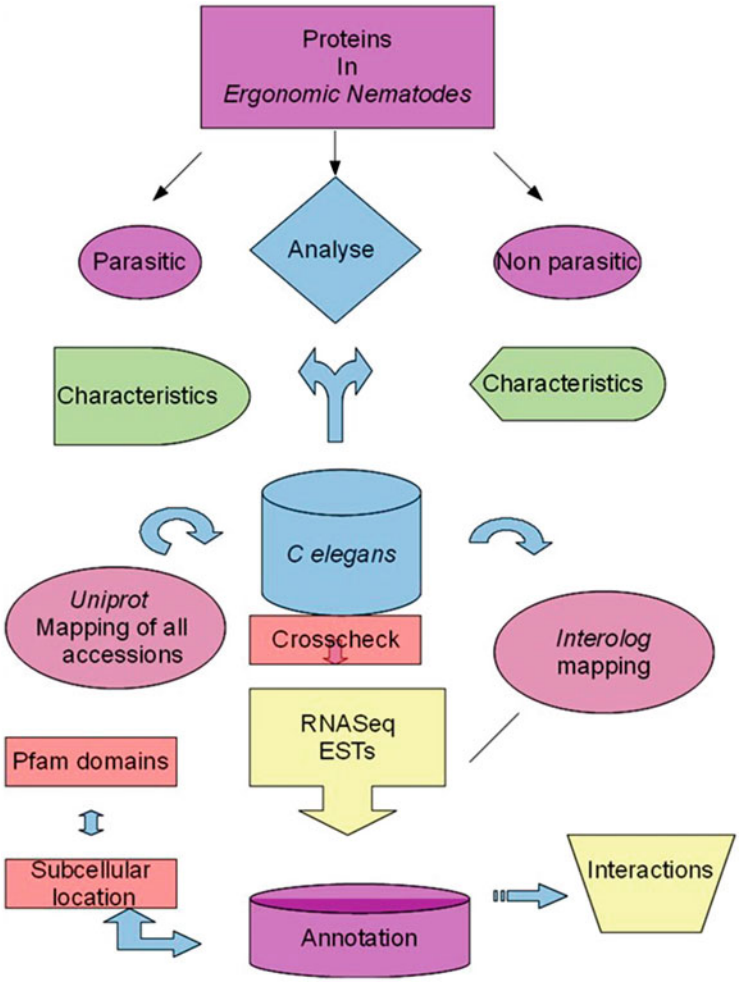


Fig. 3.1 A pipeline for making a "Nematome"

3.10 Convert Ideas and Thoughts into Action for a Strong Problem Formulation

There are changeable ideas always about how long or how short a review/paper proposal should be. Conversely, the proposal can be called as white paper if it is addressed during its preparation. But please ensure that in the former, all the points and ideation process are covered while in the latter, a concrete dialog of proposal should be summarized.

3.10.1 Address the Problem Well with Subheadings

3.10.1.1 Background

The background is an important summary of the major points addressed by the erstwhile researchers. These questions behind your research could be good compendium to address the current problem you are to address in the future and provides the context of those questions within a larger academic framework. This precisely is a kind of pre-introduction and one should be able to see who read the introduction should be able to understand what you are attempting to discern through your research and writing.

3.10.1.2 Introduction and Review of Literature

This portion should address the scope of research while listing major findings. Whether or not one sets out a specific portion of the proposal for a literature review is to the discretion, it would be nice if a fair problem is specifically indicated wherein bioinformatics has been employed to lessen the scale of experimentation. Some points on introducing a pictorial representation or Gantt chart would also be inviting wherein some of the results can interlace within the other major portions of the intending proposal. Regardless of how one would present the literature review, we could describe the findings of the review specific to the importance of the problem chosen in the area of custom research.

3.10.1.3 Problem Formulation and Objectives of the Study

It would be nice if the researcher describes how bioinformatics can leverage issues as depth and provide the background and particular context of the problem in relation to the particular academic field. The objectives can be described on a point-to-point basis not making up to one page.

3.10.1.4 Materials and Methods

What is the overall plan of the experiments that will be done and why planning these bioinformatics predictions and experiments' including annotation and curation is an important element of the dissertation. All the detailed methods of research to be demonstrated relating to the question and problem formulation should also be addressed.

3.10.1.5 Results and Discussion

This should specifically focus on what was aimed during the dissertation frame, and the results, although not preferentially discussed as per the objectives, may include the contents with pictorial and graphical representations with concise statements debating and reaching the consensus of the objectives of the study. It would be nice if the points are split and an appealing statement can be made to be inviting for the reader to ensure there is a flow of very good reading. Many a time, the reader falls in trouble in not understanding the scope and depth of the problem. Especially, if it were bioinformatics predictions, what makes the predictions *bona fide* would mean that the works are more precisely specific and with respect to the growth of the works.

3.10.1.6 Conclusions and Future Directions

Conclusions typically are written shorter and play an important role in bringing together the main areas covered in the erstwhile analyses as described until the Results and Discussion. Furthermore, it would also work on giving a kind of prefinal comment or judgment making suggestions for improvement and speculating on future directions. Although conclusions are likely to look more complex, it is to be noted that the significance of the findings and recommendations for future work are to be brought to the notice of the reader in this section wherein important implications are to be covered. The future scope and the upcoming challenges negating the line of false positives would really provide the reader and the dissertation very inviting not only toward the research but also for the reader who can cite your work in many ways.

3.10.1.7 References/Bibliography

Making a list of all source materials and properly formatting them in whatever academic style is required for a complete dissertation, which will address the problem of writing with elegance.

3.10.2 Plan Your Next Steps and Always Give Plenty of Time

One good thing that always takes us to move forward is to plan the way we write the report. As bioinformatics involve lots of predictions, the wise thing would be to estimate how many hours/days you can start working in writing the manuscript. The last thing at the last minute is to understand the fact that the write-up or proposal may even take up to 6–12 months to complete or sometimes even longer. Always ensure

that there is room for tackling questionnaires, discussion with your peers, and most important, call for constructive criticism.

3.10.3 Discuss with your Peers

Always broader understanding of the research topic helps because when it comes to the dissertation proposal, many a proposal is negated and they eventually get off the ground even without an adequate review. That will allow us to invite contacts in the form of peer reviewers and committee members to prepare the report which plays a vital step to stay in touch for the future. Furthermore, they are available anytime for open advice anytime. There is nothing wrong with asking for copies of previous reports so that these, which were once approved, would allow you to prepare with ease and satisfaction.

3.10.4 Accept Constructive Criticism

Peer reviewing and the purpose of a support group are to ensure we read each other's work and give feedback. In doing this, we will not only help improve each other's writing but also allow us time to read while evaluating the work and later providing the feedback. That said, one need not be an expert and so let not content or vocabulary intimidated. Even if not familiar, a strong argument can be made on how the manuscript has been written and addressed within the scope of the journal or subject, and so on.

3.10.5 Publish or Perish Is the Key While Citing and Cross-Referring Other Articles of Interest

After the research is well taken into, it is recognized only on account of publications. The world knows only after the work is published where people come to know about it. Adding references to such work into their publication text and list of references can be seen as a kind of good normative citation. That said, references can be divided into single units, whereby each reference turns into a citation that can be aggregated in many different ways, forming a wide range of citation impact factors/indicators (CIF). It is like many such articles are references and indexed in citation indexes, such as the Thomson Reuters' Web of Science database even as many online repositories such as Google Scholar, BioMed Experts in the form of "scientometrics" work for the cause. The key here is in the competitive field, publish, or perish.

3.10.6 Peer Review Holds an Important Community Service

Peer review is the key! The more you review, in all likelihood, the more you will be asked to review. Oftentimes you may be asked to review boring papers that are of no interest to you. While it is important to serve as a reviewer, only accept papers in which you are keenly interested, because either they are close to your area of research or you feel you can learn something. You might say that should I not know the work very well to be a reviewer? Often a perspective from someone in a slightly different area can be very effective in improving a paper. Editors would of course like to see your review papers even if you are not particularly interested in them, but the reality is that good reviewers must use their reviewing time wisely.

References

Katam, R., Basha, S.M., Suravajhala, P., Pechan, T.: Analysis of peanut leaf proteome. J. Proteome Res. **9**(5), 2236–2254 (2010)

van Bokhoven, H., Verver, J., Wellink, J., van Kammen, A.: Protoplasts transiently expressing the 200K coding sequence of cowpea mosaic virus B-RNA support replication of M-RNA. J. Gen. Virol. **74**(Pt 10), 2233–2241 (1993). https://doi.org/10.1099/0022-1317-74-10-2233

Chapter 4
The Esoteric of Bioinformatics

Prashanth N. Suravajhala

There are few people who can understand the intricacies of bioinformatics. Through myriad bioinformatics predictions, the wet laboratory observations, and experiments can be developed to illustrate ideas on problem formulation based on genes or proteins. Topics include the biology of the system, sources, and effects of bioinformatics predictions, characterizing the uncharacterized genes and genomes, and applications in medicine and agriculture.

Basic skills in integrated biology would be a plus which includes knowledge of biochemistry, bioinformatics, and molecular biology. One need not have to be an expert in all, but it helps to have at least some background in each field. Amazingly, bioinformaticists hail from many other disciplines such as statistics, computer science, and mathematics. The smartness in getting into bioinformatics depends on how good the researcher is getting acclimatized to through experience he/she had gained in his/her field. We will provide tips on how to be a winner in bioinformatics using bioinforma TICKS (*Please see Frequently Asked Questions*). Through limitless observations and discussions on experiments, we develop and illustrate the following ideas on how esoterically bioinformatics can be in nutshell.

4.1 Bioinformatics Market: Hype or Hope?

The last couple of decades have seen the bioinformatics market significantly evolving across the world on the back of the rising omics industry. With an increase in the application of these omics-es in biotechnology, there has been a commercial market

P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_4

for bioinformatics worldwide. With declining costs of per-base genome sequencing, introduction of next-generation sequencing (NGS), widespread interest in Genealogy, micro-RNA research, the introduction of aptamers replacing antibodies, and so on. Public and private sector investment has given a significant boost to the industry. It has been known that the biggest field of the global bioinformatics industry has been into sequencing, software, and services with respect to IT infrastructure. With the software segment improving its share, the database market will suffer the downturn due to the increasing popularity of innovative analysis software including that of SAS/inbuilt software for companies. Applications of bioinformatics in genomics, proteomics, and pharmacogenomics have furthered genome studies that completely transformed basic research. For example, diagnostics specific to cancer has become the leading therapeutic area wherein bioinformatics is a big support, be it outsourcing or finding drivers pushing the market. As the market has been witnessing the launches of key bioinformatics products and services in various areas, these developments might impact the biotechnology industry's future performance, and therefore, the competitive landscape for this need to be done.

4.2 Decoding Genes Using Genealogy: What Bioinformatics Can Do?

"DNA transcripts RNA and RNA translate proteins" have been the central dogma of molecular biology and with the identification of genes contributing to diseases well represented through the central dogma of bioinformatics using *sequence predicts structure predicts function*, direct method for discovering the molecular pathways involved in their pathogenesis or function of these proteins per se could be interesting to delve. What interests' researchers to delve into the genome? Let us imagine 10 years down the line what are the diseases one could predict beforehand and perhaps have a diagnostic moiety represented so that humans can beware of? That said Iceland based decode genetics have virtually established this using a built-in system of information linking medical information from patients. However, there are lots of ethics that one needs to exploit and understand and use this interpretation with specific informed consent even as disease-by-disease studies and molecular genetic information could be ascertained.

4.3 Communication Between Organelles and the Genes

Communication between organelles and the genes can be done with respect to the systems biology context. A system can be better described as an entity constituting major components and minor components. While major components, namely, tissues, organs, and organelles, compete for space, minor components such as genes, proteins, and enzymes compete for interacting with each other or compete with the

analogous components. This competence results in ascribing function. One of the major competitors is the presence of subcellular sorting signals for the proteins localized to different organelles. For example, researchers have identified key protein determinants targeting mitochondria. These protein determinants falling at least four subclasses of mitochondrial-targeted proteins containing targeting signals directed to different sites within the mitochondrion (outer membrane, intermembrane space, inner membrane, and matrix) by diverse mechanisms (Bolender et al. 2008). Conversely, some of these proteins are not known and are yet to be discovered, remaining as "hypothetical." Recently, a dog was used as a cancer model (Khanna 2006), which enabled the researchers to understand whether there are any unknown genes or genes encoding some HPs involved in diseases (Snyderman and Langheier 2006). The dog genome study has initiated exploring some diseases whose genetic linkage is not yet known. We think that some of the orthologous regions of these HPs probably were mapped, providing clues to study important diseases.

4.4 Pull-Down Assays and the Role of Bioinformatics

Earlier, researchers' idea that electrospray could spray and ionize molecules using mass spectrometry was well-conceived, and the substances were analyzed to ionize them. However, subsequently, small molecules would play an important role in analyses wherein samples from the patients could be directly analyzed. Researchers now have been working on similar ideas with proteins. However, there are several limiting factors in measuring the mass of proteins and so is the case with sequencing peptides. The Edman degradation test has been in use to determine what amino acid it was. With mass spectrometry techniques burgeoning, tests such as ED were shelved out bringing the development of more robust technologies such as tandem MS and TOF. Thanks to bioinformatics, even there are efficient bioinformatics databases employed with MS such as Mascot, a powerful search engine to identify proteins from primary sequence databases. The idea of obtaining proteins from the gel and still analyzing them very sensitively using mass spectrometry was a part of pioneering studies by Mathias Mann's research group back in 1996. With the proteins very difficult to get out of the gel, it has been eventually discovered that the short part of the protein sequences, aka peptide sequence tag, can be searched using bioinformatics algorithms that are well-known and searched for similar sets of sequences with equivocal function be ascertained. So, if one would ask whether proteomicists' view of mapping all proteins on this 2D gel electrophoresis was deployable, it must be noted that it has its own merits and demerits. Furthermore, out of the limitations and less potential use, variants of MS play an important role. The electron spin ionization MS has enabled characterizing the protein complexes where one can look at a small number of proteins that had some functional context; nevertheless, we are now able to do it at one go. Consider an example MLH1which is

involved in DNA mismatch repair. However, it has not been known that DNA mismatch repair exists in mitochondria, but we knew that from a bioinformatics point of view and interolog mapping, we could get interesting candidates and so believe that MLH1might be inherent to mitochondria. To run pull-down assays, we need antibodies raised against the bait (MLH1 here). The idea is that the proteins when pulled down would interact with its prey proteins and therefore, would be associated with each other and so the interactions and functions could be transferred. If you had the antibody, you could pull out not just the receptor, but the other players in the known mitochondrial pathway. Very recently, we could use aptamers which completely are sensitive, and *bona fidelity* is at the means of cost-effectiveness when compared to antibodies. Aptamers, until recently, were used for RNA chimeras, and not long ago we discover that these can be used for pull-down assays, especially for CoIP experiments sensitizing the tagged assays in determining the function of hypothetical proteins.

That said, determining the HPs targeted to mitochondria specific to MLH1 could be interesting. There are currently 1185 HPs in humans. While searching for human mitochondrial proteins, we augmented the fidelity of in silico selection strategy in searching for candidate HPs targeted to mitochondria. In this process, the human mismatch repair protein hMLH1 that we explored revealed that it is not localized to mitochondria, while the hMLH1 is known in nuclear extracts of human cells. A greater amount of gene diversity remains to be studied across many DNA mismatch repair proteins including that of hMLH1. In that process, we believe many HPs targeted to mitochondria (read putative DNA mismatch repair) could be interesting candidates to study the dynamic behavior of genes and establish how HPs are distributed and altered. The work could be interceded based on wet laboratory and in silico experiments. First, to check for candidate HPs targeted to mitochondria, we can run pull-down assays. Whereas the above wet laboratory experiments can be carried out, we can also work based on the interaction studies. We could deduce putative protein interactions that we are already establishing specific to hMLH 1and diseased candidates. From the interaction map, we could find the nearest interacting partners of the protein and then model the genes involved in diverse functions, and specifically, that of the HPs targeted to mitochondria. Above all, while hMLH1 is just considered as an example, we can consider any genes or set of genes to study how gene functions are altered. Finally, a web server (with script complemented) is to be developed to find how genetic variation is ascribed to genes.

As MS data analysis is endless and limitless, we feel it is quantitative especially when bioinformatics has made leaps and bounds in the recent past. Previously, one used mass spectrometry, as we are trying to do on MLH1, to find and sequence a single protein. Now thereon, quantitative measurements on it can be done with the kinds of things with proteins that people have only been doing so far with mRNA and microarrays. The advantage of doing them at a protein level is that proteins are the functional agents. When you look at mRNA on chips, you have a question as to whether the change you are seeing is at the mRNA level or down at a deeper level of regulation. Maybe, it is at the protein level. Nowadays, we can read out the proteome in a quantitative way, in a large-scale fashion. Another area of work that is more

specific to mass spectrometry is to look at modifications. We not only want to know what proteins are present in a sample but also how they are modified. Are they in active status? Are they phosphorylated, for example? This can now be done in a large-scale way by mass spectrometry. We can now look at the proteome quantitatively and examine how mitochondrial pathways change by looking at how the proteins are modified, whether they are phosphorylated. By doing so, we have a very good handle on how cells process information. This will be a big theme and esoteric in the fields of bioinformatics and systems. There is another new field called "interaction proteomics." Here you use mass spectrometry and proteomics to see which proteins talk to which other proteins. Could the MLH1 be used as a potential biomarker for diseases? And that is also finally coming within reach. We hope that with further development we can look at the proteins in a urine sample, for instance, and then use them to classify patients. What diseases would we be covering and following? What drugs do they respond to? It is limitless, isn't it?

4.5 Say "Ome" Using Essential Bioinformatical Indicators

Access to huge bioinformatics data is essential for understanding what kind of data would be useful to experiment in the laboratory. Thanks to many omics-es that have steadfastly been available. Among several impediments to generating data and accessing in the laboratory, an important incentive for scientists to publish research articles in bioinformatics with explicit recognition from wet laboratory perspective, collaborate with scientists, has tacit conversations through academic communities. Ome is many or monies. Saying "ome" is the key to become a successful bioinformaticist.

4.6 Ten Career Options to Opt Through Bioinformatics

1. Scientist

   WHAT: You chose to become a Scientist.
   WHO: A Ph.D. in Biology/Biotechnologies/Informatics is needed. However, MScs with research with an exceptional track record can apply for positions too.
   WHEN: A Ph.D. with a couple of years of postdoctoral experience can start applying for positions.
   WHERE: In research organizations, collaborative institutes, and so on.
   HOW: You will be responsible for the following:

(a) Envision and build databases/web servers for wet laboratory researchers to share their biological data. In the case of hospitals, help create personalized data warehouse medicines with individuals' genetic code and biochemistry. Create computer tools to track and analyze the patterns of viral outbreaks, such as flu, around the country.

2. Lab manager
3. Professor
4. Research fellow
5. Entrepreneur
6. Analyst
7. Consultant
8. Technology licensing officer
9. Business manager
10. Research associate

References

Bolender, N., Sickmann, A., Wagner, R., Meisinger, C., Pfanner, N.: Multiple pathways for sorting mitochondrial precursor proteins. EMBO Rep. **9**(1), 42–49 (2008). https://doi.org/10.1038/sj.embor.7401126

Khanna, C.: The dog as a cancer model. Nat. Biotechnol. **24**(9), 1065–1066 (2006)

Snyderman, R., Langheier, J.: Prospective health care: the second transformation of medicine. Genome Biol. **7**, 104 (2006). https://doi.org/10.1186/gb-2006-7-2-104

Chapter 5
Common Minimum Standards: A Syllabus for Bioinformatics Practitioners

Prashanth N. Suravajhala

Disseminating key survival messages on bioinformatics with emphasis on common minimum standards for bioinformatics education could bring a rise in awareness of the need for nonformal and formal education programs. The following are the brief subheads of what an undergraduate in bioinformatics could be taught:
Introduction

- What and the how of bioinformatics.

Bioinformatics to Computational Biology

- From a mere "Tool" to a science.

Need for bioinformatics today.
Applications of bioinformatics in various disciplines. Current prospects and future challenges.
Homology and similarity searches using bioinformatics.
Advanced similarity searching on the Web.
Using Blast on the Web.
Searching sequence databases for predicting structures.
Phylogenetic and Multiple Alignment Tools

- CLUSTAL and PHYLIP.

Sequence-based Taxonomy

P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_5

- From multiple alignments to phylogeny.

Validating consensus sequences.

Primer Designing In Silico

- Properties of a bona fide primer.

Primer Designing Tools

- Primer Blast and e-PCR

A pilot experiment for designing and synthesizing a primer in vitro. Hands-on with emphasis on the participants' favorite genes.
Introduction to computational evolutionary biology.

Bioinformatics for Evolution

- Validating novel proteins

Ks/Ka score detects evolution. Detecting selection in sequences.

Functional Genomics

- Annotating genomes to proteomes. Why need curation?

Validating complexity in sequences/conserved regions.
Genome Projects

- Comparative genomics

Advances

- The HapMap project

Poor Man's Genomes

- An EST perspective

Expression Data in Genomes

- SAGE

Web-based practices and analyzing assembled genomes.
Systems to Synthetic Biology.
What is Systems Biology?
Systems Biology of aegis.
Top-down and bottom-up.Protein–protein interactions (PPI).
Wet laboratory methods employed for analyzing PPI
Interpreting the Data

- In silico tools and visualizers for Systems Biology

Introduction to (Now) Next-Generation Sequencing

- Challenges
- Tools for next-generation sequencing

Semantic technologies for biologists.
Interpreting genes and proteins using rearrangement.

Conserved DNA sequences

- Understanding promoters and restriction sites.

Domains, motifs, patterns, and SNPs. Comparative genomics of regulatory regions. Introduction to bio programming.
Subcellular localization studies and the role of genes as probes.
Miscellaneous topics of interest.

Chapter 6
Colloquial Group Discussion on Bioinformatics: Grand Challenges

Prashanth N. Suravajhala

This is the outcome of the GD addressed by multifaceted disciplinarians. The following text is colloquial and the reader may consider it to be the voice of all the participants.

Bioinformatics was highly evolved in early 2000 and all of a sudden fallen and not much talked by the end of the decade. Various schools and universities have recently started a high-end program on bioinformatics in Western countries where as in developing countries, the taught programs are weakened on the premise there is a demand for expeditious faculty. With researchers scaling the ladder of bioinformatical progress by leaps and bounds, there is a need to identify the why and the how of lacuna for bioinformatics. Some of the excerpts from the consensus points of the GD titled 'bioinformatics: Visions and Challenges for the next decade' are as follows.

We considered how bioinformatics may evolve in the future and what challenges and research is needed to realize this evolution. With an increase of diverse resources, we suggest bioinformatics will evolve by bringing biologists together to understand what precisely the 'B' word is. In other words, we all agreed that multifaceted disciplinarians play an important role to evolve bioinformatics research in developing countries for that matter anywhere. While many researchers consider bioinformatics a threadbare term, few do not understand the application of it leaving lots to read about bioinformatics, more than a tool. How bioinformatics will help wet laboratory biologists and other cross-disciplinary scientists were discussed in great detail. It was felt that the biologists can easily understand and interpret the results of bioinformatics compared to bioinformaticists because of the girth of understandability they have. Basic skills in integrated biology would be a plus which includes

P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_6

knowledge of biochemistry, bioinformatics, and molecular biology. One need not have to be an expert in all, but it helps to have at least some background in each field. Amazingly, bioinformaticists hail from many other disciplines such as statistics, computer science, and mathematics. The smartness in getting into bioinformatics depends on how good the researcher is getting acclimatized to through the experience he/she had gained in his/her field. That said, we also discussed how to be a winner in bioinformatics and perhaps not make paraphernalia.

While all of us agreed to the fact that bioinformatics in India and developing countries has been hyped a lot, the Information technology (IT) aspect of it was considered to be one of the reasons even as a high salary market for bioinformatics was assumed to play wet blanket. Many people also agreed to the fact that it is not really a problem to digest the big B word, with no compulsion set to it. We also agreed that funding, reaching consensus; flexibility, and collaboration are the keys to bioinformatics in Indian research to move forward. On a note on whether or not we are good at writing research grants, there was a split in the opinion, wherein many opined that yes, it could be; while bureaucracy hinders the innovation and ideas. That said, we also herald the discussion for respecting open-access, which is stellar for bioinformatics development.

Bioinformatics as a subject should be taught at school and college level along with major subjects or it should be a part of computer application. That said, students who come out of college or universities will have an idea about what actual bioinformatics applications can be leveraged in solving major biological problems. The government could also take the initiative to encourage bioinformatics by giving funding to the projects, facilities, and positions in undergraduate schools and colleges. Likewise, all faculties could visit colleges and schools once in a month and cater to the understanding of bioinformatics in the schools. With the world facing a lot of problems in the environmental and medical issues, we reached a consensus that bioinformatics is the solution through combined efforts. Traditionally, we have a lot of acceptance to the novelty and inventory through the three C's in practice: Creativity, credibility, and continuity. That said, we can grow high lest we realize that the growth of bioinformatics in India is in our hands for that maxim we need to follow: To be enthused is to be infused with life!

6.1 Opinion of Bioinformatics Practitioners

Madhan Mohan opines ...

1. How should bioinformaticist makeup the blend of being a biologist as well as a programmer?
   Right from the career point, the candidate has chosen, the student should be preparing himself for problems specific to both biologists and informaticists. I suggest students give emphasis to understanding the programming logic, make an algorithm on white paper, and then try compiling the program, testing it several

times. Initially, there could be acclimatizing problems for students where biologists may not find programming logic interesting as most of them hail from pure biology background and not taught mathematics during their intermediate schooling whereas in the Indian scenario at least it has not been difficult for programmers to understand biology. Interestingly, the latter genres of students have become the most fitting entrepreneurs in bioinformatics. In conclusion, try achieving it, not thinking of failures.

2. How best to formulate a problem for research in bioinformatics? What are the challenges one should pose?

   As a bioinformatician, please understand that there are genres of multifaceted disciplinarians who have a background in physics, statistics, informatics, biochemistry, and so on. Try making a preproposal and collaborate with biologists and biochemists who do not have a background in programming. That said, problems described through programming should be understood with basic logic and then algorithms could be written. The main challenge is to understand the problem and solve it whereas if you are a programmer, simply solve it.

3. Given the high-dimensional genomic data generated from time to time, how should a bioinformaticist keep himself updated?

   Just follow the new publications that are available in reputed journals such as Nature, Science, Public Library of Science, and many more. You will find the greatest updates at the three genebanks available worldwide, namely, NCBI, DDBJ, EMBL-EBI, and from India—IISc, IMT, MKU, and University of Pune websites. The updates can be checked through "What's new" and importantly comparative analyses of genomes, with an advent of next-generation sequencing (NGS), new genomes can be studied.

4. Open-access and ethics play a very important role for the biologist. Do you agree? If so or not, how good are these to be exploited in bioinformatics? Please give valuable suggestions.

   I slightly differ considering the fact that there should be closed access and some data may be kept confidential until it gets through the file of patenting records. Ethics also clubs to closed access. Thanks to the introduction of the new patenting regime in India!

5. Making a successful principal investigator after years of postdoctoral experience is a need by choice. Please provide suggestions with an example specific to the development of bioinformatics in academia and industry.

   I would say that a Ph.D./Postdoc making a career front could go for tenure track positions which need not really affect his career. While, it need not be a choice, but it would essentially allow him to make a better PI. As discussed earlier it is not specific to bioinformatics but lest bioinformaticist work for both biology and information technology principles.

Pawan Dhar writes

1. How should bioinformaticist makeup the blend of being a biologist as well as a programmer?

Bioinformatics offers an opportunity to help understand biology more accurately. What is accurate can be programmed. What can be programmed, can be understood. Thus, bioinformatics is a reasonable way to blend biological concepts and programming tools to help understand biology better. In my opinion, first students must identify the widest variety of data that makes an organism. Second, students must know the biological context in which the data makes sense. Third, they must know the right mathematics and computational tools to play with the data. Finally, an attempt should be made to build models that represent biology as we know it. The payoff is that if we understand biology, we can compose organisms. Students should clearly understand that if they are good biologists, only then can they be good bioinformaticians.

For bioinformatics students to become good biologists, they should be trained to write algorithms that accurately abstract biological processes, from molecular expression to cell–cell interaction and onward. To appreciate the beauty of scale and complexity, it would be useful to train bioinformaticians in experimental data collection—from sequencing to expression, structure, and flux measurement technologies, and so on.

2. How best to formulate a problem for research in bioinformatics? What are the challenges one should pose?

To formulate a problem, the following steps may be helpful.

Step 1: Read the discussion section of good scholarly papers published in the last 2–3 years.

Step 2: Make a document of unanswered questions. The questions usually appear in the form of direct statements or loosely thrown hints.

Step 3: Extract common and unique questions. The common ones are those that the community is often discussing and may be important. The unique ones will be those that the authors are thinking about and could be important. One should expect to see some noise in such data.

Step 4: Match the questions with your interests and available facilities, funding, and so on.

Step 5: Add more questions with the hope to extend the intellectual front-end of the field.

Step 6: Predict the value of answer (i.e., nonobviousness and scale), if one successfully addresses the problems.

Step 7: Assemble questions into an integrated research problem and imagine practical applications that would possibly emerge at the end of the project.

With reference to the challenges, there are at least two different approaches:

1. For those who want to play safe, look for a challenge that would result in incremental but useful innovation.
2. For those who like disruptive innovations, either (i) look for simple things that have been ignored, which, if properly addressed, may result in groundbreaking work, for example, BLAST, or (ii) scale up the work to an embarrassingly complex level, for example, virtual organisms.

Since research is essentially an art to predict useful ignorance, students must be exposed to creativity and innovation exercises, in addition to getting trained in formal bioinformatics.

3. Given the high-dimensional genomic data generated from time to time, how should a bioinformaticist keep himself updated?

   Keep track of publications in key journals, follow conferences and symposia where speakers are likely to showcase the latest unpublished results and make announcements. These days one need not physically attend an event as several talks appear online, both real-time and also as a repository.

4. Open-access and ethics play a very important role for biologists. Do you agree? If so or not, how good are these to be exploited in bioinformatics? Please give valuable suggestions.

   Yes, in my opinion for every scientist. Both open-access and traditional models have advantages and limitations. However, given the financial crunch that publishers are facing in the traditional model, it seems to me that increasingly journals are going to opt for an open-access model in the future. There are many bioinformatics journals that offer open-access option. One may start more such journals provided the focus is not just to make huge profits only but to identify the niche and maintain the quality, which in my opinion and experience is extremely difficult to balance. Given the fast-paced competitive world that we are part of, I support the idea of "first publish and then defend." Another idea is to get the paper prereviewed by the scientific community and publish it straight away. I recognize that there are flaws in every publication model.

5. Making a successful principal investigator after years of postdoctoral experience is a need by choice. Please provide suggestions with an example specific to the development of bioinformatics in academia and industry.

   The following points apply both to academia and industry:

   - Focus on algorithms more than tool development
   - Move from sequence level to the pathway level and tissue level
   - Find strategies to integrate every type of data that organisms offer
   - Develop reliable literature mining tools such that one can build good molecular interaction models straight from the literature
   - Maintain the database, if you have built one. However, maintain that the database is infeasible, merge it with a more established larger database
   - Design standards of data exchange and BioCAD tools for constructing organisms

Cox Murray writes

1. How should bioinformaticist makeup the blend of being a biologist as well as a programmer?

   Bioinformatics is really made up of three parts—biology, mathematics and statistics, and computer science. While bioinformaticists obviously need some coding ability, it is far more important that they have a good grasp of biology and can frame biological questions as a logical series of analytical steps. Too many

students with an interest in bioinformatics neglect vital areas of study, such as algorithms and statistics. A given problem can be coded in many ways, but a good understanding of algorithms can inform which are most efficient, or indeed, which are computationally tractable. Similarly, it is now easy to generate vast amounts of bioinformatics' data, but telling which patterns are meaningful remains much more difficult. It is increasingly important to have a solid grasp of statistics, including Bayesian and Monte Carlo approaches, especially as datasets get bigger and it becomes easier to misinterpret spurious, nonsignificant patterns in biological datasets.

2. How best to formulate a problem for research in bioinformatics? What are the challenges one should pose?

   I would consider that there is no such thing as a bioinformatics' question. There are, however, biology questions, some of which can be addressed using computational approaches. The questions bioinformaticist addresses should always be driven by the underlying biology. In this sense, it is important to distinguish whether a question can best be answered computationally or in the laboratory, or better yet, using a combination of both approaches. Some of the best bioinformatics studies are those where computational and laboratory-based researches inform and support each other.

3. Given the high-dimensional genomic data generated from time to time, how should a bioinformaticist keep himself updated?

   Good bioinformaticists require solid working skills in biology, mathematics and statistics, and computer science. Few researchers are equally conversant in all three areas. I advise my students to focus on upskilling in their weakest subjects. It is also important to keep track of subject areas that are changing particularly fast, which in practice, unfortunately, tends to be all three. It is often tempting to find quick, project-specific solutions to individual problems, but I encourage my students to discover generic solutions wherever possible. This means you already know the answer when you invariably encounter a similar problem again in the future.

4. Open-access and ethics play a very important role for biologists. Do you agree? If so or not, how good are these to be exploited in bioinformatics? Please give valuable suggestions.

   Issues around open-access and ethics are not specific to bioinformatics. The best policy is to follow ethical guidelines for the associated biology field, which usually have the same broad underpinnings, but different specific requirements. Open-access publishing is quite a different concern and has its pros and cons. I am very supportive of making research results more accessible to the general public, as it seems unfair that taxpayers who have already funded research should have to pay again to access its results. Nevertheless, it is important to recognize that there are genuine, nontrivial costs associated with publishing, and these costs are not substantially lower for electronic-only publications. Open-access publishing pushes these costs from the reader to the author, and in doing so, introduces new problems. Publishing costs, which can easily exceed US$2000–3000, limit who can afford to publish in open-access journals, and can often make researchers

choose what research they want to release. Unfortunately, this divide is not obviously one between developing versus developed countries, but instead often falls within countries between well-resourced and poorly resourced research groups. Open-access publishing is still an emerging phenomenon, and the progression and sustainability of this model are yet to fully play out.

5. Making a successful principal investigator after years of postdoctoral experience is a need by choice. Please provide suggestions with an example specific to the development of bioinformatics in academia and industry.

   Although again not specific to bioinformatics, postdoctoral training is increasingly viewed as a necessary evil. The complexity of most modern research questions means that training beyond a Ph.D. degree is required for most scientific jobs. I encourage students to expand their research horizons during their postdoctoral training by learning new skills unrelated to their Ph.D. research. Training in a foreign country often proves extremely useful as well, particularly given the increasingly international nature of modern research programs. The Ph. D. degree can now be considered basic training, while postdoctoral positions allow students to mature into well-rounded scientific researchers.

   (a) Challenges and road ahead: Key skills and knowledge for bioinformatics
   (b) Bioinformatics as a career

Jeff W Bizzaro, President of bioinformatics.org opines

1. How should bioinformaticist makeup the blend of being a biologist as well as a programmer?

   Bioinformatics is cross-disciplinary. If your intention is to enter it from any one of the other STEM fields (Science, Technology, Engineering, and Mathematics), you will need to supplement your education with courses or background material in certain places. For example, if your studies have or had a major focus outside of the life sciences, then you should also study the basics of biology, with an emphasis on genetics, genomics, and proteomics. Likewise, if you have a strong background in the life sciences, you will need to learn about computer programming, databases, and statistical analysis.

   However, just as it would be impossible for a biologist to become familiar with every single topic in biology, no one can expect to have a fully comprehensive education in bioinformatics. If you consider the fact that the field is at the intersection of several subjects, each requiring years of mastery, you will appreciate that the makeup of a bioinformatics research group is complementary by necessity. Once you do have a solid education in one major field and have completed some cross-disciplinary studies, it would therefore be best to quickly choose an area of interest within bioinformatics—a specialty.
2. How best to formulate a problem for research in bioinformatics? What are the challenges one should pose?

   As a bioinformaticist, you are likely to collaborate with biologists and biochemists who do not have a background in programming. Nevertheless, any problem that involves programming can be described algorithmically, even

employing some brainstorming sessions at a marker board. At such times the common "language" is science, and there may be little need to elaborate on the details of any programming that will be involved.

3. Given the high-dimensional genomic data generated from time to time, how should a bioinformaticist keep himself updated?

   You will find the greatest change in the available data if your interests include comparative genomics (the comparison of genomes between species) or newly investigated species. However, the genomic data for even the most highly studied organisms may already be somewhat static, in which case your research could involve lesser-known aspects, such as gene regulation, epigenomics, protein–protein interactions, or protein structures.

4. Open-access and ethics play a very important role for biologists. Do you agree? If so or not, how good are these to be exploited in bioinformatics? Please give valuable suggestions.

   Ethics in bioinformatics can be thought of as a dichotomy between the need to reveal and the need to protect, perhaps more so than in any other field of science. On the one hand, science itself depends upon the free exchange of information, particularly for the purpose of reproducing and verifying results. Science as we know it cannot exist without such a collegial atmosphere.

   One the other hand, individuals feel a need for privacy regarding their medical information, and so data that come from human trials or medical records need to be given special consideration. Along the same lines, companies and academic institutions also share an interest in profiting from discoveries and innovations and are often required by law to protect their information if they are to claim certain rights to them.

5. Making a successful principal investigator after years of postdoctoral experience is a need by choice. Please provide suggestions with an example specific to the development of bioinformatics in academia and industry.

   A postdoctoral appointment is a necessary internship for those pursuing a career based on a cycle of grant proposals and subsequent funding. It could also be important in an industry where the research and development phase of a product would be funded in a similar fashion. The role of a bioinformaticist, however, is oftentimes in support of larger projects that are not computational by nature. Drug discovery is an example of that. In such a case, a master's degree with several years of experience would be suitable for an employer.

Chapter 7
The Bioinforma "TICKS": Frequently Asked Questions

Prashanth N. Suravajhala

Abstract What is bioinformatics?

Bioinformatics is a tool, whereas computational biology is a discipline. Bioinformatics predicts the biological outcome and can be used to compare the biological data, for example, sequence analyses and structure prediction. In a nutshell, bioinformatics predictions can lessen the scale of experimentation. Bioinformatics can be considered as a method to annotate the newly sequenced genomes. It can be well defined in the biological and computational way.

1. What is bioinformatics?

   Bioinformatics is a tool, whereas computational biology is a discipline. Bioinformatics predicts the biological outcome and can be used to compare the biological data, for example, sequence analyses and structure prediction. In a nutshell, bioinformatics predictions can lessen the scale of experimentation. Bioinformatics can be considered as a method to annotate the newly sequenced genomes. It can be well defined in the biological and computational way.

   Definition from biologists' perspective. Application of informatics and statistics to solve, analyze, annotate, and organize biological data in graphical and browsable formats.

   Computational scientists' perspective. Design computational algorithms and applications for solving biological problems. By analyzing the existing biological data using information technology, we can predict the biological outcomes. Planning the analysis by workflow using bioinformatics tools and knowing the expected output of the workflow will help to predict and solve the biological problems. The bioinformatics era has been started, and data are generated in huge

P. N. Suravajhala (✉)
Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India

Bioclues Organization, Hyderabad, India
e-mail: prash@bioclues.org; http://bioclues.org

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_7

amounts by next-generation sequencing (NGS) in every field of biology, increasing the need for bioinformatics analysis.

2. Where can I be placed? Are there any companies working in the area of core bioinformatics research?

   There are many institutes that require bioinformaticists to work with. After all, the role of bioinformaticist/bioinformatician is to help wet laboratory biologist plan his experiment or lessen his scale of experimentation using in silico methods. Say in India,

   - Indian Institute of Science (IISc), Bangalore
   - Indian Institute of Technology (IIT)
   - Indian Institute of Science and Educational Research (IISER)
   - National Center for Biological Sciences (NCBS), Bangalore
   - Institute of Bioinformatics and Applied Biotechnology (IBAB), Bangalore
   - Jawaharlal Nehru University (JNU), New Delhi
   - Jawaharlal Nehru Center for Advanced Scientific Research (JNCASR)
   - National Institute of Mental Health and Neuro-Sciences (NIMHANS).

   And a host of all bioinformatics centers developed by the Department of Biotechnology, Government of India.
3. Can I get placed in companies? Are there any companies that have bioinformatics resources/placements?

   - Astrazeneca
   - GVKBio
   - AravindBio
   - Reddy's laboratories
   - Ocimumbio.com
   - Biocon

4. What are the branches/fields of bioinformatics?

   (a) Computational Biology
   (b) Drug Designing
   (c) Phylogenetics
   (d) Structural bioinformatics
   (e) Population Genetics
   (f) Genotype Analysis
   (g) Systems Biology
   (h) Synthetic Biology
   (i) Functional Genomics

5. People say that bioinformatics has no scope. Is it true?

   No, it is not so. Research is measured with publications, and now, almost all high impact factor journals are accepting with bioinformatics analysis in the articles. This shows the importance of bioinformatics in all fields of biology.

6. Do I need programming experience to become a good bioinformaticist?

   Yes, one does need programming experience for that. It helps to understand the most bioinformatics tools and their functionality; maybe you do not write your programs, but surely it is an asset to learn more as a bioinformaticist.
7. How is chemoinformatics different from bioinformatics?

   Cheminformatics deals specifically with the chemical structures, whereas bioinformatics deals with the biological systems and signaling pathways. Both the fields are devised so as to be able to manage huge data easily and come up with respective tools and techniques to study the same.
8. I am a B.Tech graduate. How can bioinformatics help me?

   There are two options available: The first one is to go for M.Tech and then Ph.D., and the second option you can opt to work as a project assistant with funded projects or as a trainee/junior post with a bioinformatics organization. If you like to have a good grip on bioinformatics and start your career with a good level, better go for the first option.
9. Whither bioinformatics?

   When it comes to bioinformatics, the biologist has the opinion of just storing the date or searching from different databases such as doing the BLAST search, and now things have changed with the time. As the different genome project moved up and algorithmic solution needs with large data, thus biology itself has changed from a dogmatic, "disciplinary" or "pathway-based" science, to a broader, multidisciplinary exercise.
10. What's new in bioinformatics?

    According to Shankar Subramaniam of the University of California, San Diego, there is a new "central dogma"; genomes code for gene products, whose structures and functions are embedded in the pathways and physiology of biological activity. Each metabolic pathway can no longer be considered in isolation but in the context of the interlocking and cross-coupled networks in which each component of that pathway participates. So, the next solution is with a bioinformatician. According to Leroy Hood, founder of the Institute for Systems Biology in Seattle, such an approach not only needs a greater infrastructure (DNA/gene expression array technologies, proteomics, multiparameter cell sorting, mass spectroscopy, single-cell assays, etc.) than traditional disciplines (molecular/cellular biology, biochemistry) but also requires advanced computational technologies. The major challenge the bioinformatics/system biology is facing, for now, is trained bioinformatician and sufficient funding; here at Bioclues, we have taken the one challenge to have 2020 bioinformaticians by 2020.
11. Who coined the term bioinformatics?

    Paulien Hogeweg of the University of Utrecht, Netherland, coined the word Theoretical Biology in late 1980s.
12. How good the salary would be for a bioinformaticist?

    It depends from one country to the other. As far as India is concerned, for a beginner, one can expect 30k per month while medium-level scientists 40k and a Senior level 60k.

13. Can I do a Ph.D. in bioinformatics? Where?

    Of course, you can. But please understand that bioinformatics is a tool. You may have to complement the wet laboratory analysis done by someone or you need to liaise with a wet laboratory biologist.

14. Is there any integrated course/curriculum for bioinformatics?

    No. But IBAB, Bangalore has recently started the program. Many IISER institutes in India have recently started a 5-year integrated program catering to the needs of student scientists.

15. Where can I undergo training after undergraduation/graduation in bioinformatics?

    After completing your graduation/undergraduation, try to seek a position in a reputed laboratory, which is working on bioinformatics. In this way, you would learn many tools and techniques which shall be adding to your profile.

16. I want to do a project in bioinformatics. Can you suggest to me?

    For doing a bioinformatics project, try approaching the people who are actually working on it in various research institutes by surfing the internet and writing those emails. Apart from this, you can enroll for a live virtual project with Bioclues itself and get real-time problems to solve under the guidance of top-notch scientists.

17. If I take up M.Sc. bioinformatics, won't my area be more specialized and narrower? Can you suggest to me to take up a broader area for my masters?

    Your subject would be more specialized as compared to any other broad field. No doubt about it, but you would be an expert in it. If you are confused as to whether bioinformatics is your cup of tea or not, then go for a broad subject for your masters in which you may study one paper on bioinformatics and later on pursue higher studies in the same to have the expertise.

18. I have done my B.Tech in bioinformatics. I am planning for my masters. I am confused between MS (Research) and M.Tech. I would like to know your valuable opinion on the career prospects of MS in biotechnology and MS in bioinformatics as per the industry standards.

    We would suggest you to always go by "interest" because opportunities reckon by interest not necessarily by choice. MS by research is a "mentored" degree, and as a protege, you will be free to undertake a project of your interest. Typically, it lasts for 1.5 years with a small amount of time dedicated to teaching the program. Both MS by research and M.Tech allow you to gain in-depth exposure to the component parts of bioinformatics. While the former focuses on research, the latter on pure taught program.

Chapter 8
Undergraduate Education in Bioinformatics—Progress and Lessons Learnt from an Engineering Degree

Bruno A. Gaeta

More than 20 years since Russ Altman's landmark bioinformatics editorial (Altman 1998) started the discussion on bioinformatics education, a standard bioinformatics curriculum remains a moving target as the field and associated technologies continue to evolve. What has become clear is that a single curriculum for all students is unfeasible due to the wide variety of interested students. Distinctions have been made between "training" (applied, skill-specific, most often targeted at biologists) and "education" (a formal course or degree programs with an emphasis on theory) as well as between courses targeted at "tool-users" versus "tool-builders". The curriculum task force of the International Society for Computational Biology (ISCB) suggested three main "personas" for students: bioinformatics users—for example, clinicians or wet lab biologists who make use of a small subset of tools relevant to their domain expertise; the biology-focused bioinformatics scientists who work toward biological discovery primarily by computational means, and the more computing-oriented bioinformatics engineers who develop new tools and set up the necessary computational infrastructure for the other personas (Welch et al. 2014). While these personas are necessarily an oversimplification in an increasingly popular and complex field, they provided a useful tool and inspired a community effort to identify suitable core competencies that define the knowledge and skills expected of bioinformatics graduates at a range of levels. The refinement of these core competencies is an ongoing effort (Mulder et al. 2018; Welch et al. 2016), which has drawn on the feedback from an increasing number of bioinformatics educators worldwide including GOBLET, the Global Organisation for Bioinformatics Learning, Education and Training (Atwood et al. 2015).

A parallel effort has focused on the importance of bioinformatics content in undergraduate life science education. While there is a consensus that bioinformatics

B. A. Gaeta (✉)
School of Computer Science and Engineering, UNSW Sydney, Sydney, NSW, Australia
e-mail: bgaeta@unsw.edu.au

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_8

must become an essential component of the life science undergraduate curriculum, there is some disagreement as to the nature and amount of computer science and mathematics required. Pevzner (2004; Pevzner and Shamir 2009) has argued that solid theoretical foundations in computer algorithms and statistics are essential so that life science graduates can be "bioinformatics scientists", as opposed to "bioinformatics technicians" only able to use and apply existing tools. Pevzner's view of the bioinformatics scientist is close to that of the ISCB scientist persona, while their bioinformatics technician is closer to the ISCB bioinformatics user persona. Another perspective has emerged from NIBLSE, the Network for Integrating Bioinformatics into Life Sciences Education, who surveyed life science educators across universities (mainly US-based) and used the results to develop a set of core bioinformatics competencies in life science undergraduate education focusing on the informed use of tools and data sources in bioinformatics, with only a light emphasis on algorithms and computer science (Porter and Smith 2019; Wilson Sayres et al. 2018). The NIBLSE competencies are more consistent with a user persona than a scientist persona.

Core competencies and minimal standards (for example, Tan et al. 2009) provide a useful framework for evaluating courses and degree programs, as well as formulating general guidelines in the design of new programs, but their generalization to all types of bioinformatics practitioners at all levels of expertise remains difficult. The Mastery Rubric for Bioinformatics (Tractenberg et al. 2019) is an attempt to add a longitudinal element by focusing on knowledge/skills/attributes (KSAs) displayed by bioinformatics students/practitioners at multiple levels of development (Novice, Beginner, Apprentice, and Journeyman). The level of detail of the Mastery Rubric makes it well-suited to external and self-evaluation of individual skills in addition to that of curricula. The rubric is mostly tailored at bioinformatics scientist and user personas whose goal is biological discovery, rather than method or infrastructure development. This reflects the fact that these personas represent the majority of current employment opportunities in bioinformatics. Educational programs aimed at the bioinformatics engineer persona, while essential in driving the development of the field forward, are less common.

One example of an engineering-focused bioinformatics degree is the Bachelor of Engineering (Bioinformatics Engineering) program offered by UNSW Sydney (www.engineering.unsw.edu.au/study-with-us/undergraduate-degrees/bioinformatics-engineering). Starting in 2001, it is the oldest continuously running bioinformatics degree program in Australia. The program is formally accredited as an engineering degree by Engineers Australia, the engineering peak body in Australia. This means that its graduates are recognized as professional engineers by all the national engineering bodies in countries signatories to the Washington Accord (www.ieagreements.org/accords/washington/), including ABET in the USA. This accreditation introduces several constraints on the program but also results in graduates with a strong set of technical and practical design skills.

Difficulty in designing the bioinformatics engineering program was the large number of foundation courses required in bioinformatics. Students need to study introductory biology, computer science, engineering, mathematics, chemistry, and

physics, as well as sophomore or senior statistics, algorithms, programming, databases, molecular biology, and genetics. Bioinformatics is taught at the sophomore and senior level both from "tool-user" and "tool-builder" perspectives. Students also take courses in engineering ethics, design, and project management.

A number of lessons were learned over the course of designing, running, and revising the program over the 20 years since its creation.

1. Keep options open for students: When the program started, bioinformatics employment opportunities were still rare in Australia and the program was designed to provide enough versatility for graduates to be highly employable in other fields including IT and computing, biomedical engineering, and biotechnology. Over the course of the program around 25% of graduates have gone to work in bioinformatics, as engineers or scientists, and 50% have chosen employment as engineers and consultants in the IT industry or computer science with no biology applications. The remainder has gone on to postgraduate retraining, for example as medical doctors and biomedical professionals. The fact that many students are keen to keep their options open is shown by the fact that more than half of the students elect to study bioinformatics jointly with another degree such as biomedical engineering, commerce, or science.
2. Involve colleagues from other faculties and industry: Bioinformatics is interdisciplinary, whereas universities are typically organized in disciplinary "silos" in departments, schools, and faculties. Providing a good interdisciplinary experience for the students will require drawing on expertise from more than one department. Since its inception, the bioinformatics engineering program has drawn on expertise not only from its home school (Computer Science and Engineering) but on lecturers from biotechnology and biomolecular sciences, mathematics and statistics, medical sciences, and physics. One feature of the program is a number of practical workshop courses where the students work in teams on engineering projects provided by "customers" from research groups in the university and associated medical research institutes. A number of software applications and pipelines have resulted from these courses, as well as several publications, and the students continuing their work with the customers as undergraduate or postgraduate thesis students.
3. The pros and cons of an undergraduate degree: As an undergraduate program, the bioinformatics engineering program accepts students straight out of high school. This is a strength of the program as students learn from the beginning to think both as life scientists and as quantitative scientists/engineers. Many bioinformatics programs around the world are taught at the master's level, by which stage the students have already specialized either in the computer or life sciences for their bachelor's degree, and may require significant retraining and refresher courses. Some of the bioinformatics courses in the program are offered to both undergraduate and postgraduate students (coursework and research), which allows a direct comparison of their performance. As a rule, undergraduate students in the courses get much better marks than the postgraduate students in the same course, who often have difficulties with either the biology or the computing components.

However, the lack of awareness in high school in the field of bioinformatics means that relatively few students choose to study undergraduate bioinformatics engineering. Enrolments have been only around 10–15 students a year until recently. In order to diversify the offering and recruit additional students, UNSW also introduced a bioinformatics major as part of its standard bachelor of science degree. Since science students typically do not choose their major until their second year, this allows more students to become aware of bioinformatics as an option. The bioinformatics science major shares many courses with the engineering program but has less technical content. It aims to graduate "bioinformatics scientists" as defined by the ISCB personas.

4. Review and revise: As a field, bioinformatics has gone through a number of changes since the 1990s, as new omics technologies and data types emerged and took over the field, and the bioinformatics content of the program must keep up with these changes. The program must also be altered periodically in response to constraints imposed by the university administration, course changes in other schools/faculties, and by external accreditation bodies such as Engineers Australia. In addition to these university-mandated revisions, both the engineering and science bioinformatics programs are periodically reviewed using the ISCB core competencies. The process used is to map the learning objectives and assessment items in individual courses in the program to the core competencies and building a competency matrix that allows identifying areas of strength and weaknesses in the curriculum. The program is then adjusted when possible to address the weaknesses.
5. Focus on long-term, transferable skills: The demand is high in the biology research community for graduates trained in currently "fashionable" bioinformatics techniques, and this demand is most often served using short focused training courses which are best targeted at the bioinformatics user persona. However, the field also needs graduates with solid foundational knowledge to adapt to, adopt, and develop the methods of the future. A program aimed at bioinformatics scientists and engineers should aim to provide foundations and skills that the graduates can build on for the rest of their careers in addition to current trends.
6. Network with other educators: Unlike established disciplines, there is no "standard" curriculum for bioinformatics. While there is now an abundance of textbooks, they are often already out of date by the time they are published. Resources for practical teaching of bioinformatics also change very quickly. It is therefore important to keep in touch with other bioinformatics educators to exchange information about best practice and useful tools. Organizations such as GOBLET (www.mygoblet.org) and ISCB (www.iscb.org) are extremely useful for meeting others who are facing similar challenges, and for sharing resources and ideas that ultimately feedback into the development of bioinformatics beyond a niche discipline.

A recent surge of interest in the UNSW bioinformatics undergraduate degrees suggests that students are becoming increasingly aware of the importance of the field. This is good news for bioinformatics educators but also provides a challenge to

keep the curriculum relevant and suitable for current and future developments in the field.

References

Altman, R.B.: A curriculum for bioinformatics: the time is ripe. Bioinformatics. ***14***(7), 1–2 (1998)

Atwood, T.K., Bongcam-Rudloff, E., Brazas, M.E., Corpas, M., Gaudet, P., Lewitter, F., et al.: GOBLET: the global organisation for bioinformatics learning, education and training. PLoS Comput. Biol. **11**(4), e1004143–e1004110 (2015). https://doi.org/10.1371/journal.pcbi.1004143

Mulder, N., Schwartz, R., Brazas, M.D., Brooksbank, C., Gaeta, B., Morgan, S.L., et al.: The development and application of bioinformatics core competencies to improve bioinformatics training and education. PLoS Comput. Biol. **14**(2), e1005772–e1005714 (2018). https://doi.org/10.1371/journal.pcbi.1005772

Pevzner, P.A.: Educating biologists in the 21st century: bioinformatics scientists versus bioinformatics technicians. Bioinformatics. **20**(14), 2159–2161 (2004). https://doi.org/10.1093/bioinformatics/bth217

Pevzner, P., Shamir, R.: Computing has changed biology—biology education must catch up. Science. **325**(5940), 541–542 (2009). https://doi.org/10.1126/science.1173876

Porter, S.G., Smith, T.M.: Bioinformatics for the masses: the need for practical data science in undergraduate biology. Omics. **23**(6), 297–299 (2019). https://doi.org/10.1089/omi.2019.0080

Tan, T., Lim, S., Khan, A.M., Ranganathan, S.: A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. BMC Genomics. **10** (Suppl 3), S36–S36 (2009). https://doi.org/10.1186/1471-2164-10-S3-S36

Tractenberg, R.E., Lindvall, J.M., Attwood, T.K., Via, A.: The mastery rubric for bioinformatics: a tool to support design and evaluation of career-spanning education and training. PLoS One. **14** (11), e0225256–e0225229 (2019). https://doi.org/10.1371/journal.pone.0225256

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., Schneider, M.V.: Bioinformatics curriculum guidelines: toward a definition of core competencies. PLoS Comput. Biol. **10**(3), e1003496 (2014). https://doi.org/10.1371/journal.pcbi.1003496

Welch, L., Brooksbank, C., Schwartz, R., Morgan, S.L., Gaeta, B., Kilpatrick, A.M., et al.: Applying, evaluating and refining bioinformatics core competencies (an update from the Curriculum Task Force of ISCB's Education Committee). PLoS Comput. Biol. **12**(5), e1004943–e1004944 (2016). https://doi.org/10.1371/journal.pcbi.1004943

Wilson Sayres, M.A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A.G., Smith, T.M., et al.: Bioinformatics core competencies for undergraduate life sciences education. PLoS One. **13**(6), e0196878–e0196820 (2018). https://doi.org/10.1371/journal.pone.0196878

Chapter 9
Engineering Minds for Biologists

Alfredo Benso, Stefano Di Carlo, and Gianfranco Politano

9.1 Why Bioinformatics Needs an Engineering Mind

In the last decades, the advances in data gathering technologies, from high-throughput sequencing to "omics" analyses, has changed the approach to biology. From an engineering point of view, biological research is a "reverse-engineering problem", where scientists try to unravel the mechanisms that allow and support life in living organisms. Traditionally, biology was approached using a bottom-up approach, where each individual actor (molecules, cells, organs) was individually studied, with the hope of later understanding the functioning of the whole biological system by "assembling" the functionalities of its individual components. Unfortunately, the data gathered in the last decades demonstrated that this is not possible because biological systems are complex systems. Mathematics tells us that to understand a complex system we must understand the relations between the parts. The system as a whole determines how the parts behave. Doing otherwise would be like trying to understand how a flock of birds move by individually studying each bird (Fig. 9.1). There are two other very important characteristics of complex systems that we need to mention. The first is that their behavior heavily depends on their initial conditions, and this is incredibly important if we think that inside biological systems a lot of random events (e.g., the proximity of two molecules or cells) can drastically change these "initial conditions" from one moment to the next. The second, directly related to the first, is that the study of the systems "dynamics" (how the system behaves in time) cannot be overlooked. As a simple example, imagine that we want to observe the behavior, in time, of a certain biological reaction (e.g., a cell cycle). As with every dynamic, this reaction will be regulated by a time frequency. At each step, the reaction will take some time to pass from one state to the

A. Benso (✉) · S. Di Carlo · G. Politano
Control and Computer Engineering Department, Politecnico di Torino, Turin, Italy
e-mail: alfredo.benso@polito.it

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_9

Fig. 9.1 A flock of birds shows patterns and behaviors that emerge from the interactions among the birds, and that could not be seen nor understood by studying each bird individually

next. The number of state-changes over a unit of time determines the system's "frequency". Now let's suppose that our technology allows us to take a "snapshot" of the reaction every second. After collecting our reaction timeline, we then want to understand, reconstruct, the whole reaction by studying the individual snapshots. Is this possible? Unfortunately, the answer is not "yes" every time. The Nyquist-law states that a signal can be reconstructed only if we sample it with a frequency that is at least double the system's own frequency. So, in our example, the data would allow us to reconstruct the reaction only if its frequency were of at least 2 s. In any other case, the data we gathered would be completely useless and any deduction obtained from it flawed.

The "engineering mind" is, in our view, the tool to overcome these methodological flaws and approach biological research from a much more efficient and promising angle.

Systems biology, a research methodology (not only a discipline) based on the cross-fertilization between biology, physics, computer science, mathematics, chemistry, and engineering, emerged as the most promising tool for biological research. It is not by chance that biological networks became a "hit" in 1998 when a physicist (Barabasi) and a biologist (Oltvai) became neighbors. At the time, Barabasi was studying the structure of the internet and had already shown that the internet was a nonrandom network and that its connectivity structure was influencing its functions. One year later, in 1999, together they proved that the metabolic pathways of yeast define a network whose structure is very similar to that of the internet. Starting from this discovery, the step to demonstrate that recurrent topological structures of

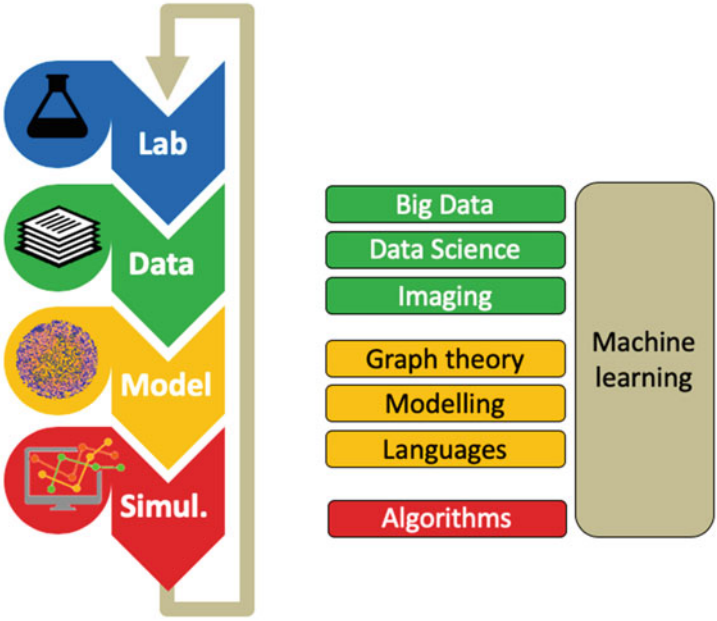


Fig. 9.2 The Systems Biology pipeline is built on a feedback-loop, where Computer modeling and simulations are continuously used to improve Experimental data and protocols which, in turn, allow to create more and more refined models

biological networks were clearly linked to biological functions took little time (Barabasi and Oltvai 2004; Alon 2003; Emmert-Streib and Glazko 2011; Kaplan et al. 2008; Konagurthu and Lesk 2008). The "subliminal" message here is that communication, not only collaboration, is a fundamental aspect of being a bioinformatician. Being at the frontier between two incredibly different communities requires learning to speak (fluently) both "languages" and to become an extremely patient mediator.

The typical systems biology pipeline is sketched in Fig. 9.2. It is made of three main stages: lab experiments, data gathering, and modeling and simulation. Nevertheless, its most important characteristic is the continuous feedback-loop from computational results back to the lab. This loop creates a situation in which labs produce data used to optimize the model of the biological system under study and make predictions that drive the next set of experiments. Each of the three phases of the systems biology pipeline poses, for different reasons, important "engineering" challenges. Experiments are expensive and time-consuming and need to be driven by plausible hypotheses. Despite that, biological data will still be extremely noisy, unreliable, biased by many factors like lab-equipment, operators, and so on. Moreover, labs are not the only data source; a lot of information is also gathered and integrated from publicly available databases. The problem is that the life science research community is still understanding the need for solid formalisms for data

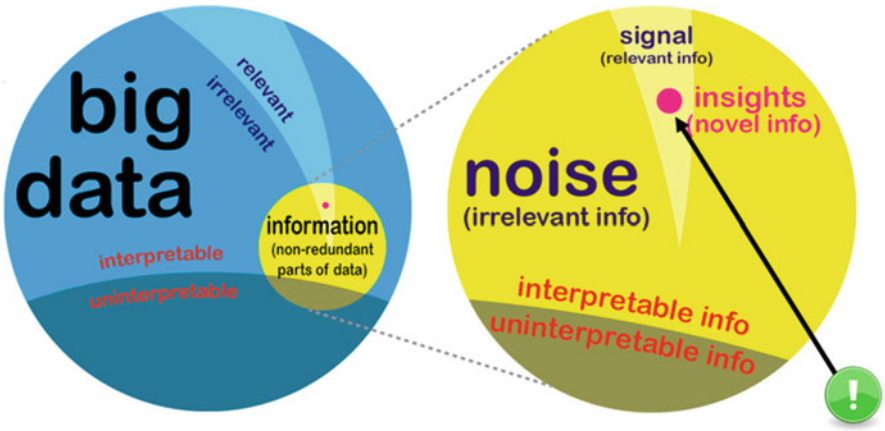


Fig 9.3 Interesting and useful insights are usually deeply buried into vast amounts of irrelevant, uninterpretable, and noisy information, which in turn constitutes a small percentage of the available data

storage, handling, and exchange. The result is that a huge percentage of available data is not reliably reusable outside of the scope in which it was generated, and is therefore useless. Even if the typical systems biology pipeline intrinsically partially mitigates these problems (thanks to the feedback loop), a lot of work still needs to be done.

Obviously, gathering data is not enough if we are then not able to extract information and new knowledge from it (see Fig. 9.3). Therefore, data has to be used to build more and more refined models, which then can be analyzed and simulated to both validate the model itself, and to come up with new hypotheses. Simulation results are the feedback loop to the lab, where they need to be used to drive the experiments necessary to validate the hypotheses.

Networks (or "graphs" in engineering terms) are very simple yet powerful models to represent interactions among systems' components, and for this reason, they emerged as one of the most used formalisms used to models biological systems (protein-to-protein interactions, metabolic reactions, gene-regulatory mechanisms, bacterial populations, etc.). But networks alone are not enough.

It is obvious that each step in the systems biology pipeline requires expertise from different disciplines. This is why we advocate the need for an "engineering mind": to train researchers able to understand, coordinate, and address all these problems.

9.2 The Importance of Data Handling

Often, the work of a pure systems biologist, especially when a wet lab is not available for custom data generation or validation, only relies on data available online. It is therefore a mandatory requirement to validate the data on top of which

algorithms are developed. Nowadays, a large amount of publicly available sources of high throughput/dimensionality data is available, which is good news for researchers facing the systems biology field. However, all that simplicity and availability may dangerously hide some often-neglected caveats that should be properly considered in order to guarantee the overall reliability of the entire pipeline/model and its results. Some practical hints: do not just integrate data without any critical thinking, do not simply trust the homepage claims of a high-quality repository, go in-depth, read the related paper (don't forget additional materials), understand the scope, the underlying assumptions, the experimental bias, if any, and don't forget to check the date of the last update, bearing in mind that not always a recently updated data source may be easily integrated with another with less frequent updates; don't ever assume that one update is backward-compatible with the previous one.

Such caveats should be always critically considered before starting downloading data from a new source. Otherwise, there is the risk to incur some hidden and misunderstood errors, very hard to be identified a posteriori.

When accessing a new data source, thus, critical analysis and a proper assessment should be done by verifying a set of main aspects: (1) the scope of the data source, (2) the naming standardization adopted, (3) the data exchange format, and (4) their updates timing.

9.3 Scope

The first thing we should consider when accessing a new data source is its scope. Taking gene regulation as an example, many databases are very specialized in a limited set of interaction types, possibly providing a biased perspective on the broad scenario of regulations that concurrently take place in a biological entity. Let us consider for instance a database containing posttranscriptional (e.g., microRNAs) regulations such as miRTarBase (Chou et al. 2018) or Targetscan (Agarwal et al. 2015). Information provided by those two databases are in the form of punctual miRNA–mRNA interactions, with no reference to a broader context like a signaling cascade or a pathway (Politano et al. 2014, 2016). Similarly, we have detailed information about host genes relations with their co-transcribed intergenic or intragenic miRNAs (in the shape of mRNA-miRNA interaction, see miRiad (Hinske et al. 2010) thus lacking again an integrated scenario comprising gene-gene regulations. Protein–protein interaction (PPI) databases (e.g., Fisingene (Wu et al. 2010), Mentha (Calderone et al. 2013), String (Szklarczyk et al. 2015)) usually provide information about undirected protein interactions, which often reflects into clusters of protein subunits that may cooperate for activation or assembly of a protein complex, lacking any form of causality. And finally, pathway and regulatory databases (like KEGG (Du et al. 2014) or iRefIndex (Razick et al. 2008)), which do provide a representation of directed causal regulations that take place in a regulation cascade. Furthermore, other specialized databases do exist representing

Transcription Factor regulations (Mathelier et al. 2014), Transcription Factor Co-regulation (Schaefer et al. 2011), or ceRNA neighborhoods (Bhattacharya and Cui 2016).

Overall, though all these databases are sources of reliable information, they provide different concepts of relation, with different customs and very specialized approaches. Not a single database, on his own, is capable of providing a holistic perspective on the complex set of regulations that concurrently take place. Therefore, when designing a reliable system-level analysis or simulation, these co-occurring regulations should not be neglected.

9.4 Naming and Standardization

The overall lack of centralized standards and naming conventions often makes it not trivial to cross-match information among different databases (e.g., consider for instance Unichem (Chambers et al. 2013), which reports up to 35 different aliases for each chemical).

Data so far stored in publicly available databases often reflect overlapping and sometimes conflicting conventions adopted in different subcommunities, with different scopes. Unique identifiers are commonly redefined from scratch for every single database and rarely are adopted widely enough in the community to guarantee a reliable cross match. This lack of consensus in the naming convention requires users to check at first the viability to properly translate entity names from a data source to another, to avoid further lookup problems.

9.5 Data Format

Data from the original repositories are often downloadable for further, high-throughput, processing. Although this is a common feature among public databases, things may become complex just considering the large amount of data formats currently available, which often requires a custom parser, and are not always "true" standards.[1] We may encounter standard plaintext formats like csv/tsv or more complex formats like XML, which can be further declined into specific XML formalism, like OWL, Biopax. We may face raw SQL or non-relational DB dumps, along with several other custom structured formats with proprietary

[1] A "standard" is certified by a standard certification authority (like IEEE), not defined or "customized" by each individual research group. There are also "de-facto" standards, not officially recognized but widely used in scientific communities. Nevertheless, this does not to be the case in the Life Science community.

formalisms and naming conventions. This often results in difficult data interoperability and high management complexity.

Given the large amounts of file formats and sometimes the lack of a proper parser to extract custom information, and considering the average dimension of the files to parse (in the range of tens of megabytes to hundreds of gigabytes), this time consuming and error prone step should be considered in advance, to guarantee its correctness and feasibility. Another solution may be to build intermediate data storage containers and subgrouping data, in order to reduce the overall dimensionality thus allowing for faster querying and more immediate record-level access (Politano et al. 2019).

9.6 Timing of Updates

Last but not least, the probably most neglected problem in data integration is related to data versioning. Due to asynchronous updates in data sources, direct linking to databases may rapidly become obsolete and, even more dangerously, a source of erroneous assumptions. This problem is particularly evident when dealing with miRNAs. miRbase, for instance, the current main reference for the naming of microRNAs, during any update (e.g., see change log between releases 20 and 21) cleans up dubious and mis-annotated sequences and reassigns previously used ids. This resulted in 169 hairpins and 353 mature sequences that have changed names (21), thus getting out of synch any work or data collection built before this update. Reassigning miRNA ids means that any previous reference to a given miRNA id, may actually refer to another one according to the new id reassignment. By considering the large amounts of updates during the year and often the lack of evident information regarding data synchronism, a lot of attention should be spent to properly investigate data aggregation timing and synchronism in order to avoid dealing with obsolete, or worse, wrong references.

Keeping track of the consistency of cross-references among different databases is not trivial and must be taken into account every time data from multiple sources must be integrated.

9.7 Modeling for Simulation

After collecting relevant data regarding a target biological phenomenon, one of the most important goals in systems biology is to build models able to simulate the dynamics of the target system in order to further understand its mechanics and its biological details. Simulations often aim at predicting in silico the outcome of an experiment (e.g., administration of a particular drug or molecule) and, once promising results are identified, predictions can be tested by wet lab experiments (Loewe and Hillston 2008).

How these models are constructed and how predictions are computed strongly depends on the target problem but, often, on the researcher's background. With the increasing availability of computing power, a simplistic solution to this problem could be to capture everything that is known about a system and simulate it in supercomputers. While this could be feasible for specific problems, it is often complex and infeasible due to the complexity of biological systems, which often demand simplifications to make them amenable to modeling. However, simplifications are a dangerous instrument, they have to capture the essence of the processes of interest while neglecting the less important details all taking into account the capability of the formalism and computing platform available for the simulation.

Systems biology models must be able to handle different scales of representation, to model the system and its sub-parts into a complex hierarchical structure, and to handle various types of information represented with different formalisms. Among the different categories of models used in engineering, multilevel computational models are a powerful instrument to handle complexity. Informally, multilevel computational models describe a system considering at least two hierarchical levels, with interactions that take place within and between these levels (Degenring et al. 2004). In the last decade, several methods have been proposed to properly represent and simulate complex biological systems using multilevel computational models (Bardini et al. 2017a).

The choice of the best modeling and simulation approach is not trivial. However, we believe that a set of six relevant key performance indicators (KPIs) can be used to fairly compare different modeling approaches:

1. Scalability with system's complexity
2. Readability and visualization of the simulation results
3. Discrete simulation versus model solving
4. Model constructability starting from other formalisms devoted to data, representation
5. Modeling complexity
6. Complexity management features

Taking these six KPIs in mind we can have a look at common approaches that have been largely applied in the literature to model and simulate complex biological systems:

- Ordinary differential equations (ODE)
- π-calculus
- Rule-based languages
- Agent-based models
- Petri nets

The ODEs are among the most exploited approaches to multilevel modeling and simulation in systems biology (Jasim Mohammed et al. 2017; Suzuki et al. 2009), although they are a powerful mathematical model that can describe how different quantities (e.g., concentration of specific molecules) continuously changes in time. The ODEs fit stoichiometry-based chemical problems in which relations and

gradients are well known, however, their complexity may explode when applied to contexts with a 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.

π-calculus is an attempt to cope with the major drawback of ODE (i.e., computational complexity) (Regev et al. 2001). π-calculus is very well suited to model concurrency, communication, and stochasticity that are all important features of cellular systems. However, the simplification introduced to handle the complexity of the simulation limits its applicability whenever the target system shows complex dependencies among the different parts and simultaneous exchange of information.

While ODE and π-calculus are significant examples of approaches derived from mathematical theories, other approaches find their foundation in the software engineering and computing simulation fields.

Rule-based languages are such an example of approaches that have been used to create multilevel models of complex biological systems (Maus et al. 2011). These languages offer a very compact multilevel representation of a complex system making the modeling effort easy and allowing step by step simulation. Moreover, they are easy to derive from other formalisms. However, 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.

Agent-based systems are another interesting simulation approach that particularly focuses on the simulation problem and its complexity and scalability. They offer a high degree of detail about the agent functions, which correspond to the rules governing the underlying biological interactions. Moreover, they allow for computational parallelization, thus scaling up to very complex systems. Some models (e.g., Repast (North et al. 2006) and CompuCell3D (Fortuna et al. 2020)) offer graphical user interfaces, but these unfortunately, lack significant features of relevance for biologists (Gorochowski 2016).

When looking at the proposed examples it seems that mathematical models and software engineering models both fail at providing all major KPIs required to answer the systems biology requirements.

When facing multilevel hybrid modeling, Petri Nets and their improvements formulations (i.e., Nets-Within-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

Table 9.1 Main topics in a bioinformatic/system biology curriculum

Topic	Prerequisites	Cannot do without
Graphs models	Math logic/differential equations	Graph theory Graph algorithms Modeling and simulation
Programming	Foundations of math Programming logic	Python Modeling and simulation Basics of algorithm complexity
Big data	Relational database design	Machine learning Deep learning basics Graph databases

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. An 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 2003). NWNs implicitly enable observing a system in a zoom-in/zoom-out fashion. NWNs can be used to model properties such as process synchronization, asynchronous events, concurrent operations, and conflicts or resource sharing. These properties characterize discrete- event systems and look promisingly coping with the synthetic biology complexity (Bardini et al. 2017b).

9.8 Conclusions

Systems biology is, in our view, the most promising methodological approach to study biological systems which, given their complex nature, cannot be thoroughly understood using a traditional bottom-up approach. It requires cross-fertilization between biology, physics, computer science, mathematics, chemistry, and engineering, but, most of all, it requires an "engineering mind" able to understand, coordinate, and exploit this challenging mix of competences. In this chapter, while being well conscious of the extreme complexity of biological systems, we supported our claim by discussing many practical issues that appear in the systems biology pipeline, and that are well known (and in several cases have been already addressed and solved) in other engineering areas. To conclude, we summarize which, in our opinion, are the main topics to be included in a bioinformatic/system biology curriculum (Table 9.1).

References

Agarwal, V., Bell, G.W., Nam, J.-W., et al.: Predicting effective microRNA target sites in mammalian mRNAs. Elife. **4**, e05005 (2015)

Alon, U.: Biological networks: the tinkerer as an engineer. Science. **301**(5641), 1866–1867 (2003)

Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**(2), 101–113 (2004)

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

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

Bhattacharya, A., Cui, Y.: SomamiR 2.0: a database of cancer somatic mutations altering microRNA-ceRNA interactions. Nucleic Acids Res. **44**(D1), D1005–D1010 (2016)

Bonzanni, N., Feenstra, K.A., Fokkink, W., et al.: Petri Nets are a Biologist's Best Friend. Springer, Cham (2014)

Calderone, A., Castagnoli, L., Cesareni, G.: Mentha: a resource for browsing integrated protein-interaction networks. Nat. Methods. **10**(8), 690–691 (2013)

Chambers, J., Davies, M., Gaulton, A., et al.: UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform. **5**(1), 3 (2013)

Chou, C.-H., Shrestha, S., Yang, C.-D., et al.: miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. **46**(D1), D296–D302 (2018)

Degenring, D., Rohl, M., Uhrmacher, A.M.: Discrete event, multi-level simulation of metabolite channeling. Biosystems. **75**(1–3), 29–41 (2004)

Du, J., Yuan, Z., Ma, Z., et al.: KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a PATH analysis model. Mol. BioSyst. **10**(9), 2441–2447 (2014)

Emmert-Streib, F., Glazko, G.V.: Pathway analysis of expression data: deciphering functional building blocks of complex diseases. PLoS Comput. Biol. **7**(5), e1002053 (2011)

Fortuna, I., Perrone, G.C., Krug, M.S., et al.: CompuCell3D simulations reproduce mesenchymal cell migration on flat substrates. Biophys. J. **118**, 2801 (2020)

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

Heiner, M., Gilbert, D.: How Might Petri Nets Enhance your Systems Biology Toolkit. Springer Berlin Heidelberg, Berlin (2011)

Heiner, M., Gilbert, D., Donaldson, R.: Petri Nets for Systems and Synthetic Biology. Springer Berlin Heidelberg, Berlin (2008)

Hinske, L.C.G., Galante, P.A.F., Kuo, W.P., et al.: A potential role for intragenic miRNAs on their hosts' interactome. BMC Genomics. **11**, 533 (2010)

Jasim Mohammed, M., Ibrahim, R.W., Ahmad, M.Z.: Periodicity computation of generalized mathematical biology problems involving delay differential equations. Saudi J. Biol. Sci. **24**(3), 737–740 (2017)

Kaplan, S., Bren, A., Dekel, E., et al.: The incoherent feed-forward loop can generate non-monotonic input functions for genes. Mol. Syst. Biol. **4**, 203 (2008)

Konagurthu, A.S., Lesk, A.M.: Single and multiple input modules in regulatory networks. Proteins. **73**(2), 320–324 (2008)

Loewe, L., Hillston, J.: Computational models in systems biology. Genome Biol. **9**(12), 328 (2008)

Mathelier, A., Zhao, X., Zhang, A.W., et al.: JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. **42**(Database issue), D142–D147 (2014)

Maus, C., Rybacki, S., Uhrmacher, A.M.: Rule-based multi-level modeling of cell biological systems. BMC Syst. Biol. **5**, 166 (2011)

North, M.J., Collier, N.T., Vos, J.R.: Experiences creating three implementations of the repast agent modeling toolkit. ACM Trans. Model. Comput. Simul. **16**(1), 1–25 (2006)

Politano, G., Benso, A., Savino, A., et al.: ReNE: a cytoscape plugin for regulatory network enhancement. PLoS One. **9**(12), e115585 (2014)

Politano, G., Orso, F., Raimo, M., et al.: CyTRANSFINDER: a Cytoscape 3.3 plugin for three-component (TF, gene, miRNA) signal transduction pathway construction. BMC Bioinformatics. **17**, 157 (2016)

Politano, G., Di Carlo, S., Benso, A.: 'One DB to rule them all'-the RING: a Regulatory INteraction Graph combining TFs, genes/proteins, SNPs, diseases and drugs. Database (Oxford). **2019**, baz108 (2019)

Razick, S., Magklaras, G., Donaldson, I.M.: iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics. **9**, 405–405 (2008)

Regev, A., Silverman, W., Shapiro, E.: Representation and simulation of biochemical processes using the pi-calculus process algebra. In Pac Symp Biocomput. p. 459–470 (2001)

Schaefer, U., Schmeier, S., Bajic, V.B.: TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. Nucleic Acids Res. **39**(Database issue), D106–D110 (2011)

Suzuki, Y., Asai, Y., Oka, H., et al.: A platform for in silico modeling of physiological systems III. Conf. Proc. IEEE Eng. Med. Biol. Soc. **2009**, 2803–2806 (2009)

Szklarczyk, D., Franceschini, A., Wyder, S., et al.: STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. **43**(Database issue), D447–D452 (2015)

Valk, R.: Object petri nets: using the nets-within-nets paradigm. In Lectures on Concurrency and Petri Nets (2003)

Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. Genome Biol. **11**(5), R53 (2010)

Chapter 10
Design Bioinformatics Curriculum Guidelines: Perspectives

Qanita Bani Baker and Maryam S. Nuser

10.1 Introduction

Bioinformatics is a highly dynamic, demanding, and evolving interdisciplinary field that requires a good level of understanding of life sciences and good computational skills. There are critical needs for bioinformatics for several users and researchers in several fields in biology, biomedical, computer science, and medicine. While offering training and educational programs in bioinformatics are so crucial for several educational institutes all around the world, there are still many challenges and research that are going to meet the educational demand for this dynamic field of study.

Learning methods and practices that contribute to bioinformatics courses training success is critical (Emery and Morgan 2017). Several recent studies such as in (Mulder et al. 2018; Attwood et al. 2019) have identified the big needs to define and enhance the bioinformatics curricula and their contents to be taught in a suitable pedagogical environment. Many works provided a range of approaches for bioinformatics training courses with several focuses. Emery and Morgan (2017) explored the application of project-based learning, in a 5-day training course, as a part of bioinformatics summer school. While, for example, in (Gurwitz et al. 2017), Gurwitz et al. provided sustainable approaches to develop the bioinformatics domain in Africa by offering an introductory course in bioinformatics that includes several fundamental bioinformatics topics as the introductory level course.

Q. Bani Baker (✉)
Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan
e-mail: qmbanibaker@just.edu.jo

M. S. Nuser
Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan

Department of Information Systems, Yarmouk University, Irbid, Jordan

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_10

Many bioinformatics training programs have been established to address the increasing needs of computational and quantitative learning in bioinformatics as shown in (Cohen 2003; Schneider et al. 2012; Atwood et al. 2015). Despite efforts in enhancing and developing these training programs and degrees, the adoption of them is usually limited either to a small and limited number of institutions or to specific courses within a given curriculum. Moreover, the established courses were primarily directed to bioinformatics majors and no other related majors such as life sciences or computer science as presented in several studies (Schneider et al. 2012; Atwood et al. 2015; Wilson Sayres et al. 2018).

Unfortunately, these courses are either short, or/and concentrate on one subject, or are provided as self-learning with no support (Faria et al. 2018). ELIXIR is another example of a distributed sustainable infrastructure for collecting and maintaining biological data in European countries. It provides comprehensive training programs in bioinformatics and computational biology for professionals (Crosswell and Thornton 2012). With the increasing amount of training and teaching information distributed across several locations, a new platform called the TeSS that collects geographically dispersed information is developed by ELIXIR. It presents information in a central portal that helps researchers to find training opportunities that are relevant to their demands (Sansone et al. 2020). Moreover, efforts have been made to create networks to improve education and training in bioinformatics such as the African Bioinformatics Network for the Human Heredity and Health in Africa (H3ABioNet), Global Organisation for Bioinformatics Learning, Education, & Training (GOBLET), and Bioinformatics Club for Experimenting Scientists (Bioclues).

In this chapter, we aim to report the current main perspectives in designing and developing bioinformatics curriculum guidelines. These perspectives are based on comprehensive investigations of the previous and most recent studies related to bioinformatics to discuss the learning outcomes to take into consideration when designing the bioinformatics curriculum and discuss them based on the educational level and the program specialization. Moreover, in this chapter, the main issues that currently face the development and implementation of bioinformatics training programs are presented. Finally, the main recommendation that could help researchers and educators to enhance offering bioinformatics curriculums and planned courses are reported.

The remainder of the chapter includes the following sections: in Sect. 10.2, we discuss curricular goals and learning outcomes. Then, in Sect. 10.3, we present the targeted educational level for the developed programs and their focuses, and in Sect. 10.4, we explain the works done for several targeted specializations based on demands. After that, in Sect. 10.5, we discuss designing and running courses in bioinformatics. Section 10.6 presents the bioinformatics courses' designing and offering issues. Following that, Sect. 10.7 shows the recommendations. Finally, we present the chapter conclusion in Sect. 10.8.

10.2 Curricular Goals and Learning Outcomes

In this section, we summarize the basic conceptual framework used to provide an educational program that can meet diverse target audience populations. This framework depends on competencies and the learning outcomes (LOs) expected from the educational frameworks. As shown in (Mulder et al. 2018), demonstrating the competencies is used to drive the bioinformatics curriculum development. Due to the diversity of the potential training audience, there are different needs in terms of required skills or knowledge and at which depth. In studies (Mulder et al. 2018; Welch et al. 2016), three major user profiles were considered which are: the bioinformatics engineer, the bioinformatics user, and the bioinformatics scientist. ISCB Curriculum Task Force constructed a set of competencies that can be used to develop a curriculum of bioinformatics training programs.

In several works such as (Machluf et al. 2017; Welch et al. 2014), the main process is shown in Fig. 10.1, it is suggested to be used to build several case studies. This is a part of continuing works to refine and update the building curricular guidelines to serve and assist the bioinformatics community. In the first step, surveys of educational and training programs' needs are refined through a multiyear process of community engagement. This includes surveying and asking questions to participants to identify their computational skills levels and bioinformatics knowledge and background. This community engagement will be the data-driven source to refine the required competencies needed based on the audience's needs and background. Then, based on potential recipients of the training program's needs, the competencies need to be refined for each set of personas. The competency framework needs to be developed using an iterative process with input from several parties from varieties of backgrounds related to bioinformatics. After that, Bloom's Taxonomy is used in mapping competency levels for each of the user groups, with Bloom's Taxonomy terms that include: knowledge, comprehension, analysis, synthesis, application, and

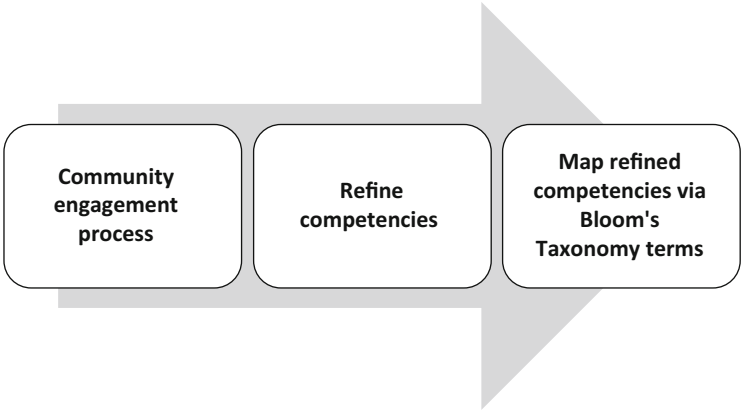


Fig. 10.1 Setting up the learning outcomes based on competencies based on the current studies

evaluation. The curriculum task force of the "International Society of Computational Biology" identified a set of 16 essential competencies setup using an iterative process of inputs and surveys from learners with interest in bioinformatics and computational biology domain (Mulder et al. 2018).

For training excellence and as shown by (Via et al. 2013), providing up-to-data and high quality for satisfying the expectations of potential learning audience. Several studies confirmed that LOs should be formulated based on competencies and using Bloom's Taxonomy Illustrative Verbs terms such as "analyze," "reproduce," "apply," "design," "predict," "develop," "compare," rather than "know." This gives direct action into practical exercises and tasks, which represent essential and core tools to achieve defined LOs.

10.3 Targeted Educational Level

With the growing demand for bioinformaticians, there is a persistent and continuing need for bioinformatics education at all levels starting with students at schools going through undergraduate students at universities to postgraduate students and might continue as a form of training for postdoctorate and bioinformaticians. In this section, we present and discuss the current efforts that cover several educational levels and we divide them into three levels as shown in the following Sects 10.3.1, 10.3.2, and 10.3.3.

10.3.1 High Schools Level

It is important to integrate bioinformatics into high school classrooms. This integration enhances the teaching of concepts of evolution, human biology, molecular biology, and genetics (Form and Lewitter 2011). Students can be trained in using bioinformatics tools and on approaching real-world problems. Final year students might be able to develop tools that they might continue using during college and beyond. Moreover, this will introduce career awareness to students (Form and Lewitter 2011).

Bioinformatics@-school (Marques et al. 2014) is one example of a project that serves the life sciences curriculum that uses bioinformatics activities in high schools. The project helps teachers incorporate and use the latest advances in science into their teaching. It targets both students and teachers. Bioinformatics@-school project includes a teachers-training program where teachers follow the same activities given to students, with the help and support of a teacher manual and under the supervision of qualified bioinformaticians. The main goal of the training program is to enable teachers to guide students in the required activities and to understand the basics of the bioinformatics methods, tools, and resources underlying given activities. On the other hand, bioinformatics@-school includes web-based research tasks that students

can practice alone or under teacher supervision. Also, the project enables the discussion of key results between students and teachers (Marques et al. 2014). Another study (Barker et al. 2015) supports the practicality of bringing university-level, bioinformatics activities, and exercises to school students. A measurement survey provides proof of an increase in awareness of the importance of computers within biology. It also shows students' ability to use computers in analyzing DNA sequences.

10.3.2 Undergraduate Level

There are a great number of undergraduate students who do not have the chance to learn bioinformatics or/and computational biology skills structured in their educational curriculum. Undergraduate students in life sciences and biology, in many cases, need to analyze big data; they rely on bioinformaticians to help them do such tasks because bioinformatics education is not well integrated at the undergraduate level for these majors (Zhan et al. 2019). Therefore, a need arises to educate university students of these majors to analyze data by themselves through integrating bioinformatics in these majors.

Barriers to the integration of the bioinformatics domain into life sciences education were reported in (Williams et al. 2019). The most frequently reported barrier was that most current life science professors don't have the required bioinformatics analysis skills. Other issues included were the big lack of students who have an interest in bioinformatics study and research; the weakness in student preparation in computer science, mathematics, and statistics; limited access to required resources, vetted teaching materials, and overly full curriculum. One way to face some of these barriers is by implementing teaching modules that can assist both teachers and students in this area. The University of Puget Sound, Washington State (US) implemented such a module. The module was incorporated into undergraduate biology majors, who have little or no knowledge of computer science, and programming. The module makes achievements in building students' skills in basic command-line computing and bioinformatics. It motivates students to investigate more in the bioinformatics field by approaching real-world biological problems and to appreciate the use of bioinformatics in modern biology (Madlung 2018).

Another implementation of such modules was a short course that was developed in some colleges and regional universities in the United States to train professors with essential bioinformatics skills that were later adopted by the professors in their classes (Zhan et al. 2019). At Lancaster University, they integrate bioinformatics skills into undergraduate biology degree portfolio (Gatherer 2020).

10.3.3 Postgraduate Level

To identify the required qualifications suitable for employers in the bioinformatics field, a study by (Shang and Ghriga 2019) indicated that of all the 38 jobs analyzed in their research, a master's degree is required or preferred. Most graduate programs, and especially bioinformatics as an interdisciplinary field, have students from various backgrounds with undergraduate degrees from a variety of majors. Therefore, special attention needs to be taken in designing such curricula (Huanmei et al. 2016).

Some courses were introduced and served both undergraduate and graduate students (Zhan et al. 2019) while others targeted only postgraduate students (Guerfali et al. 2019) such as the advanced course in bioinformatics and genome analyses offered in the Institut Pasteur de Tunis. Consequently, to spread the chance to learn bioinformatics or computational biology skills, a well-designed curriculum should start with students from schools going to undergrad level and through postgrad levels.

10.4 Targeted Specialization Needs

Bioinformatics is an interdisciplinary field of study. Bioinformatics courses can be offered under several departments based on the content of the course and the background of the students. Life scientists, for example, need bioinformatics to help them analyze big datasets. This requires courses that teach students how to deal with computer programs starting from the formatting and parsing data files going through writing computer scripts and programs that can connect existing software applications. Also, training on using high-power computer clusters needs to be prioritized (Madlung 2018).

As discussed in (Khuri et al. 2020), several bioinformatics topics are usually incorporated with undergraduate computer science courses such as studies in (Cohen 2005; Unay et al. 2010). Also, many computer science departments offer several undergraduate and graduate bioinformatics as electives, and/or specializations such as (Ericson et al. 2014; Fetrow and John 2006; Khuri 2008; Ritz 2018). Computer scientists need to have introductory information (course) on biological data. They need to have the skills to work with and extract requirements from biologists and life scientists who usually don't have the required enough programming skills. A suggested two course in concert was presented in (Goodman and Dekhtyar 2014). The courses focused on teamwork where collaboration from both CS students and biology students is required to solve the problems.

Medical students need bioinformatics essentials to succeed in medical research. For example, at the Washington University School of Medicine (WUSM), a module for teaching the bioinformatics essentials to first-year students is developed. The module utilizes clinical cases as a platform to access information stored in Online

Mendelian Inheritance in Man (OMIM), GenBank, and PubMed databases at NCBI. The module proved its efficiency in introducing medical students to well-known genetic databases. The developed module successfully integrated the basic principles of molecular and clinical genetics using real clinical cases (Magee et al. 2001). Later, a joint course was developed between the University of Washington and the Universidad Peruana Cayetano Heredia. It served a new bioinformatics track along with an established Medical/Public Health Informatics track for participants in Peru. Assessment results showed the success of this course (Magee et al. 2001).

Genetics is another discipline that uses bioinformatics tools. Several universities realized the importance of bioinformatics use in their departments and included their needs for that field in their curriculum. At Clark Atlanta University, they designed and implemented a new bioinformatics component to supplement the undergraduate genetics course they have and which the students appreciated (David Holtzclaw et al. 2006). At Centenary College of Louisiana, they developed a project for the genetics course, which is a research-based laboratory course, where students were introduced to the use of bioinformatics methods and tools, and genetic and molecular biology approaches (Brame et al. 2008).

10.5 Designing and Running Course in Bioinformatics

Several courses have been designed and run to address the many gaps in bioinformatics education. Some of these courses are focusing on project-based (Emery and Morgan 2017; Achappa et al. 2020), or with problem-based focus as (Ling 2017), or it can be a combination of both (Saarunya and Ely 2018). Courses can be provided as workshops or as a semester-long course.

As shown in Fig. 10.2, we summarize the basic steps that can be used to develop courses in bioinformatics fields as recommended by several studies as in (Saarunya and Ely 2018). The first step in designing the courses is to define the needs of students based on their working domain. Then, the course's participants have to be defined including the number of participants, participants' educational level, and participants' specializations. As explained in the above section. The needed competencies have to be established through surveys and in an iterative manner of inputs from participants, and the LOs have to be setup based on the defined competencies. Based on the responses of the participants, the learners can be assigned to three groups as recommended by (Mulder et al. 2018) and as done by many studies (Saarunya and Ely 2018). Then, a developed course structure needs to be determined and, in this step, the syllabuses and the type of courses should be clearly defined. Identifying a course's types includes a clear stating: if the course is a workshop or

Fig. 10.2 Summary of the basic steps that can be used to develop courses in bioinformatics fields

long semester course? Is it one semester or multi-semesters? Is it a problem-based or project-based or both? Using multi-projects or a single project? Identifying the practical part?

Some studies suggested having a small class, especially if the course has different levels of students (Saarunya and Ely 2018). Small size class makes it able to work with the students individually and to help them to achieve the learning outcomes successfully. Conducting surveys includes a pre-survey that needs to be completed at the very beginning of the training educational courses and a post-survey that should be completed at the end of the course. Pre-survey could have some questions about student demographics as shown in (Madlung 2018). Many research can be used to help in building these surveys such as (Mendez et al. 2016). These surveys can include questions asking the participant to identify their computational skills and bioinformatics knowledge. By comparing these skills' levels, which are reported by participants in pre-survey and post-survey, the improvements can be measured after completing a course.

10.6 Courses Design and Offering Issues

In designing and offering bioinformatics courses, many issues and challenges are expected. Different studies have provided use cases for the training experiences and showed many learned lessons and discuss many of these challenges. In this section, we discuss and report the main challenges that can be taken into account when offering and designing bioinformatics educational courses.

One of the major challenges that are expected is the heterogeneity of the backgrounds and/or skills of the course participants. Many of the provided courses have several levels of participants, for example, undergraduate and graduate levels, or sometimes from different specialization, for example, with biology background or computer science backgrounds. This causes different participants' needs in terms of knowledge or skills they require and at which depth. Many studies (Mulder et al. 2018) recommended to cluster the participant into groups and subgroups and identify the competencies for each group. Other studies suggest providing a small size class so it will be easy to make individual attention to courses' participants.

Another challenge that was also reported in the previous studies (Ahmed et al. 2018) is offering bioinformatics training in a limited-resource setting (Ahmed et al. 2018), Awadallah et al. reported two types of resource-limited settings which are location and timing. To solve these challenges the researchers in (Ahmed et al. 2018) suggested that the location challenge can be solved by working more closely with the universities to gain more support in allocating more resources and infrastructure. Also, they suggested a more active collaboration with government agencies and entities like ministries of higher education that will provide more sustained training.

Another important issue mentioned is the lack of expertise in the bioinformatics domain. Finding enough expertise is crucial to develop new biocomputing-intensive training programs (Gurwitz et al. 2017; Madlung 2018). This is still a challenge since

many faculties received their training before the rapid expansion and evolution of big and high-throughput bioinformatics data and research methods.

10.7 Recommendation

Designing a bioinformatics course should adhere to the highest educational quality standards that aim to empower students with the necessary knowledge, skills, and competencies that qualify them for future job positions, as well as in continuing their study in the field of bioinformatics. To face many of the above challenges and issues, some design issues need to be considered when designing bioinformatics courses (Gurwitz et al. 2017). The following are some recommendation that we think could help to offer a better bioinformatics training program and curriculum:

- Appropriate infrastructure is needed, especially if distance learning is an option. Therefore, if a developing country is an issue an appropriate university needs to be chosen as a starting place (Ahmed et al. 2018).
- The fundamentals of bioinformatics should be started at the school level. This requires training of academic staff members on bioinformatics-related courses, software packages, and tools and hardware components that are necessary to build and deploy various algorithms, techniques, and applications that serve the needs of different stakeholders across multiple bioinformatics domains. This training should include academic members of both schools and universities.
- Subjects of the course content should go in line with the learning outcomes. Participants' backgrounds determine the depth of the subjects that should be taught.
- Real-world up-to-date examples that in the long run will meet the market requirements should be a focus.
- It is extremely important to organize the contents of the individual course or the content of successive courses in a logical order to attain a high level of cohesion and complementarity. This can be accomplished by reusing datasets or building on a previous project that the student used in a previous course (Faria et al. 2018).
- Balancing between theory and practice materials and exercises. The benefit from most theoretical courses can be gained when applying practically what the students have learned. Therefore, this point should be an important guideline for bioinformatics courses where applications are mostly used (Faria et al. 2018).
- Lab exercises should be introduced in order to enforce interaction and communication between students across and within classrooms. This will teach students how to collaborate with others especially if they are of different backgrounds, which is a need in such a multidisciplinary field.
- Redundancy in course delivery methods is desired. A course might be given in a real classroom in addition to a virtual classroom so that the student can access the material in a way that is more appropriate to him/her (Gurwitz et al. 2017).
- The course should include real-world examples not artificially produced ones.

- For postgraduate courses, the scope of the suggested bioinformatics courses should be expanded in response to the emerging research to include applications of several related areas (Welch et al. 2016)
- The course should be evolving constantly as different kinds of real-world problems are introduced (Welch et al. 2016)
- Such cross-disciplinary courses should be well-formed to attract students of several majors.

10.8 Conclusions

There are growing needs to prepare bioinformatics professionals to meet the biomedical revolution demand. Several bioinformatics training programs have been established in order to address these increasing demands in computational and quantitative learning in bioinformatics. Comprehensive investigations of the previous and most recent studies related to bioinformatics training and education topics provide us with many lessons to learn from the experiences. This chapter reports the current main perspectives in designing and developing bioinformatics curriculum guidelines. In this chapter, the main issues that currently face the development and implementation of bioinformatics training programs are presented. The main recommendation that could help researchers to enhance offering bioinformatics curriculums and planned courses are reported. These guidelines represent an aggregate of the recommendations suggested in most existing courses and are fairly flexible to meet the objectives and goals of the teaching educational institutions. More case studies could be provided to describe the current needs to provide more professional degrees especially to deal with the complexity of biomedical big data revolution.

References

Achappa, S., Patil, L.R., Hombalimath, V.S., Shet, A.R.: Implementation of project-based-learning (pbl) approach for bioinformatics laboratory course. J. Eng. Educ. Transform. **33**, 247–252 (2020)

Ahmed, A., Awadallah, A.A., Elmahdi, M.T., Suliman, M.A., Khalil, A.E., Elsafi, H., Hamdelnile, B.D., Abdullateif, M., Fadlelmola, F.M.: Blended bioinformatics training in resource-limited settings: A case study of challenges and opportunities for implementation. BioRxiv, 431361 (2018)

Attwood, T.K., Blackford, S., Brazas, M.D., Davies, A., Schneider, M.V.: A global perspective on evolving bioinformatics and data science training needs. Brief. Bioinform. **20**(2), 398–404 (2019)

Teresa K Atwood, Erik Bongcam-Rudlo, Michelle E Brazas, Manuel Corpas, Pascale Gaudet, Fran Lewitter, Nicola Mulder, Patricia M Palagi, Maria Victoria Schneider, Celia WG van Gelder, et al. Goblet: the global organisation for bioinformatics learning, education and training. PLoS Comput. Biol., 11(4):e1004143, 2015

Barker, D., Alderson, R.G., McDonagh, J.L., Plaisier, H., Comrie, M.M., Duncan, L., Muirhead, G. T.P., Sweeney, S.D.: University level practical activities in bioinformatics benefit voluntary groups of pupils in the last 2 years of school. Int. J. STEM Educ. **2**(1), 17 (2015)

Brame, C.J., Pruitt, W.M., Robinson, L.C.: A molecular genetics laboratory course applying bioinformatics and cell biology in the context of original research. CBE Life Sci. Educ. **7**(4), 410–421 (2008)

Cohen, J.: Guidelines for establishing undergraduate bioinformatics courses. J. Sci. Educ. Technol. **12**(4), 449–456 (2003)

Cohen, J.: Computer science and bioinformatics. Communications of the ACM. **48**(3), 72–78 (2005)

Crosswell, L.C., Thornton, J.M.: Elixir: a distributed infrastructure for European biological data. Trends Biotechnol. **30**(5), 241–242 (2012)

David Holtzclaw, J., Eisen, A., Whitney, E.M., Penumetcha, M., Joseph Hoey, J., Sean Kimbro, K.: Incorporating a new bioinformatics component into genetics at a historically black college: outcomes and lessons. CBE Life Sci. Educ. **5**(1), 52–64 (2006)

Emery, L.R., Morgan, S.L.: The application of project-based learning in bioinformatics training. PLoS Comput. Biol. **13**(8), e1005620 (2017)

Ericson, B.J., Guzdial, M., McKlin, T.: Preparing secondary computer science teachers through an iterative development process. In Proceedings of the 9th Workshop in Primary and Secondary Computing Education, p. 116–119 (2014)

Faria, R., Triant, D., Perdomo-Sabogal, A., Overduin, B., Bleidorn, C., Santana, C.I.B., Langenberger, D., Dall'Olio, G.M., Indrischek, H., Aerts, J., et al.: Introducing evolutionary biologists to the analysis of big data: guidelines to organize extended bioinformatics training courses. Evol. Educ. Outreach. **11**(1), 8 (2018)

Fetrow, J.S., John, D.J.: Bioinformatics and computing curriculum: a new model for interdisciplinary courses. ACM SIGCSE Bulletin. **38**(1), 185–189 (2006)

Form, D., Lewitter, F.: Ten simple rules for teaching bioinformatics at the high school level. PLoS Comput. Biol. **7**(10), e1002243 (2011)

Gatherer, D.: Reflections on integrating bioinformatics into the undergraduate curriculum: the Lancaster experience. Biochem. Mol. Biol. Educ. **48**(2), 118–127 (2020)

Goodman, A.L., Dekhtyar, A.: Teaching bioinformatics in concert. PLoS Comput. Biol. **10**(11), e1003896 (2014)

Guerfali, F.Z., Laouini, D., Boudabous, A., Tekaia, F.: Designing and running an advanced bioinformatics and genome analyses course in Tunisia. PLoS Comput. Biol. **15**(1), e1006373 (2019)

Gurwitz, K.T., Aron, S., Panji, S., Maslamoney, S., Fernandes, P.L., Judge, D.P., Ghouila, A., Entfellner, J.-B.D., Guerfali, F.Z., Saunders, C., et al.: Designing a course model for distance-based online bioinformatics training in Africa: The H3ABioNet experience. PLoS Comput. Biol. **13**(10), e1005715 (2017)

Huanmei, W., Raha, O., Zhang, J.: Customizing bioinformatics graduate programs for diversified student backgrounds. In: 2016 IEEE Frontiers in Education Conference (FIE), pp. 1–7. IEEE (2016)

Khuri, S.: A bioinformatics track in computer science. In Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education, p. 508–512 (2008)

Khuri, N., Lee, W., Virginia Lehmkuhl-Dakhwe, K., VanHoven, M., Khuri, S.: Interdisciplinary minor in bioinformatics: first results and outlook. In: Proceedings of the 51st ACM Technical Symposium on Computer Science Education, pp. 407–412 (2020)

Ling, M.H.T.: Problem-based learning (pbl), an important paradigm for bioinformatics education. MOJ Proteom. Bioinform. **5**(4), 00166 (2017)

Machluf, Y., Gelbart, H., Ben-Dor, S., Yarden, A.: Making authentic science accessible the benefits and challenges of integrating bioinformatics into a high-school science curriculum. Brief. Bioinform. **18**(1), 145–159 (2017)

Madlung, A.: Assessing an effective undergraduate module teaching applied bioinformatics to biology students. PLoS Comput. Biol. **14**(1), e1005872 (2018)

Magee, J., Gordon, J.I., Whelan, A.: Bringing the human genome and the revolution in bioinformatics to the medical school classroom: a case report from Washington university school of medicine. Acad Med. **76**(8), 852–855 (2001)

Marques, I., Almeida, P., Alves, R., Dias, M.J., Godinho, A., Pereira-Leal, J.B.: Bioinformatics projects supporting life-sciences learning in high schools. PLoS Comput. Biol. **10**(1), e1003404 (2014)

Mendez, R.G., Torres, J., Ishwad, P., Nicholas, H.B., Ropelewski, A.: Assisting bioinformatics programs at minority institutions: needs assessment, and lessons learned–a look at an internship program. In Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale, p. 1–8 (2016)

Mulder, N., Schwartz, R., Brazas, M.D., Brooksbank, C., Gaeta, B., Morgan, S.L., Pauley, M.A., Rosenwald, A., Rustici, G., Sierk, M., et al.: The development and application of bioinformatics core competencies to improve bioinformatics training and education. PLoS Comput. Biol. **14** (2), e1005772 (2018)

Ritz, A.: Programming the central dogma: an integrated unit on computer science and molecular biology concepts. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education, p. 239–244 (2018)

Saarunya, G., Ely, B.: Design and implementation of semester long project and problem based bioinformatics course. F1000 Res. **7**(1547), 1547 (2018)

Sansone, S.A., Attwood, T.K., Nenadic, A., Bacall, F., Beard, N., Goble, C.A., Thurston, M.: Tess: a platform for discovering life science training opportunities. Bioinformatics. **36**(10), 3290–3291 (2020)

Schneider, M.V., Walter, P., Blatter, M.-C., Watson, J., Brazas, M.D., Rother, K., Budd, A., Via, A., van Gelder, C.W.G., Jacob, J., et al.: Bioinformatics training network (btn): a community resource for bioinformatics trainers. Brief. Bioinform. **13**(3), 383–389 (2012)

Shang, R., Ghriga, M.: Identifying skill sets for bioinformatics graduate students–a text mining approach. J. Comput. Sci. Coll. **34**(6), 160–162 (2019)

Unay, D., Çataltepe, Z., Aksoy, S.: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): preface, 2010

Via, A., Blicher, T., Bongcam-Rudlof, E., Brazas, M.D., Brooksbank, C., Budd, A., Rivas, J.D.L., Dreyer, J., Fernandes, P.L., van Gelder, C., et al.: Best practices in bioinformatics training for life scientists. Brief. Bioinform. **14**(5), 528–537 (2013)

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., Schneider, M.V.: Bioinformatics curriculum guidelines: toward a definition of core competencies. PLoS Comput. Biol. **10**(3), e1003496 (2014)

Welch, L., Brooksbank, C., Schwartz, R., Morgan, S.L., Gaeta, B., Kilpatrick, A.M., Mietchen, D., Moore, B.L., Mulder, N., Pauley, M., et al.: Applying, evaluating and refining bioinformatics core competencies (an update from the curriculum task force of ISCB's Education Committee). PLoS Comput. Biol. **12**(5), e1004943 (2016)

Williams, J.J., Drew, J.C., Galindo-Gonzalez, S., Robic, S., Dinsdale, E., Morgan, W.R., Triplett, E. W., Burnette III, J.M., Donovan, S.S., Fowlks, E.R., et al.: Barriers to integration of bioinformatics into undergraduate life sciences education: a national study of us life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. PLoS One. **14**(11), e0224288 (2019)

Wilson Sayres, M.A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A.G., Smith, T.M., Triplett, E. W., Williams, J.J., Dinsdale, E., Morgan, W.R., et al.: Bioinformatics core competencies for undergraduate life sciences education. PLoS One. **13**(6), e0196878 (2018)

Zhan, Y.A., Wray, C.G., Namburi, S., Glantz, S.T., Laubenbacher, R., Chuang, J.H.: Fostering bioinformatics education through skill development of professors: big genomic data skills training for professors. PLoS Comput. Biol. **15**(6), e1007026 (2019)

Chapter 11
Machine Learning for Bioinformatics

Harshita Bhargava, Amita Sharma, and Jayaraman K. Valadi

11.1 Introduction

Machine learning algorithms have gathered the attention of every individual with applications in astronomy, online shopping, social media, medical diagnostics, online trading, smart devices, online education, and so on. These algorithms differ from traditional problem-solving algorithms with the ability to learn from the data without being explicitly programmed. These data exist in different formats and are typically generated from a variety of sources including data from primary or secondary research, image, video, location, geospatial and sensory data (Alonso et al. 2017). The availability of such data not only provides opportunities for data-driven decision making through data analytics but also poses serious challenges in terms of data management and storage. With the advent of cloud computing data management and storage issues can be handled with greater ease and flexibility while the data analytics part can be well addressed by the use of machine learning algorithms.

Bioinformatics is an interdisciplinary active research field that requires and derives knowledge from the fields of chemistry, physics, biology mathematics, biotechnology, pharmacology, computer science, and so on (Manisekhar et al. 2020). In recent years bioinformatics data has also grown at a very fast pace along with the availability of open-source databases and low cost, time-saving next generation sequencing or high throughput sequencing techniques. These data can be further analyzed either to develop hypotheses, draw inferences, or generate predictions. These predictions can be in the form of gene sequence prediction,

H. Bhargava · A. Sharma (✉)
Department of Computer Science, IIS University, Jaipur, India
e-mail: amita.1983@iisuniv.ac.in

J. K. Valadi
Department of Computer Science, FLAME University, Pune, India

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8_11

protein function and structure predictions, drug target interaction, drug side effect prediction, and so on.

The students, researchers, and academicians in bioinformatics have equal opportunities to collaborate to study and address the basic problems of drug repositioning, drug development, and drug discovery, and so on. This equality is justified by the availability of open-source databases in bioinformatics including genomic, proteomic, transcriptomic, and metabolomic data that can be directly utilized by the students for machine learning applications. The datasets obtained using these open-source databases can assist the students to develop conventional machine learning or deep learning models and generate outputs while evaluating the same using well-defined metrics as per the studied problem. The learning aspects of the students increase with the interdisciplinary nature of the bioinformatics field and further analysis using machine learning or deep learning approaches. As far as education in bioinformatics is concerned one must be taught in a sequential manner starting from the basic theoretical domain concepts to in-depth analysis and modeling the available data thereby extracting information in the form of classifications or predictions.

11.2 Machine Learning Algorithms for Bioinformatics

Biological systems are too complex and so are the data, hence the conventional approach of wet lab experimentation proves to be a tedious and time-consuming task. The costs involved in this approach can be reduced to a great extent by using machine learning or deep learning based computational models. As an example, if we are given the task of discovering drugs for COVID-19 then this entire drug development task takes years to complete. Here the concept of reusing the existing drugs can be utilized which can greatly speed up the process. The concept of reusing existing drugs also called drug repurposing relies on finding all other targets except the one for which it was originally designed (Masoudi-Sobhanzadeh et al. 2020). These targets can be shortlisted by machine learning methods rapidly. These scanned targets can be further tested using wet lab experiments to ensure their reliability and accuracy.

Machine learning algorithms can be categorized as:

- Supervised learning algorithms: These algorithms use existing domain knowledge to build models. In other words, they take the labeled data from prior annotations as training examples and learn the mapping between the inputs and outputs in order to classify or label the new input examples. Supervised learning tasks can take the form of classification where the output has a class label either binary or some discrete value. It may also take the form of regression where the output has a definite continuous value within a specific range. In the case of regression, the difference between the expected and predicted output specifies the error. The major concern with supervised learning is that while classifying the data one should ensure that the dataset is balanced else the model will learn the majority class and the output would be biased (Kaur et al. 2019).
- Unsupervised learning algorithms: These algorithms are used primarily when the data are neither labeled nor defined as separate classes. It basically learns hidden patterns and valuable domain knowledge using similarities/dissimilarities out of the input data with the help of clustering or density estimation algorithms.
- Semi-supervised learning algorithms: These algorithms are well suited for problems where the labeled/classified data are far less than the unlabeled/unclassified one. This kind of learning is used when labeling the examples is too expensive and time-consuming as compared to utilizing unlabeled examples.
- Reinforcement learning algorithms: These algorithms enable the software agents to learn the actions depending upon the feedback received using a trial and error approach each time the agent interacts with the environment. This enables the agent to learn the optimized states that are associated with the maximization of rewards as per the defined problem. While supervised learning models learn from data, reinforcement learning models learn from experience.

11.3 Deep Learning Algorithms for Bioinformatics

The main disadvantage with the classical machine learning algorithms is that they require a typical feature engineering effort before their application to the respective domain. The output of the machine learning algorithm thereby largely depends upon the features selected for the respective problem. In order to mimic the idea of human perception in arriving at conclusions directly from raw data without learning features explicitly, deep learning was introduced (Zhang et al. 2017). Deep learning is a subset of machine learning and has found applications in various fields including social network data analysis, video streaming services, image classification, speech recognition, precision medicine, recommender systems, drug development, and so on. The voluminous bioinformatics data including genomic, proteomic, and microarray data can be considered as a perfect candidate for training deep learning algorithms.

A significant class of deep learning algorithms includes:

- Deep neural network (DNN): It is an inherited version of the ANN model with multiple hidden layers between the usual input and output layers that learn a hierarchy of concepts directly from the raw inputs. The inputs for the DNN are applied at the input layer and each layer, the input multiplied by a weight vector is calculated and a nonlinear function such as sigmoid or rectified linear unit (RELU) is applied to produce outputs. The outputs are again used as inputs and fed to the subsequent layer to generate the final output.
- Convolutional neural network (CNN): CNN has been a well-known class of deep learning algorithms for analyzing images. In terms of architecture, a CNN consists of a series of overlapping convolution layers, activation layers, pooling layers, and fully connected layers with the appropriate configuration of filters at each convolution layer. The architecture may differ but in general, the learned features become more abstract with each layer. The max pooling or average pooling is used to further reduce the dimension of the input matrix of pixels, say an image as received from the previous convolution layer. The pooling operation ensures the location invariance property suggesting the presence of the feature irrespective of its location in the image.
- Recurrent neural network (RNN): RNN are used to analyze temporal sequences or sequence-based input data either organized as text or gene sequences in the form of DNA, RNA, or protein. They involve the computation of the hidden state for each "entity" in the sequence. The output of the hidden state depends not only on the normal input fed into the network serially from the sequence but also on the previous hidden state.
- Autoencoders: Autoencoders are classified as unsupervised learning, generative models that can learn to generate the outputs that resemble the inputs. The unlabeled data is taken as input and encoded by the encoder into latent, compressed, and low dimensional representations called code or latent space. The decoder takes these low dimensional hidden and compressed representations from the latent space and converts them back into original input data. The reduced dimensional features can be used as input to any supervised learning algorithm. Autoencoders have found applications in image coloring, image regeneration from noisy image data, data compression (image, audio or video), and so on. In bioinformatics, they are used for extracting interpretable factor models and biologically relevant latent spaces.

11.4 Applications of Machine Learning (ML) in Bioinformatics

In bioinformatics machine learning (ML) is frequently used in varied problems. ML has become a strong tool to handle perplexing datasets in genomics. The students can understand applications based on the ML algorithm's task which are mainly classified as:

- Association rule based:

  This category comprises those problems where the association between binary or multiple sets is determined. Such problems are disease symptoms mapping, drug target interaction, mapping of molecule structure with protein reaction mapping, drug discovery from plant compounds, and so on. Association rule based ML generates relationships between the known to known set and unknown to a known set.
- Clustering:

  Clustering algorithms are generally used whenever the prior information about the problem is limited or unavailable. There are a large number of clustering algorithms that provide significant support in developing knowledge for such problems. Problems like determining possible sectors of infections or the presence of microorganisms, gene-based clustering to know the gene expression, feature or sample-based clustering, and so on.
- Dimensionality reduction:

  Dimensionality reduction algorithms are an intrinsic element of bioinformatics problems because of the datasets curated from multiple databases. Problems like gene sequencing, image processing, molecular structure processing, drug protein interaction prediction employ basic dimensionality reduction using algorithms such as principal component analysis, SVD, matrix factorization, and so on.
- Classification and regression:

  In bioinformatics, classification and regression algorithms are commonly used to predict the class or object. There are problems like disease prediction and digital diagnosis, presence and possibility of infection, computational screening of molecular fragments, virtual screening of compounds, computer aided drug design, large-scale protein interactions, protein structure prediction, and many more.

There are some complex problems where ML can be applied like fragment-based de novo design, fragment linking to design novel inhibitors, molecular docking analysis with virtual screening, construction of homology models, designing of the linear discriminant analysis model, and designing of analogs. The career prospects of ML in bioinformatics are increasing day by day. The application of ML to genetic data and neuroimaging data opens new frontiers for novice engineering. It has improved the understanding of complex diseases, genetic transformations, and genetic disorders. With the help of ML different kinds of application software and automated tools have been devised. Precision medicine and recommendation systems are the latest applications of ML. These applications assist medical experts in the treatment of chronic diseases. ML algorithms can easily adapt to new data that is generated each day and also they have the ability to handle noisy and missing biological data.

11.5 Conclusions

In the big data era transformation of huge unstructured data into structured information and to valuable knowledge has been the key challenge. Recent advances in machine learning and deep learning have effectively addressed this with improved and cost-effective handling methods. Recent efforts to handle model interpretability and overfitting issues will pave way for the development of more reliable models. We have described a very brief and lucid introduction to machine learning and its components. The role of machine learning in bioinformatics education is further strengthened by the availability of open-source development packages, libraries, and frameworks. There is however an urgent need for constant revision and upgradation of the course curriculum.

References

Alonso, S.G., de la Torre Diez, I., Rodrigues, J.J., Hamrioui, S., Lopez-Coronado, M.: A systematic review of techniques and sources of big data in the healthcare sector. J. Med. Syst. **41**(11), 183 (2017)

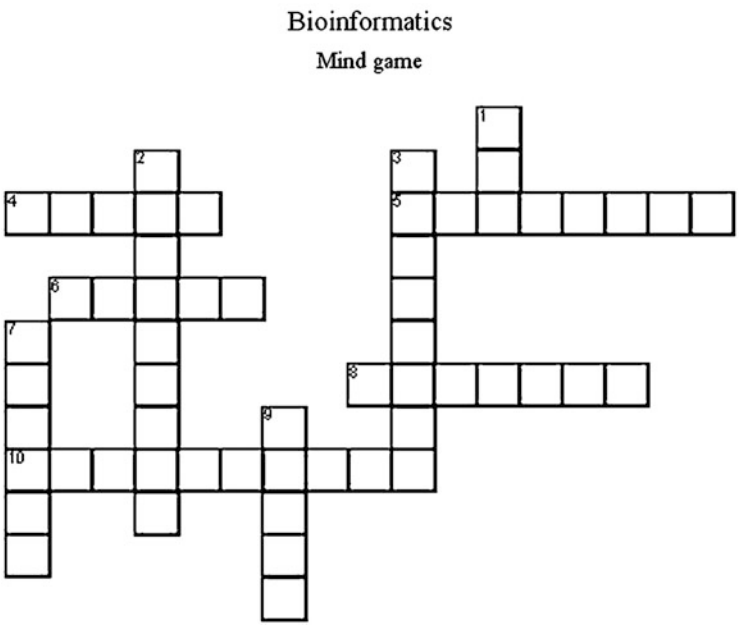
Kaur, H., Pannu, H.S., Malhi, A.K.: A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Computing Surveys (CSUR). **52**(4), 1–36 (2019)

Manisekhar, S.R., Siddesh, G.M., Manvi, S.S.: Introduction to bioinformatics. In: Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications, pp. 3–9. Springer, Singapore (2020)

Masoudi-Sobhanzadeh, Y., Omidi, Y., Amanlou, M., Masoudi-Nejad, A.: Drug databases and their contributions to drug repurposing. Genomics. **112**(2), 1087–1095 (2020)

Zhang, L., Tan, J., Han, D., Zhu, H.: From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov. Today. **22**(11), 1680–1685 (2017)

Bioinformatics Cross Word



ACROSS	
4	Primerprimer
5	Homologs duplicated
6	Well-known sequence format
8	NCBI
10	Orthologous sets of interacting proteins

(continued)

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8

DOWN	
1	Thermal cycler
2	A database but predicts
3	An alternative to antibodies
7	Conserved sequences
9	Cluster of computers virtually

Epilogue

Through this book, the author overtly conveyed to the readers from day-to-day experiences, he has traversed whence his illustrious travel to several countries. Otherwise, the author hopes it has tried to justify and confirm the traditional way of providing guidance to the beginners in bioinformatics. Thank you for reading.

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8

References

Altman, D.G., Bland, J.M.: Correlation, regression and repeated data. BMJ. **309**, 102 (1994)

Bruggeman, F.J., Westerhoff, H.V.: The nature of systems biology. Trends Microbiol. **15**(1), 45–50 (2007)

Calvo, S., Jain, M., Xie, X., Sheth, S.A., et al.: Systematic identification of human mitochondrial disease genes through integrative genomics. Nat. Genet. **38**, 576–582 (2006)

Chen, Y., et al.: Variations in DNA elucidate molecular networks that cause disease. Nature. **452**, 429–435 (2008)

Chotani, G., et al.: The commercial production of chemicals using pathway engineering. Biochim. Biophys. Acta. **1543**(2), 434–455 (2000)

Claudino, W.M., Quattrone, A., Biganzoli, L., Pestrin, M., et al.: Metabolomics: available results, current research projects in breast cancer, and future applications. J. Clin. Oncol. **25**(19), 2840–2846 (2007)

Cornell, M., Paton, N.W., Oliver, S.G.: A critical and integrated view of the yeast interactome. Comp. Funct. Genomics. **5**(5), 382–402 (2004)

Cornish-Bowden, A., Cárdenas, M.L., Letelier, J.C., Soto-Andrade, J.: Beyond reductionism: metabolic circularity as a guiding vision for a real biology of systems. Proteomics. **7**(6), 839–845 (2007)

Dell, H.G., Scott, R., et al.: A Greek-English Lexicon (1996)

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., et al.: Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U S A. **102**(40), 14338–14343 (2005)

Edwards, D., Stajich, J., Hansen, D. (eds.): Bioinformatics Tools and Applications, vol. XII, p. 451, 90 illus. Springer, New York (2009)

Fiehn, O.: Metabolomics—the link between genotypes and phenotypes. Plant Mol. Biol. **48**(1–2), 155–171 (2002)

Fiehn, O., Kristal, B., Van Ommen, B., Sumner, L.W., et al.: Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. Omics. **10**(2), 158–163 (2006)

Fierz, W.: Basic problems of serological laboratory diagnosis. Methods Mol. Med. **94**, 393–427 (2004)

Figeys, D.: Combining different 'omics' technologies to map and validate protein-protein interactions in humans. Brief. Funct. Genomic. Proteomic. **2**(4), 357–365 (2004)

Fox Keller, E., Harel, D.: Beyond the gene. PLoS One. **2**(11), e1231 (2007)

Gomase, V.S., Tagore, S.: Physiomics. Curr. Drug Metab. **9**(3), 259–262 (2008)

Govorun, V.M., Archakov, A.I.: Proteomic technologies in modern biomedical science. Biochemistry (Mosc.). **67**(10), 1109–1123 (2002)

Greenbaum, D., Luscombe, N.M., Jansen, R., Qian, J., et al.: Interrelating different types of genomic data, from proteome to secretome: 'oming in on function'. Genome Res. **11**(9), 1463–1468 (2001)

P. N. Suravajhala (ed.), *Your Passport to a Career in Bioinformatics*,
https://doi.org/10.1007/978-981-15-9544-8

Gronow, S., Brade, H.: Lipopolysaccharide biosynthesis: which steps do bacteria need to survive? J. Endotoxin Res. **7**(1), 3–23 (2001)
Han, X., Gross, R.W.: Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. J. Lipid Res. **4**(6), 1071–1079 (2003)
Huang, S.: Back to the biology in systems biology: what can we learn from biomolecular networks? Brief. Funct. Genomic. Proteomic. **2**(4), 279–297 (2004)
Huang, S., Wikswo, J.: Dimensions of systems biology. Rev. Physiol. Biochem. Pharmacol. **157**, 81–104 (2006)
Hucka, M., Finney, A., Bornstein, B.J., Keating, S.M., et al.: Evolving a lingua franca and associated software infrastructure for computational systems biology: the systems biology markup language (SBML) project. Syst. Biol. (Stevenage). **1**(1), 41–53 (2004)
Ideker, T., Hood, L.: A blueprint for systems biology. Clin Chem. **65**(2), 342–344 (2019). https://doi.org/10.1373/clinchem.2018.291062
Kaeberlein, M.: Aging-related research in the "-omics" age. Sci. Aging Knowledge Environ. **4**(42), pe39 (2004)
Kasper, L., Karlberg, E.O., Størling, Z.M., Ólason, P.Í., et al.: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. **25**, 309–316 (2007)
Katam, R., Basha, S.M., Suravajhala, P., Pechan, T.: Analysis of peanut leaf proteome. J. Proteome Res. **9**(5), 2236–2254 (2010)
Khanna, C.: The dog as a cancer model. Nat. Biotechnol. **24**(9), 1065–1066 (2006)
Kim, T.Y., Sohn, S.B., Kim, H.U., Lee, S.Y.: Strategies for systems-level metabolic engineering. Biotechnol. J. **3**(5), 612–623 (2008)
Kuiper, H.A., Kleter, G.A., Noteborn, H.P., Kok, E.J.: Assessment of the food safety issues related to genetically modified foods. Plant J. **27**(6), 503–528 (2001)
McGuire, J.N., Overgaard, J., Pociot, F.: Mass spectrometry is only one piece of the puzzle in clinical proteomics. Brief. Funct. Genomic Proteomic. **7**(1), 74–83 (2008)
Mehta, T.S., Zakharkin, S.O., Gadbury, G.L., Allison, D.B.: Epistemological issues in omics and high-dimensional biology: give the people what they want. Physiol. Genomics. **28**(1), 24–32 (2006)
Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., et al.: Calling on a million minds for community annotation in WikiProteins. Genome Biol. **9**(5), R89 (2008)
Morel, N.M., Holland, J.M., van der Greef, J., Marple, E.W., et al.: Primer on medical genomics. Part XIV: introduction to systems biology—a new approach to understanding disease and treatment. Clin. Proc. **79**(5), 651–658 (2004)
Morrison, N., Cochrane, G., Faruque, N., Tatusova, T., et al.: Concept of sample in OMICS technology. Omics. **10**(2), 127–137 (2006)
O'Rourke, N.A., Meyer, T., Chandy, G.: Protein localization studies in the age of 'Omics'. Curr. Opin. Chem. Biol. **9**(1), 82–87 (2005)
Oldiges, M., Lütz, S., Pflug, S., Schroer, K., et al.: Metabolomics: current state and evolving methodologies and tools. Appl. Microbiol. Biotechnol. **76**(3), 495–511 (2007)
Proceedings of the 3rd World Congress on Alternatives and Animal Use in the Life Sciences, Bologna, Italy, 29 Aug–2 Sept 1999
Reichert, A.S., Neupert, W.: Mitochondriomics or what makes us breathe. Trends Genet. **20**, 555–562 (2004)
Rochfort, S.: Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. J. Nat. Prod. **68**(12), 1813–1820 (2005)
Schloss, P.D., Handelsman, J.: Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. **14**(3), 303–310 (2003)
Schork, N.J.: Genetics of complex disease: approaches, problems and solutions. Am. J. Respir. Crit. Care Med. **156**(4 Pt 2), S103–S109 (1997)

Steinfath, M., Repsilber, D., Scholz, M., Walther, D., et al.: Integrated data analysis for genome-wide research. EXS. **97**, 309–329 (2007)
Stransky, B., Barrera, J., Ohno-Machado, L., De Souza, S.J.: Modeling cancer: integration of "omics" information in dynamic systems. J. Bioinform. Comput. Biol. **5**(4), 977–986 (2007)
Strömbäck, L., Jakoniene, V., Tan, H., Lambrix, P.: Representing, storing and accessing molecular interaction data: a review of models and tools. Brief. Bioinform. **7**(4), 331–338 (2006)
Suravajhala, P.: Hypo, hype and 'hyp' human proteins. Bioinformation. **2**(1), 31–33 (2007)
Taylor, D.L., Woo, E.S., Giuliano, K.A.: Real-time molecular and cellular analysis: the new frontier of drug discovery. Curr. Opin. Biotechnol. **12**(1), 75–81 (2001)
Tracy, R.P.: 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. Curr. Opin. Lipidol. **19**(2), 151–157 (2008)
Ward, N.: New directions and interactions in metagenomics research. FEMS Microbiol. Ecol. **55** (3), 331–338 (2006)
Werner, T.: Proteomics and regulomics: the yin and yang of functional genomics. Mass Spectrom. Rev. **23**(1), 25–33 (2004)

Web References

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1994.tb01287.x
https://www.nature.com/news/2011/110831/full/477020a.html
https://www.rncos.com/Market-Analysis-Reports/Bioinformatics-Market-Outlook-to-2015-IM382.htm
https://www.genecards.org