Advances in Biochemical Engineering/Biotechnology 160

Series Editor: T. Scheper

Intawat Nookaew *Editor*

Network Biology

Springer

160
Advances in Biochemical Engineering/Biotechnology

Aims and Scope

This book series reviews current trends in modern biotechnology and biochemical engineering. Its aim is to cover all aspects of these interdisciplinary disciplines, where knowledge, methods and expertise are required from chemistry, biochemistry, microbiology, molecular biology, chemical engineering and computer science.

Volumes are organized topically and provide a comprehensive discussion of developments in the field over the past 3–5 years. The series also discusses new discoveries and applications. Special volumes are dedicated to selected topics which focus on new biotechnological products and new processes for their synthesis and purification.

In general, volumes are edited by well-known guest editors. The series editor and publisher will, however, always be pleased to receive suggestions and supplementary information. Manuscripts are accepted in English.

In references, Advances in Biochemical Engineering/Biotechnology is abbreviated as *Adv. Biochem. Engin./Biotechnol.* and cited as a journal.

More information about this series at http://www.springer.com/series/10

Intawat Nookaew
Editor

Network Biology

With contributions by

P. Ajawatanawong · Y. Akiyama · S. Dahal · G. Hu ·
D. Jacobson · S. Kalapanulak · A. Klanchui ·
K. Kusonmano · Y. Li · Y. Li · Y. Matsuzaki · A. Meechai ·
M. Ohue · G. Pavesi · S. Poudel · P. Prommeenate ·
N. Raethong · T. Saithong · C. Thammarongtham ·
R.A. Thompson · N. Uchikoga · W. Vongsangnak ·
D.A. Weighill · F. Xiao

Springer

Editor
Intawat Nookaew
Department of Biomedical Informatics
College of Medicine
University of Arkansas for Medical Science
Little Rock, Arkansas, USA

Computational Biomolecular Modeling and Bioinformatics Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

ISSN 0724-6145 ISSN 1616-8542 (electronic)
Advances in Biochemical Engineering/Biotechnology
ISBN 978-3-319-56459-3 ISBN 978-3-319-56460-9 (eBook)
DOI 10.1007/978-3-319-56460-9

Library of Congress Control Number: 2017938643

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Biological systems are extremely complex and contain millions of molecules within the system. The rapid development of high throughput technologies enables us to capture the molecular interplays of molecules in the system as so-called 'OMICS' data. This leads to the need for systematic cataloguing and organization of the enormous amount of data generated and shared within the scientific community. Linking these molecules and evaluating their interactions following "Network Biology" approaches enable the insightful understanding of cellular functions from the emerging properties of the network. This special volume focuses on the state of the art, current status, and applications of Network Biology.

The volume covers broad topics on network biology such as gene networks, transcription networks, regulatory networks, protein–protein interaction networks, metabolic networks, and phylogenetic networks. I am very grateful to the authors who have contributed to this special volume by sharing their experience and expertise in the different chapters. These diverse topics should be very useful for readers to gain an overview of Network Biology.

Oak Ridge, TN, USA Intawat Nookaew

Contents

ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks 1
Giulio Pavesi

Gene Expression Analysis Through Network Biology: Bioinformatics Approaches 15
Kanthida Kusonmano

Rigid-Docking Approaches to Explore Protein–Protein Interaction Space 33
Yuri Matsuzaki, Nobuyuki Uchikoga, Masahito Ohue, and Yutaka Akiyama

Protein–Protein Interface and Disease: Perspective from Biomolecular Networks 57
Guang Hu, Fei Xiao, Yuqian Li, Yuan Li, and Wanwipa Vongsangnak

Cyanobacterial Biofuels: Strategies and Developments on Network and Modeling 75
Amornpan Klanchui, Nachon Raethong, Peerada Prommeenate, Wanwipa Vongsangnak, and Asawin Meechai

Genome-Scale Modeling of Thermophilic Microorganisms 103
Sanjeev Dahal, Suresh Poudel, and R. Adam Thompson

Networking Omic Data to Envisage Systems Biological Regulation 121
Saowalak Kalapanulak, Treenut Saithong, and Chinae Thammarongtham

Network Metamodeling: Effect of Correlation Metric Choice on Phylogenomic and Transcriptomic Network Topology 143
Deborah A. Weighill and Daniel Jacobson

Molecular Phylogenetics: Concepts for a Newcomer 185
Pravech Ajawatanawong

Index 197

Adv Biochem Eng Biotechnol (2017) 160: 1–14
DOI: 10.1007/10_2016_43

Published online: 10 January 2017

ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks

Giulio Pavesi

Abstract The first step in the definition of transcriptional regulatory networks is to establish correct relationships between transcription factors (TFs) and their target genes, together with the effect of their regulatory activity (activator or repressor). Fundamental advances in this direction have been made possible by the introduction of experimental techniques such as Chromatin Immunoprecipitation, which, coupled with next-generation sequencing technologies (ChIP-Seq), permit the genome-wide identification of TF binding sites. This chapter provides a survey on how data of this kind are to be processed and integrated with expression and other types of data to infer transcriptional regulatory rules and codes.

Keywords ChIP-Seq, RNA-Seq, Transcription factors, Transcription regulation

Contents

1 Introduction: Chromatin Immunoprecipitation and Next-Generation Sequencing 2
2 Finding Transcription Factor Binding Sites 6
3 Associating Binding Sites with Target Genes 7
4 Assessing TF Activity from Expression Data 8
5 Mining Available Data 10
6 Conclusions 10
References 11

G. Pavesi (✉)
Department of Biosciences, University of Milan, Via Celoria 26, 20133 Milan, Italy
e-mail: giulio.pavesi@unimi.it

1 Introduction: Chromatin Immunoprecipitation and Next-Generation Sequencing

The introduction of *next-generation sequencing* (NGS) technologies has opened up new avenues for every type of genetic and genomic research [1, 2]. One of the fields in which the impact of NGS has been more relevant is perhaps the study of gene regulation at the transcriptional level, and the subsequent analysis steps such as the construction of regulatory networks.

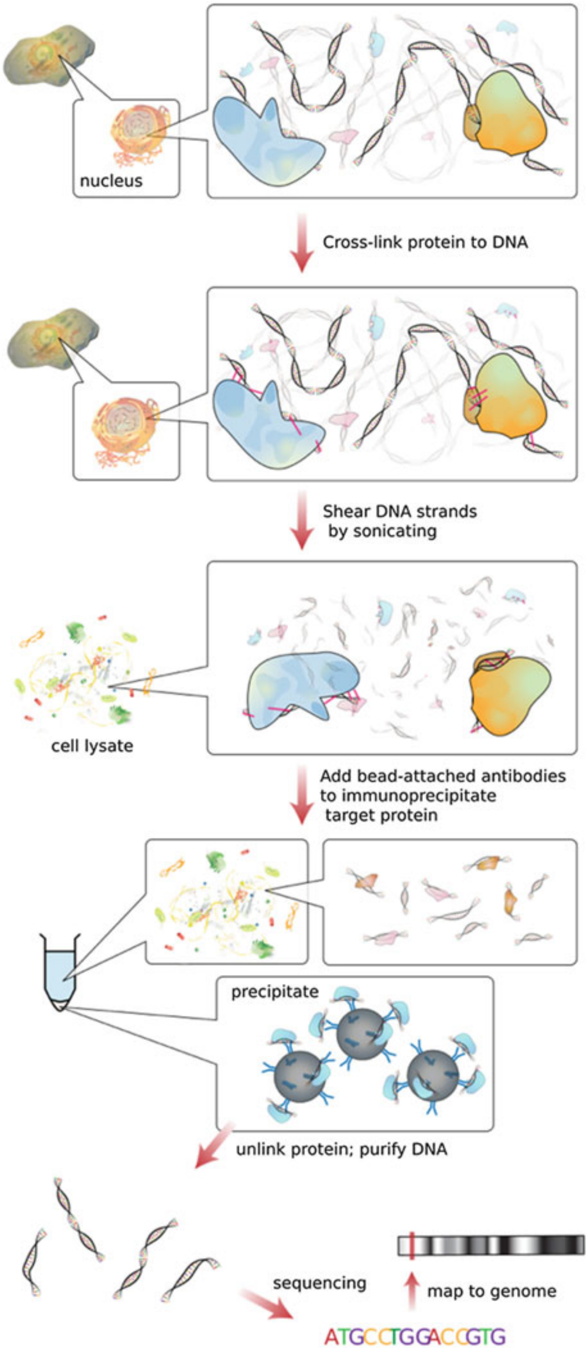
It is essential for the definition of transcription regulatory networks to establish correct relationships between regulators such as transcription factors (TFs) and the genes they regulate [3], together with the effect of the activity of the TFs (activator or repressor) [4]. A fundamental step forward in this direction has been made possible by lab techniques enabling the large-scale identification of TF-DNA binding sites on the genome, with experiments simply impossible to perform just a few years ago.

Chromatin is a complex of DNA and proteins that forms chromosomes within the nucleus of eukaryotic cells. *Chromatin Immunoprecipitation* (ChIP) [5] is a technique enabling the extraction from the cell nucleus of a specific protein-DNA chromatin complex, including DNA binding proteins such as TFs. The different steps of a ChIP experiment are summarized in Fig. 1. First of all, the DNA-bound proteins are cross-linked, that is, fixed to the DNA. The cross-linked chromatin is usually sheared by sonication, providing fragments of 300–1,000 base pairs (bps) in length. Then a specific antibody that recognizes only the protein (TF) of interest is employed, and the antibody, bound to the TF which in turn is bound to the DNA, permits the selective extraction and isolation of the chromatin complex. At this point, DNA is released from the TF by reverse-crosslinking and purified, and the result is a DNA sample enriched in regions corresponding to the genomic locations of the sites that were bound in vivo by the TF (or, in general, the DNA-binding protein) studied. The experiment is performed on thousands of cells at the same time so as to have a quantity of DNA suitable for further analysis and to have enough "enrichment" in the sample, that is, enough copies of each of the DNA regions bound by the TF, to discriminate them from experimental noise.

The next phase is quite logically the identification of the DNA regions themselves – and of their corresponding location in the genome. The introduction of "tiling arrays" had permitted for the first time the analysis of the DNA extracted on a whole-genome scale (ChIP on Chip [4, 6]) by using probes designed to cover the sequence of a whole genome, or a subset of genomic regions of interest (such as with promoter arrays). The introduction of NGS technologies has enabled this type of experiment to move one step further by providing at reasonable cost perhaps the simplest solution: to identify the DNA extracted by the cell by immunoprecipitation, sequence the DNA itself (ChIP Sequencing, or ChIP-Seq [5, 7]).

Without delving into technical details, given a double-stranded DNA fragment derived as just described, sequencing determines the nucleotide sequence on either strand, moving from the $5'$ to $3'$ direction, or both strands simultaneously (paired-

Fig. 1 Chromatin immunoprecipitation workflow (adapted from Wikipedia)


nucleus
Cross-link protein to DNA
Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

end sequencing). For technical limitations, current NGS platforms can determine the sequence of only a fragment of each region, usually ranging from 50 to 150 bps. Thus, the output is a huge collection of millions of short sequences (called *reads*), which mark the beginning of either or both strands of a DNA region of the sample. The overall number of sequence reads obtained varies from experiment to experiment, and depends on several factors such as the TF involved, sample preparation, experiment replicates, and so on. Suffice it to say that it usually ranges from a few to dozens of millions of short sequence reads.

Once the sequencing has been completed, computational analysis of the data determines which were the DNA regions enriched in the sample (see Fig. 2). First of all, the reads are aligned or "mapped" on the genome to determine their original

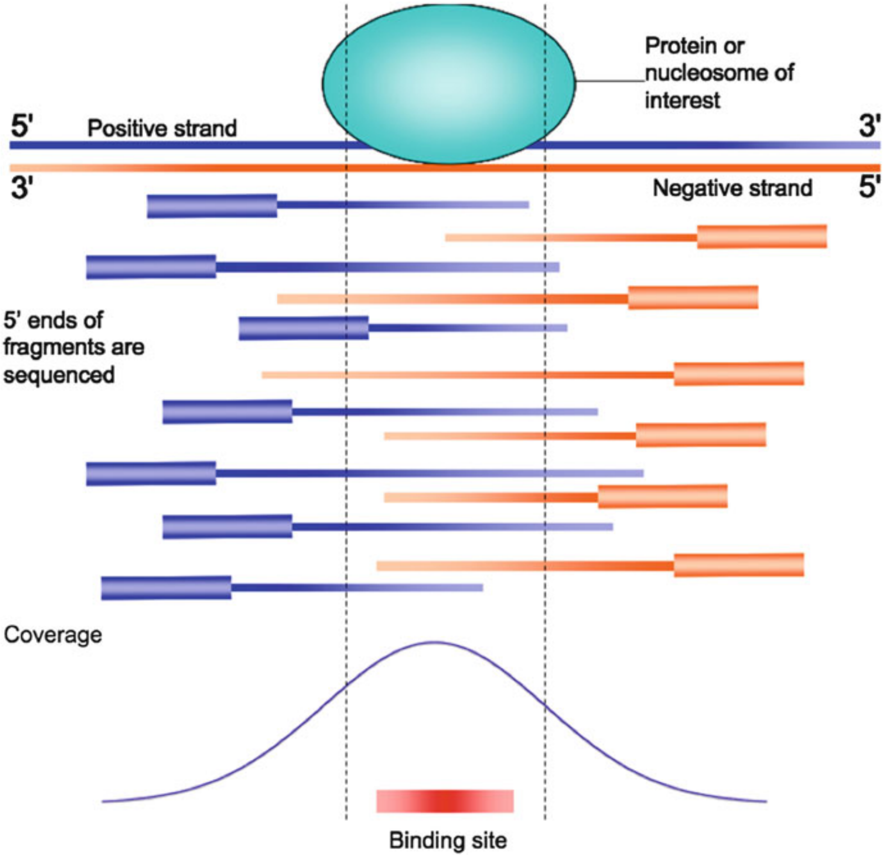

Protein or nucleosome of interest
5'
Positive strand
3'
3'
Negative strand
5'
5' ends of fragments are sequenced
Coverage
Binding site

Fig. 2 Schematic view of the result of a ChIP-Seq experiment on a genomic region bound by a TF. DNA is fragmented at random by sonication, and thus the ends of sequenced DNA fragments map on different positions on the genome. Each fragment is assumed to be the 5′ of a 200–300 bps region, and therefore extended. The resulting signal plot ("coverage") shows a typical "peak" shape. The actual DNA sequence bound by the TF should be located in correspondence of the point of maximum of the coverage plot (*bottom*)

position, using one of the several tools available for this task [8]. It is common, at this stage, to have mismatches in the alignment, that is, sequence reads differ from the reference genome sequence usually in single nucleotides. This is for both biological (sequence polymorphisms) and technical (sequencing errors) reasons. Thus, alignment is usually performed allowing for two or three substitutions per read, with no insertions or deletions. In addition, a non-negligible number of sequence reads align at multiple positions, that is, correspond to repetitive regions of the genome. Although originally these were discarded from further processing, it has indeed been shown that TFs can bind repetitive elements of the genome [9]. Thus, reads mapping at multiple positions should also be considered in the remainder of the analysis, for example also keeping those that map at most in ten different positions.

Once read mapping is complete, regions bordered by reads on both ends (on opposite strands) in numbers high enough to represent a "significant enrichment" and not sample contamination or random noise are singled out. This latter step should be performed with respect to a "control" experiment, aimed at producing "random DNA" and thus a random background model. In other words, if "random" genomic DNA was included in immunoprecipitated samples, another experiment producing only "random" DNA from the same type of cell should give the opportunity to filter the results from false positives and artifacts. The control experiment can be performed in different ways by using an antibody not specific for any TF or, if possible, by using a cell in which the gene encoding the TF studied has been "knocked out," or its expression "knocked down" in order to remove the immunoprecipitated protein from cells [10].

An ideal example of enriched region is shown in Fig. 2. A "true positive" should correspond to a genomic region bordered by several reads on both strands, and the reads on the two ends should be at a distance "typical" of experiments of this kind, that is, a few hundred bps. By plotting the number of reads falling in each genomic position, the region should be comprised between two "peaks," one made by reads on the positive strand and one on the negative. Each read mapped on the genome can also be extended by the estimated length of the immunoprecipitated DNA fragments. The latter, following a size-selection step before sequencing, is usually about 200 bps. The result is a signal plot estimating how many times each nucleotide of the genome is covered by an "extended read." Then a "significantly enriched" region should correspond to a peak in the signal plot, usually located in the middle of the region itself. As in experiments such as ChIP-Seq enrichment is essential to obtain reliable results, single-end sequencing is preferred over paired-end, which would produce at the same cost exactly one half of the sequences, and thus less enrichment.

On the other hand, the same region should not appear – at least with the same number of bordering reads or with the same height of the central peak – in the control experiment. Given the shape of the enriched regions as shown in Fig. 2, this part of the analysis is usually referred to as "peak calling," that is, identifying all the "peak shaped" regions whose enrichment can be considered to be statistically significant. From the introduction of ChIP-Seq experiments, several different methods for peak calling have been introduced, all following the above considerations but differing in the statistical approaches employed in the definition of significant enrichment. The latter is computed according to the overall number of reads that can be associated with a candidate peak, their distribution on the two DNA strands, and the height of the peak summit. These values are in turn compared to background expected values that might or might not be derived from a control experiment. In a quite ample literature, a few methods have emerged over the years as de facto standards, such as, for example, MACS [11, 12], SPP [13], and PeakSeq [14], which have been employed in the large scale analysis of hundreds of ChIP-Seq experiments performed in the framework of the ENCODE project [15, 16].

The output of peak-calling is a list of genomic regions, likely to be bound by the TF studied in vivo, with p-values and false discovery rates (FDRs) associated with each one. Thus, not only is a "yes/no" output provided but also an estimate of the probability of each region to be considered a false positive call, and hence an estimate of its actual enrichment in the sample. The latter can be employed to restrict, for example, downstream analyses only to the "most likely" or "most significantly enriched" candidates (e.g., only those for which the estimated FDR is under a given threshold). In addition, the "summit" point of each region is usually included in the output, that is, the genomic coordinate of the single base pair where the signal plot associated with the peak is maximum (see Fig. 2). As the actual point of contact with DNA of the TF or the complex investigated should be present in all the regions extracted, the latter should be close to the summit point, which can thus be used to approximate the binding site of the TF within the region for downstream analyses.

2 Finding Transcription Factor Binding Sites

The actual DNA region bound by a TF usually ranges in size from 8–10 to 16–20 bps [3]. TFs bind the DNA in a sequence-specific fashion, that is, they recognize sequences that are similar but not identical, differing in a few nucleotides from one another. As peak regions bound by a TF identified through ChIP-Seq are usually several hundreds of bps long, further processing is needed to identify the actual binding sites within them. Motif discovery or enrichment tools can be employed for this task [17, 18]. The general idea is that the regions identified by the ChIP-Seq, should contain a subset of oligos appearing in all or most of the sequences (thus allowing for experimental errors and the presence of false positives in the set) similar enough to one another to be instances of sites recognized by the same TF. The same set of similar oligos should also not appear with the same frequency and/or the same degree of similarity in a set of sequences selected at random or built at random with a generator of "biologically feasible" DNA sequences [19]. This set of similar and over-represented oligos collectively build a *motif* recurring in the input sequences, describing the binding specificity of the TF itself. Instances of the motif within the enriched regions can then be used to identify the actual binding sites within them. A motif enrichment analysis might also be useful for the identification of additional motifs enriched within the regions which could correspond to binding sites for additional TFs binding DNA in close proximity to the one investigated [20], and thus likely to co-associate with it forming regulatory modules.

3 Associating Binding Sites with Target Genes

The results of ChIP-Seq experiments provide a map of the binding sites on the genome for the TF investigated, but obviously no information regarding genes whose transcription is affected by each of the binding sites. For building regulatory networks it is therefore essential to associate each region with one or more "target" genes.

The first logical step is to single out binding sites located within promoters. There is no unique definition of what constitutes the "promoter" of a gene or of its size. It is usually described as a region of a few hundred or thousand base pairs located upstream of its transcription start site (TSS). ChIP-Seq experiments performed on histone modifications, however, revealed that active promoters have a very precise chromatin signature, that is, a pattern of modifications such as H3K4me3 or H3K9ac covering a few nucleosomes upstream and downstream of the TSS itself [21]. Hence, even if it narrows down the number of binding sites that can be assigned to promoters, it is advisable not to define a region too broad around TSSs as "promoter" and avoid going beyond 1 kbp upstream or downstream of the TSS. Indeed, TF binding regions outside these "core promoters" (e.g., within the first intron or further than 1 kbp upstream of the TSS) exhibit a different chromatin signature, with modifications such as H3K27ac or H3K4me1 that are indicators of distal "enhancer" or "silencer" regions but not of promoters.

Associating distal binding sites, not close to TSSs, with the "right" target genes is perhaps the hardest part of this type of analysis. Even factors usually associated with promoters and TSSs such as NF-Y [9, 22] have the majority of their binding sites located in distal regulatory regions. Thus, restricting the analysis only to binding sites located in promoters has the effect of missing several target genes regulated by the binding of the TF to distal elements; on the other hand, associating a distal regulatory element with the wrong gene produces wrong data.

In the absence of further information, this step usually follows the "nearest neighbor rule": a distal binding site is associated with the closest TSS on the genome. If the binding site is within a gene body (the transcribed region of a gene) then it is attributed to the gene itself. Given a reference annotation providing the genomic coordinates of genes that can be retrieved from any genome browser [23, 24], this analysis can be performed with in-house developed scripts, or with tools such as HOMER [25] or GREAT [26]. On the other hand, as a typical ChIP-Seq experiment returns several thousands of bound regions, associating every peak with the closest TSS results in a very sizable portion of the annotated genes to be considered targets of the TF investigated. Hence, further criteria are employed to reduce their number, usually by establishing a threshold on the distance from the TSS of the binding sites. For example, in the large-scale analysis performed in the Roadmap Epigenomics project [21], an enhancer region was associated with the closest gene if its TSS was located at less than 30 kbp from the enhancer itself. Otherwise, no association was defined.

Modern experimental techniques based on immunoprecipitation and NGS, the most relevant being ChIA-PET or ChIA-Seq experiments [27, 28], have enabled light to be shed on this aspect too. The ChIA-PET (or -Seq) method combines ChIP-Seq methods and Chromosome conformation capture techniques such as 3C [29] for the identification of long-range chromatin regulatory interactions [30]. The immunoprecipitation is performed against a protein usually found in complexes connecting enhancers to the respective TSSs, such as p300, to be pulled down together with all the DNA regions bound to it. Before sequencing, linker sequences are incorporated onto the free ends of the DNA fragments tethered to the protein complexes. To build connectivity of the DNA fragments, the linker sequences are ligated by nuclear proximity ligation. The resulting DNA sequences is thus formed by both the enhancer and the promoter, connected by a linker sequence. Application of NGS paired-end sequencing produces sequence pairs coming from each of the two connected regions. Subsequent mapping on the genome finally results not in single peaks but in “paired” peaks, located at different positions of the genome, where reads in one peak are found to be paired in sequencing with reads in the other. Paired peaks correspond to pairs of genomic regions connected by the protein complex immunoprecipitated. These experiments thus enable the identification of unique, functional chromatin interactions between distal and proximal regulatory transcription-factor binding sites and the promoters of the genes with which they interact. Remarkably, their application has revealed the serious limitations of the application of the “nearest neighbor” rule introduced before: for example, in mouse stem cells only about one-third of the long-distance enhancer-promoter interactions have been shown to be associated with the gene nearest to the enhancer [31]. An enhancer located within a transcribed region can also regulate a distal gene. Finally, a sizable number of the enhancers (about 30% of the total) were even associated with genes located on different chromosomes. All in all, then, in the absence of long-distance interaction data, all the enhancer-promoter associations should be taken with a pinch of salt.

4 Assessing TF Activity from Expression Data

TFs can have the effect of both activating and repressing the transcription of target genes. Thus, the activity of any TF can be assessed by performing experiments in which the expression of the TF is limited or, vice versa, amplified. Then the activity of the TF on target genes can be measured by identifying those genes that change their expression level as a consequence of the TF inactivation or over-expression.

Before the introduction of genome-wide techniques such as ChIP-Seq this was indeed the method of choice for the identification of putative target genes for TFs. It is, however, important to stress the fact that this approach, alone, might also identify genes that are not direct targets. In other words, the TF directly affects the expression of a subset of differentially expressed genes; some of the direct targets can in turn regulate further genes, also found to be differentially expressed, and so on.

Before the advent of NGS technologies, expression studies were usually performed with oligonucleotide microarrays. Then the application of NGS to RNA (RNA-Seq) was shown to be able not only to reconstruct and assemble whole transcriptomes, but also to provide a reliable quantification of the expression level of each gene [32, 33].

One of the key advantages of RNA-Seq over microarrays is that they enable one to identify and reconstruct the single alternative transcripts of the same gene, as well as estimate their expression level. This, in turn, has revealed alternative splicing and alternative transcript production to be ubiquitous features of eukaryotic genes [34]. From the viewpoint of transcription regulation it is worth mentioning that alternative promoters and transcription start sites have emerged as a widespread feature. This is a very important point in the association between TF binding and promoters, as a TF-gene association could be missed if the alternative promoter bound by the TF is not included in the analysis. For a TF binding only one of the alternative promoters of a gene, its effect on gene transcription should be assessed only for the corresponding transcripts. Techniques such as Cap Analysis Gene Expression (CAGE [35]), coupled with NGS sequencing [36], enable one to identify more reliably alternative TSSs and the relative transcription level.

It is worth mentioning that the usual measures of transcript level employed are concentration measures. That is, the "expression level" of a transcript or gene is an estimate of the fraction of the RNA sample that can be assigned to it, described by normalized measures such as "reads per kilobase of exon per million reads" (RPKM) or "transcripts per million" (TPM). This, in turn, can produce incorrect conclusions when applied to experiments resulting from TF inactivation or overexpression. Suppose, for example, that a TF acts purely as an activator, targeting 10% of the genes of the genome studied. Upon inactivation of the TF, the transcript level of its target genes is decreased and the rest of the genome remains unchanged. As expression measures used are relative and describe concentration with respect to the overall sample, we observe a marked reduction of the transcript levels for the target genes, but at the same time an increase of the expression estimate of non-target genes, some of which might also finally be "significantly overexpressed" by statistical analysis. Hence, the TF is incorrectly observed to act both as an activator and a repressor. Other than previous knowledge about the TF activity, indicators of the possible presence of this effect for an activator TF are a large majority of genes significantly down-regulated with just a few overexpressed, the latter having very high expression estimates. Vice versa for repressor TFs. In case of doubt, special techniques should be employed in the design and analysis of the expression experiment, as shown, for example, in [37].

5 Mining Available Data

The ever decreasing cost of next-generation sequencing has led to the widespread application of the techniques described in this chapter, such as ChIP- and RNA-Seq. It has indeed become common practice to study simultaneously more than one TF in a given condition in order to have more meaningful results and to identify co-associations and modules of key regulators [38, 39]. The last few years have also witnessed the completion of large-scale general purpose projects in which hundreds of TFs have been tested in several different cell lines. The most relevant example is perhaps the (still ongoing) ENCODE project [40], in which hundreds of human and mouse TFs have been analyzed through ChIP-Seq in several different cell lines, or the modENCODE project for model organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans* [41]. TF ChIP-Seq data are integrated by other data relevant for transcriptional regulation analysis such as chromatin structure, histone modifications, DNA methylation, expression profiles from RNA-Seq and CAGE experiments, and ChIA-PET data for long-distance chromosomal interactions. Analysis of co-occurrence of TF binding sites of the genome revealed that TFs tend to associate, forming distinct co-regulatory modules [15], giving rise to many enriched regulatory network motifs (e.g., noise-buffering feed-forward loops). Hence, any TF should not be viewed as a separate entity whose interactions with other regulatory factors happen only by chance, but should be considered as part of more complex regulatory modules, and the construction of regulatory networks should consider this point.

Other than the deluge of information they contain, these data, or those contained in large repositories such as Cistrome [42], constitute a perfect benchmark set for any bioinformatics or systems biology approach to the study of transcriptional regulation. They can also be retrieved to complement data produced locally. There also exist resources in which data have already been processed, for example tools such as Cscan [43] or Enrichr [44], which already have pre-computed associations between TFs and target genes for hundreds of experiments.

6 Conclusions

The introduction and the creative use of next-generation sequencing technologies have opened new avenues for every aspect of genetic and epigenetic research. Perhaps the field that has benefited most from them is regulation of gene expression at the transcriptional and post-transcriptional level. This chapter provides a brief survey of the experimental and bioinformatic techniques currently employed for the study of transcription factors, summarized in Fig. 3, from the identification of target genes to the characterization of their activity, and all fundamental steps for subsequent studies such as the definition and analysis of transcriptional regulatory networks.

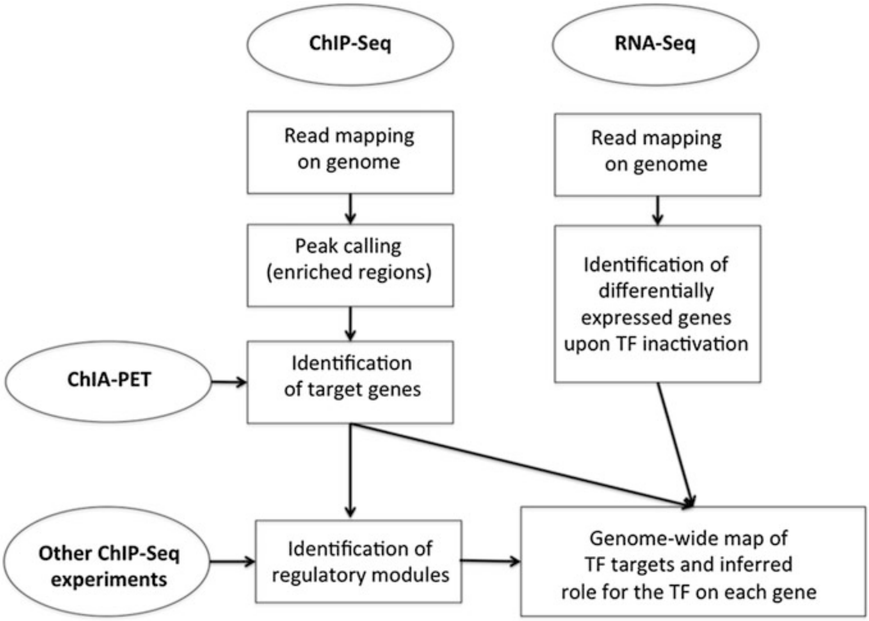

ChIP-Seq
RNA-Seq
Read mapping on genome
Read mapping on genome
Peak calling (enriched regions)
Identification of differentially expressed genes upon TF inactivation
ChIA-PET
Identification of target genes
Other ChIP-Seq experiments
Identification of regulatory modules
Genome-wide map of TF targets and inferred role for the TF on each gene

Fig. 3 Combining different NGS-based experiments for building the regulatory map of a given TF. ChIP-Seq identifies genomic regions bound by the TF, and further processing the corresponding target genes. The latter can in turn be more easily singled out by capturing long-distance interactions with experiments such as ChIA-PET. RNA-Seq experiments assess significant changes of gene expression upon TF inactivation. Different ChIP-Seq experiments performed in the same condition can be combined to identify regulatory modules

References

1. Horner DS, Pavesi G, Castrignano T, De Meo PD, Liuni S, Sammeth M, Picardi E, Pesole G (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform 11(2):181–197. doi:10.1093/bib/bbp046
2. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24(3):133–141. doi:10.1016/j.tig.2007.12.007
3. Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424 (6945):147–151. doi:10.1038/nature01763
4. Blais A, Dynlacht BD (2005) Constructing transcriptional regulatory networks. Genes Dev 19 (13):1499–1511. doi:10.1101/gad.1325605
5. Collas P, Dahl JA (2008) Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Front Biosci 13:929–943
6. Pillai S, Chellappan SP (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. Methods Mol Biol 523:341–366
7. Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4(8):613–614. doi:10.1038/nmeth0807-613
8. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. doi:10.1186/gb-2009-10-3-r25

9. Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K (2013) NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res 23 (8):1195–1209. doi:10.1101/gr.148080.112
10. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6(11 Suppl):S22–S32. doi:10.1038/nmeth.1371
11. Feng J, Liu T, Zhang Y (2011) Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics Chapter 2:Unit 2. 14. doi:10.1002/0471250953.bi0214s34
12. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9 (9):R137. doi:10.1186/gb-2008-9-9-r137
13. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5(9):829–834. doi:10.1038/nmeth.1246
14. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27(1):66–75. doi:10.1038/nbt.1518
15. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489(7414):91–100. doi:10.1038/nature11245
16. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22(9):1813–1831. doi:10.1101/gr.136184.111
17. Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Res 43 (W1):W39–W49. doi:10.1093/nar/gkv416
18. Zambelli F, Pesole G, Pavesi G (2014) Using Weeder, Pscan, and PscanChIP for the discovery of enriched transcription factor binding site motifs in nucleotide sequences. Curr Protoc Bioinformatics 47:2. 11. 11–12. 11. 31. doi:10.1002/0471250953.bi0211s47
19. Zambelli F, Pesole G, Pavesi G (2013) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief Bioinform 14(2):225–237. doi:10.1093/bib/bbs016
20. Zambelli F, Pesole G, Pavesi G (2013) PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res 41(Web Server issue):W535–W543. doi:10.1093/nar/gkt448
21. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M,

Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M (2015) Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317–330. doi:10.1038/nature14248

22. Ceribelli M, Dolfini D, Merico D, Gatta R, Vigano AM, Pavesi G, Mantovani R (2008) The histone-like NF-Y is a bifunctional transcription factor. Mol Cell Biol 28(6):2047–2058. doi:10.1128/MCB.01861-07
23. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, Humphrey J, Kerhornou A, Khobova J, Aranganathan NK, Langridge N, Lowy E, McDowall MD, Maheswari U, Nuhn M, Ong CK, Overduin B, Paulini M, Pedro H, Perry E, Spudich G, Tapanari E, Walts B, Williams G, Tello-Ruiz M, Stein J, Wei S, Ware D, Bolser DM, Howe KL, Kulesha E, Lawson D, Maslen G, Staines DM (2015) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. doi:10.1093/nar/gkv1209
24. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Fujita PA, Eisenhart C, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ (2015) The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. doi:10.1093/nar/gkv1275
25. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38 (4):576–589. doi:10.1016/j.molcel.2010.05.004
26. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28(5):495–501. doi:10.1038/nbt.1630
27. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi HS, Tennakoon C, Wei CL, Ruan Y, Sung WK (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol 11(2):R22. doi:10.1186/gb-2010-11-2-r22
28. Paulsen J, Rodland EA, Holden L, Holden M, Hovig E (2014) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. Nucleic Acids Res 42(18), e143. doi:10.1093/nar/gku738
29. Simonis M, Kooren J, de Laat W (2007) An evaluation of 3C-based methods to capture DNA interactions. Nat Methods 4(11):895–901. doi:10.1038/nmeth1114
30. Li G, Cai L, Chang H, Hong P, Zhou Q, Kulakova EV, Kolchanov NA, Ruan Y (2014) Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. BMC Genomics 15(Suppl 12):S11. doi:10.1186/1471-2164-15-S12-S11
31. Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, Mulawadi FH, Sung WK, Nicolis S, Ahituv N, Ruan Y, Wei CL (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504 (7479):306–310. doi:10.1038/nature12716
32. Fonseca NA, Marioni J, Brazma A (2014) RNA-Seq gene profiling—a systematic empirical comparison. PLoS One 9(9), e107026. doi:10.1371/journal.pone.0107026
33. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18 (9):1509–1517. doi:10.1101/gr.079558.108
34. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456 (7221):470–476. doi:10.1038/nature07509

35. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100 (26):15776–15781. doi:10.1073/pnas.2136655100
36. Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7(3):542–561. doi:10.1038/nprot.2012.005
37. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA (2012) Revisiting global gene expression analysis. Cell 151(3):476–482. doi:10.1016/j.cell.2012.10.012
38. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133(6):1106–1117. doi:10.1016/j.cell.2008.04.043
39. Hutchins AP, Diez D, Takahashi Y, Ahmad S, Jauch R, Tremblay ML, Miranda-Saavedra D (2013) Distinct transcriptional regulatory modules underlie STAT3’s cell type-independent and cell type-specific functions. Nucleic Acids Res 41(4):2155–2170. doi:10.1093/nar/gks1300
40. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 22(9):1798–1812. doi:10.1101/gr.139105.112
41. Brown JB, Celniker SE (2015) Lessons from modENCODE. Annu Rev Genomics Hum Genet 16:31–53. doi:10.1146/annurev-genom-090413-025448
42. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12(8):R83. doi:10.1186/gb-2011-12-8-r83
43. Zambelli F, Prazzoli GM, Pesole G, Pavesi G (2012) Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. Nucleic Acids Res 40 (Web Server issue):W510–W515. doi:10.1093/nar/gks483
44. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma’ayan A (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14:128. doi:10.1186/1471-2105-14-128

Adv Biochem Eng Biotechnol (2017) 160: 15–32
DOI: 10.1007/10_2016_44

Published online: 10 November 2016

Gene Expression Analysis Through Network Biology: Bioinformatics Approaches

Kanthida Kusonmano

Abstract Following the availability of high-throughput technologies, vast amounts of biological data have been generated. Gene expression is one example of the popular data that has been utilized for studying cellular systems in the transcriptional level. Several bioinformatics approaches have been developed to analyze such data. A typical expression analysis identifies a ranked list of individual significant differentially expressed genes between two conditions of interest. However, it has been accepted that biomolecules in a living organism are working together and interacting with each other. Study through network analysis could be complementary to typical expression analysis and provides more contexts to understanding the biological systems. Conversely, expression data could provide clues to functional links between biomolecules in biological networks. In this chapter, bioinformatics approaches to analyze expression data in network levels including basic concepts of network biology are described. Different concepts to integrate expression data with interactome data and example studies are explained.

Keywords Biological network analysis, Data integration, Gene expression analysis, Interactome, Network biology

Contents

1 Introduction 16
2 Interactome Networks 17
2.1 Protein–Protein Interaction Networks 18
2.2 Regulatory Networks 18
2.3 Metabolic Networks 21

K. Kusonmano (✉)
Bioinformatics and Systems Biology Program, School of Bioresources and Technology,
King Mongkut’s University of Technology Thonburi, Bangkhuntien, Bangkok, Thailand
e-mail: kanthida.kus@kmutt.ac.th

2.4 Functional Interaction Networks ... 21
3 Graph Concepts for Network Analysis ... 22
3.1 Graph Definition ... 22
3.2 Network Properties ... 23
4 Identification of Modules or Subnetworks ... 24
5 Expression Analysis Through Network Biology and Data Integration ... 24
5.1 Identification of Differential Expression Subnetworks ... 25
5.2 Identification of Co-expression Subnetworks ... 26
5.3 Identification of Differential Co-expression Subnetworks ... 27
6 Conclusion and Perspectives ... 28
References ... 29

1 Introduction

Mapping relationships between genotype and phenotype helps us to understand the biological mechanisms in an organism. The availability of high-throughput technologies allows us to study biomolecules (e.g., RNAs, proteins, metabolites) in a living cell as a whole under specified conditions. The measurements result in high-dimensional data of thousand of biomolecules (or variables) of a number of samples, which is usually much smaller than the number of biomolecules. In the last decade these data have been produced in vast amounts and the analysis is far from easy. Bioinformatics approaches play a key role in analyzing such data to extract biological information, leading to a better understanding of molecular processes.

Gene expression data are one of the most popular and has been utilized for studying cellular systems at the transcriptional level, also known as transcriptomics. Well-known techniques to measure RNAs are microarray and RNA sequencing. Thousands of transcripts could be measured in one sample. Several bioinformatics methods have been developed to analyze the expression data. Similar principles, especially in downstream analyses, have also been applied to analyze other omics data. The typical approach of expression analysis is to identify differential expressed genes between two conditions of interest. The approach provides a ranked list of differentially expressed genes, where each gene is individually tested for significant difference. Functional enrichment of these genes can then be performed to provide biological context for interpretation [1, 2].

Network analysis is another way to study biological system. The approach facilitates studies of biomolecules and their interactions, which could be physical interactions, functional relations, and/or co-regulations. In network analysis, a system of biomolecules could be represented as a graph where a node is a biomolecule (e.g., DNA, RNA, protein, and metabolite) and an edge or a link between nodes is an interaction between them. The methods mainly try to identify distinct modules or subnetworks driving a common biological process [3]. These modules are extracted from the resulting responses of perturbed systems [4].

Network analysis can complement typical expression analysis as it provides more information on the relations between biomolecules. Rather than getting a ranked list of individual differentially expressed genes, functional modules of interacting biomolecules (or genes if expression data is used) can be identified. For example, differentially expressed subnetworks could be detected, representing modules containing differentially expressed genes and their interactions. In addition, by using the network-based approach, genes that are not individually differentially expressed but are interacting with differentially expressed genes and still important for the process can be detected.

Conversely, expression data can provide context of functional links between biomolecules in interactome networks. By overlaying expression data on interaction networks, for example, protein-protein interactions (PPI) networks, regulatory networks, and metabolic networks, the functional relations between biomolecules can be revealed based on expression patterns under the studied conditions. Thus, both expression and network analyses complement each other to study biological processes. Even though this chapter mainly discusses the integration between expression data and interaction networks, other types of genomics and other omics data can also be integrated to study biological system as a network.

The concept and different types of interactome networks are first described including techniques to detect the interactions. Then the principle of graph theory important for biological networks analysis is introduced. Here different bioinformatics approaches to utilize expression data in network analysis are explained. This includes integration of expression data with other interactome networks and example studies. Finally, the perspective of expression and network analyses are discussed.

2 Interactome Networks

It has been accepted that biomolecules in a living organism work together and interact with each other [5]. The word interactome refers to a whole set of interactions between biomolecules within a cell. In some contexts, interactome specifically refers only to physical interactions. There are several types of interactions, for example, protein–protein interactions, protein–DNA interactions, and RNA interactions [6]. These biomolecules and their interactions can be represented as networks or graphs where nodes are biomolecules and edges (i.e., links between nodes) are interactions between these biomolecules.

Nodes and edges in biological networks represent different types of biomolecules and their interactions according to the types of the networks. Here different types of interactome networks are briefly explained, namely protein–protein interaction networks, regulatory networks, and metabolic networks, respectively. The contents include experimental techniques to study the networks and available databases containing the interactome data. The databases of these interactions are invaluable resources, facilitating analyses and studies of biology as a linking system. Functional interaction networks are also described, being inferred from expression data and providing functional relations between biomolecules.

2.1 Protein–Protein Interaction Networks

Protein–protein interaction (PPI) networks describe proteins and their interactions in a cellular system. The networks contain information on how proteins operate with each other to enable the biological process. In PPI networks, nodes represent proteins and edges represent physical interactions between two proteins.

One popular approach to detect PPI is yeast two-hybrid (Y2H) [7]. The method detects physical interaction between two protein pairs. Several projects have utilized Y2H technologies to construct the PPI maps, mainly in model organisms, for example, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, including *Homo sapiens* [8–12]. Another approach is affinity purification followed by mass spectrometry (AP-MS), which isolates protein complexes and identifies the constituents of the identified complexes, respectively [13]. Comprehensive efforts to generate PPI networks have been developed. The AP-MS technique is mainly used in unicellular organisms such as yeast [14] and mycoplasma [15], whereas Y2H is implemented in both unicellular and complex multicellular organisms [16].

With the availability of experimental techniques and a large amount of PPI data, several databases have been constructed to collect and provide such interaction data (Table 1). There has been much effort to curate PPIs data from the literature and collect it into databases [17–25]. Some databases provide the interactions specifically for organisms such as drosophila [22] and humans [19, 20]. STRING [26] is a more unique database and contains both experimental and predicted PPIs. The predicted interactions from computational methods for both physical and functional interactions are, for example, from knowledge transfer between organisms, from interactions aggregated from other databases, and from functional association. Despite these intensive efforts, the PPI data are still incomplete and require further investigation and systematic ways of detection [8]. The completeness of the data is an invaluable factor for studying living organisms.

2.2 Regulatory Networks

Regulatory networks contain information about the regulation of biomolecules in a cellular system such as regulation of gene expression, post-translational modification, and regulation by RNAs. This type of network is represented by a directed graph as the relations between nodes show the directions of regulation (see the graph definition in the next section).

Table 1 Interactome databases

Databases	URLs	Reference
Protein–protein interaction networks		
Biological General Repository for Interaction Database (BioGRID)	http://thebiogrid.org/	[18]
Biomolecular Interaction Network Database (BIND)	http://binddb.org	[17]
CCSB Interactome Database	http://interactome.dfci.harvard.edu	[8]
Database of Interacting Proteins (DIP)	http://dip.mbi.ucla.edu/dip/	[24]
Drosophila Interactions Database (DroID)	http://www.droidb.org/	[22]
IntAct Molecular Interaction Database	http://www.ebi.ac.uk/intact/	[23]
Molecular Interaction Database (MINT)	http://mint.bio.uniroma2.it/mint/Welcome.do	[21]
MIPS mammalian protein–protein interaction database (MPPI)	http://mips.gsf.de/proj/ppi/	[25]
The Human Protein Interaction Database (HPID)	http://www.hpid.org	[19]
Human Protein Reference Database (HPRD)	http://www.hprd.org/	[20]
STRING	http://string-db.org/	[26]
Regulatory networks		
Gene regulatory networks		
JASPAR	http://jaspar.genereg.net/	[35]
TRANSFAC	http://www.biobase-international.com/gene-regulation	[36]
Universal PBM Resource for Oligonucleotide Binding Evaluation (UniPROBE)	http://thebrain.bwh.harvard.edu/uniprobe/	[34]
Post-translational modification networks		
NetPhorest	http://netphorest.info/	[39]
PhosphoNetworks	http://www.phosphonetworks.org/	[30]
PhosphoSitePlus	http://www.phosphosite.org/	[31]
Phospho.ELM	http://phospho.elm.eu.org/	[37]
Posttranslational Modification Database (PHOSIDA)	http://www.phosida.com	[38]
RNA networks		
miRecords	http://miRecords.umn.edu/miRecords	[41]
miRBase	http://www.mirbase.org/	[43]
miRDB	http://mirdb.org/miRDB/	[44]
TarBase	http://microrna.gr/tarbase	[40]
TargetScan	http://www.targetscan.org/	[42]
Metabolic networks		
Biochemical Genetic and Genomic knowledgebase (BiGG)	http://bigg.ucsd.edu	[50]
BioCyc	http://biocyc.org/	[48]

(continued)

Table 1 (continued)

Databases	URLs	Reference
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/	[49]
MetaCyc	http://metacyc.org/	[48]
metaTIGER	http://www.bioinformatics.leeds.ac.uk/metatiger/	[51]

For gene regulatory networks, nodes are either transcription factors or DNA regulatory elements, and edges are physical bindings of a transcription factor and a regulatory element. To map protein–DNA interactions, two main techniques have been utilized, which are yeast one-hybrid (Y1H) [27] and chromatin immunoprecipitation (ChIP) [28]. The ChIP technique can then be followed by microarray (ChIP-chip) or sequencing (ChIP-seq). Y1H and ChIP methods can be complementary. The Y1H approach can discover novel transcription factors relying on known regulatory regions, and the ChIP method can discover novel regulatory regions based on the availability of reagents specific to transcription factors [6].

Phosphorylation network is one of well-known networks for post-translational modifications [29, 30]. The relations in this type of network are interactions between kinases and their substrates. Mapping of kinases and phosphorylation sites are displayed in phosphorylation networks. The networks provide global insight into kinase-mediated signaling pathways, which open up opportunities to a better understanding of cellular signaling processes. Furthermore, other types of post-translational modifications, such as ubiquitination, acetylation and methylation, have also been studied [31].

RNA networks have been investigated [32, 33]. This network displays expression regulation by RNAs such as small non-coding microRNAs (miRNAs) and small interfering RNAs (siRNAs). The mapping between RNA–RNA interactions and RNA–DNA interactions can be demonstrated in the RNA networks such as interactions between miRNAs and their targets. Protein-RNA networks have also been revealed [33].

Various types of databases providing regulatory interactions are currently available (Table 1). For example, UniPROBE [34], JASPAR [35], and TRANSFAC [36] collect useful data of gene regulatory networks, mainly from ChIP experiments. Several databases provide the data of post-translational modifications [31, 37–39]. For RNA networks, some databases, such as miRecords and TarBase, contain experimentally supported miRNA-target interactions [40, 41], whereas some contain only predicted targets [42–44].

2.3 Metabolic Networks

Metabolic networks explain the system of biochemical reactions in a particular cell or organism [45]. There are two main graph types of metabolic network, a reaction graph and a substrate graph [46]. For a reaction graph, nodes represent enzymes and edges represent metabolites that are substrates or products of the enzymes. In a substrate graph, nodes are biochemical metabolites, and edges represent reactions converting one metabolite into another or enzymes that catalyze the reaction.

The metabolic networks seem to be the most likely comprehensive networks containing discovered biochemical pathways and reactions. However, the completion with curation of metabolic network maps is still required. This could be fulfilled by having full genome sequencing with gene annotation and experimental investigations [47]. Several databases provide information on metabolic mapping as shown in Table 1. They provide comprehensive information on metabolic pathway, metabolites, enzymes, and reactions [48–51].

2.4 Functional Interaction Networks

The three network types detailed above contain physical interactions or biochemical interactions representing scaffold information of cellular systems [6]. Another main type of network focuses on functional links between biomolecules. This type of network infers functional associations between biomolecules that contain patterns of molecular profiles, for example, gene expression profiles. Very many bioinformatics attempts have been carried out to study and infer the functional interactions [26, 52–54]. Gene expression data have been widely applied to detect such relationships. STRING [26] is one of the databases that provide functional interactions data. As mentioned above, the database covers both experimentally detected interactions and computational predicted interactions. For computational prediction, STRING infers interactions between two proteins by, for example, co-expression and gene context analyses.

Functional interaction networks are often integrated with other types of networks. The three types of networks described above (PPI, regulatory, and metabolic networks) containing physical interactions or biochemical interactions are used as a network scaffold. Functional links can be overlaid on the scaffold network to infer functional associated modules according to the data of the conditions of interest. For example, co-expression networks (i.e., networks having nodes as genes and edges as correlations between genes) can be integrated with PPI networks. Edges in a network can represent both co-expression and PPIs. This can reveal PPI subnetworks, which are also co-expressed under the studied conditions.

Furthermore, the genetic interactions, showing relation of mutations that are related to the studied phenotype, could also be studied. Even though the genetic interaction network is not focused in this chapter, it contains another type of information that could be integrated into other types of networks as well.

3 Graph Concepts for Network Analysis

To analyze the biological networks, the system of biomolecules can be represented in a graph format. In computer science or mathematics, a graph represents a structure that models pairwise relations between items. Graph theory has been applied to analyze biological system, revealing network features and biological properties. Here the graph concepts that have been widely used in network analysis are briefly explained.

3.1 Graph Definition

A graph $G = (V, E)$ consists of a set of V and E, where V represents a set of vertices or nodes and E represents a set of unordered pairs of distinct elements of V called edges (i.e., links between nodes). As described previously, in interactome networks, nodes represent biomolecules and edges represent interactions between them. This type of graph is known as a simple graph or *undirected graph* (Fig. 1a). Edges between nodes do not have a direction. This type of graph is used to demonstrate PPI networks and some functional interaction networks, for example, co-expression networks.

A *directed graph* (V, E) consists of a set of vertices V and a set of edges E that are ordered pairs of elements of V (Fig. 1b). In other words, an edge in a directed graph has a direction. The directed graph is used to describe regulatory networks and metabolic networks, displaying direction of regulation and sequential pathways, respectively.

In an undirected graph G, two nodes u and v are adjacent or *neighbors* if $e = \{v, u\}$ is an edge of G. The edge e connects u and v. The *degree* of a node in an undirected graph is the number of edges incident with it, or the number of neighbor nodes. For a directed graph, each node has two types of degree, which are *in-degree*

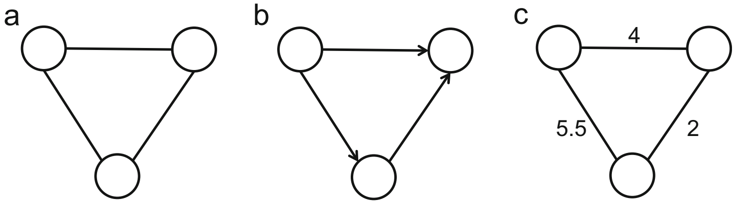


Fig. 1 Examples of (**a**) undirected, (**b**) directed, and (**c**) weighted graphs

and *out-degree*. For a given node in a directed graph, in-degree is the number of incoming edges and out-degree is the number of outgoing edges.

A *weighted graph* is a graph that has a number assigned to its edges (Fig. 1c). The weighted graph is often used in network analysis as a way of giving importance or confidence of interactions. For example, computational prediction interactions might be weighted with a smaller score than experimentally detection interaction [26].

A *subgraph* of a graph $G = (V, E)$ is a graph $H = (W, F)$ where $W \subseteq V$ and $F \subseteq E$. This concept is an essence for studying biological networks, as a molecular system is believed to be modular [3, 55]. A functional *module* is a subgraph or a *subnetwork* from a whole network that contains nodes having a joint role or a common function. The concept of identification of subnetworks is explained below.

3.2 Network Properties

Network topology, an arrangement of a network, often reveals information and characteristic of the network. *Scale-free topology* is usually found in biological networks [6, 16]. For a scale-free network, the degree distribution follows a power-law tail $P(k) \sim k^{\gamma}$, where k is a node degree (i.e., the number of links on a given node), $P(k)$ is the degree distribution, and γ is the degree exponent. The topology is often found in real world networks including biological networks [56], which contain a small number of highly connected hubs that hold the whole network together [57]. A *hub* is defined as a node that has a high degree. A scale-free network is different from a random network, where most nodes have approximately the same degree and the highly connected nodes or hubs are rare [56].

Hubs have been found to play an important role in biological systems. Hub proteins were found to be essential in a biological process, having slower evolution and yielding a larger diversity of resulting phenotypes with their deletions [57–59]. In addition, hubs have been found to arise as a disease-related gene, for example, in cancer [6].

Furthermore, hubs have been categorized into *date hubs* and *party hubs* [60]. Party hubs are highly co-expressed with their interacting partners and likely to interact with their partners at all times and under all conditions. Date hubs are more dynamically regulated relative to their partners and interact at different times and under different conditions [16]. In addition, date hubs and party hubs are sometimes called inter-module and intra-module hubs, respectively [61]. This is because date hubs seem to be functional modules connected to each other whereas party hubs appear to be inside functional modules.

These network properties have been discovered, leading to a better understanding of biological networks. Importantly, the concepts, especially scale-free topology and hubs, have been used for network analysis. For instance, Zhang and Horvath [52] proposed a computational method to identify functional modules based on scale-free topology assumptions.

4 Identification of Modules or Subnetworks

From large-scale interaction networks, it has been accepted that a cellular system is modular [3, 55]. The concept of functional modules has been introduced to provide meaningful interacting subgroups of biomolecules that share the same functions. Several computational methods for extracting informative subnetworks have been introduced. These methods are mainly based on topological or functional modularity of the networks, or both.

A *topological module* is defined as part of network that tends to link to a node within the same local neighborhood rather than to a node outside it. The methods search for subgraphs containing nodes that share locally dense neighbors. The topological modules are believed to carry specific functions. This also leads to the concept of the functional module, where nodes have related functions [62].

The concept of functional modularity is often considered to identify subnetworks. Many studies utilize molecular profiles, for example, expression profiles to reveal functionally related subnetworks. Often scoring and searching for high score/significant subnetworks are carried out. Different approaches define scores on nodes, edges, both nodes and edges, or some signal contents. It is known that searching for active modules is computational expensive. Heuristic approaches for searching to optimize computing time have been utilized, for instance, greedy algorithms [63, 64], simulated annealing [65], genetic algorithms [53], and exact methods [66, 67].

Another strategy identifies subnetworks by seeding node-containing genetic information related to the studied phenotype (e.g., it could be a disease-related gene). Both topological and functional modularity can be utilized by searching genes associated with the conditions of interest through interactome networks. The method aims to identify subnetworks that contain most of the condition-associated genes and have a compact structure of topological modules [62].

5 Expression Analysis Through Network Biology and Data Integration

As mentioned above, the typical approach of expression analysis is to identify a list of differential expressed genes between two conditions of interest. The differential expression of each individual gene is usually measured by using a statistical method to determine mean expression changes between the two conditions [68]. These genes are then ranked according to their differentially expressed scores and a significant set of genes that passes a cut-off criterion is considered. However, deriving only a ranked list of differentially expressed genes is still difficult for interpretation. Functional analyses such as mapping to biological pathway or gene set analyses (e.g., enrichment analysis [1] and Gene Set Enrichment analysis (GSEA) [2]) play a role to provide more meaningful biological results.

A network-based approach is one of the key approaches that could complement the typical differential expression analysis. The approach can identify a group of functionally related genes with their interactions. It is known that genes or biomolecules often work together, and detecting genes as a functional group (i.e., submodules or subnetworks) helps us to understand biology mechanisms and makes things easier to interpret. The approach can also identify genes that are not individually differentially expressed but are still important for dysregulation processes and interacting dysregulated genes. Furthermore, detecting genes as a module could enhance statistical power to detect differentially expressed submodules, even though individual genes might not be statistically significant [54].

Network-based approaches for utilizing expression data are described based on gene expression patterns and correlations. The methods can be performed based on expression data only to construct functional interaction networks. However, the methods are often applied and integrate expression data with other interactome networks. Usually the interactome networks are used as a network template and then overlaid with the expression data to identify functional modules of conditions of interests. With the power of data integration, the detected subnetworks reveal more layers of information, providing stronger evidence of the discovered physical interactions or reactions and a layer of correlated gene expression of the study. Even though the interactome and functional networks might not be from the same conditional experiments, integration analysis is still believed to provide some clues to study cellular systems at molecular levels.

Although the analysis of expression data or transcriptome analysis is focused upon in this chapter, it should be noted that the method is not limited just to this level. The same principles can be applied to analyze the data at other levels, for example, proteome and metabolome. In addition, more than one type of interaction network could be integrated, depending on the biological question and expected output of the study.

5.1 Identification of Differential Expression Subnetworks

The main idea of an approach for identification of differentially expressed subnetworks is to identify subnetworks/modules containing connections of differentially expressed genes. These approaches require computational methods for scoring and searching for candidate modules. The methods measure a significance of differentially expressed genes as a connected set. The measurement relies on scoring of nodes and sometimes their connectivity. Aggregation of scoring subnetworks of linking differentially expressed genes can be computed. Various searching methodologies have been applied to identify high score candidate modules of differentially expressed genes.

One of the very first examples of scoring-based methods to identify active subnetworks has been developed by Ideker et al. [65]. They integrated protein–

protein and protein–DNA interactions with expression data, and identified connected subnetworks showing significant changes in expression over particular subsets of conditions. The method combines measurement of scoring subnetworks with searching algorithm to find high score subnetworks. The algorithm, jActiveModules, is also provided as a Cytoscape [69] plugin.

Another example shows the application of differentially expressed subnetworks for classification purposes. Chuang et al. [63] integrated PPI network and expression data to identify differentially expressed subnetworks that give high discrimination power between metastatic and non-metastatic in breast cancer. They overlayed expression profiles between the two states of cancer on a PPI network. The scoring differentially expressed subnetworks according to their discrimination power and was measured and searched using a greedy algorithm. The identified subnetworks were used for the disease classification.

Other than the scoring-based approach mentioned above, another approach for identification of differential expressed subnetworks is a set cover-based approach [54]. The methods have proved to be successful in capturing heterogeneity among patients in complex diseases such as cancer [70, 71]. The set cover-based methods take into account connected sets of genes that are significantly enriched with genes that are differentially expressed in samples of the disease. The method is based on a concept that each disease sample has some differentially expressed genes, and in heterogeneous diseases different samples have different covering genes. A gene is considered to cover a disease sample if it is differentially expressed in the sample. The methods search for a representative set of connected covering genes.

An example of set cover-based approach is the study of Ulitksy et al., named DEGAS (DysrEgulated Gene set Analysis via Subnetworks) [70]. They integrated expression data and interaction networks. The method aims to find the smallest subset of genes covering disease samples on connected subnetworks. The method was demonstrated in analyzing human diseases such as Parkinson's disease, Alzheimer's disease, and cancer.

5.2 *Identification of Co-expression Subnetworks*

Instead of focusing on differential expression, several approaches consider the co-expression pattern or expression correlation between genes. If the expression changes between two genes are correlated (or anti-correlated), it may be assumed that the two genes have functional relations. In this type of network, a link between nodes indicates an expression correlation between them. The correlation between genes in each sample group can be measured using statistical methods, usually Pearson's correlation. The method measures a linear correlation between two genes, giving a correlation coefficient between +1 and -1, where +1 is a total positive correlation, 0 indicates no correlation, and -1 is a total negative correlation or anti-correlation. One could define only correlated subnetworks and/or anti-correlated subnetworks. However, several studies consider subnetworks with both types of correlation. In this case, a correlation value could be considered as an absolute value of correlation coefficient ranging from 0 to 1. The network can be represented as a weighted graph having edges as correlation values.

An important issue in analyzing correlation-based networks is to determine a threshold for drawing a link of correlation between genes' hard and soft thresholds. In the hard threshold criterion, a link between nodes is determined to be either 1 or 0, connected or unconnected. A threshold is set to define a link, usually based on statistical significance of a correlation between nodes across samples. However, it has been questioned whether the binary information is meaningful enough to encode a biological network. Instead of using a binary value, another method called soft threshold suggested a way to weight a correlation having a value range in [0,1]. The method has been found to provide more robust results [52] with more information than using hard threshold.

One example of a method to identify weighted co-expression subnetworks is the study of Zhang and Horvath [52]. They proposed a soft threshold criterion as an adjacency function converting co-expression measure to a connection weight. Parameters of the adjacency function (e.g., power or sigmoid functions) were determined by using scale-free topology criteria. The chosen parameter values should lead to scale-free topology networks, as they have been known to provide meaningful biological results. The method is also implemented and provided in an R package [72].

5.3 Identification of Differential Co-expression Subnetworks

Other than focusing on only differentially expressed genes/nodes or co-expression patterns between nodes, another approach focuses on the changing of co-expression patterns between nodes. The approach is called differential co-expression analysis, searching for loss and gain of correlations in different states. An edge in this type of network indicates a change of correlation between two genes in different conditions. Several scenarios could be counted as differential co-expression between genes, for example, two genes both have correlation in each sample group but different sign (correlated and anti-correlated), and one gene has a correlation in one sample group but no correlation in another, or vice versa. For instance, if we study expression patterns between healthy and disease samples, gene *a* might have a positive correlation with gene *b* in healthy samples but both genes have a negative correlation in disease samples, or vice versa (a negative correlation in healthy samples, but a positive correlation in disease samples). Another scenario might be that gene *a* might have a positive (or negative) correlation with gene *b* in healthy samples but there is no correlation among them in disease samples, or vice versa.

The differential co-expression analysis itself has been shown to complement differential expression analysis [73]. Some known transcriptional regulators involved in cancer appeared to be not significantly differently expressed but were highly differentially co-expressed [61, 74, 75]. Hudson et al. showed an example of identification of casual mutations and perturbations using expression data by contrasting network connections and examining regulators in the network changes [74]. They demonstrated the method in microarray data between two stages of myostatin mutation. The gene did not significantly differ as its regulation is post-translational; however, the gene happened to be in a top rank among transcriptional regulators when considering differential co-expression. The method was suggested as another way to identify important transcription factors which might be overlooked by differential expression analysis.

Another example of a network-based method for considering differential co-expression is Interactome Dysregulation Enrichment Analysis (IDEA) [76]. Mani et al. developed an approach to identify genes enriched for perturbed interactions displaying changes in co-expression patterns. They integrated interactome (predicted protein–protein, protein–DNA interactions and post-translational modifications) with expression data of B-cell lymphomas. By utilizing the expression data, network edges of gain and loss of correlations were drawn. Genes were scored according to the enrichment of the perturbed edges. They demonstrated the identification of known oncogenic lesions and downstream effectors for three malignant B-cell phenotypes.

6 Conclusion and Perspectives

Integration of expression data and interactome networks provide a very powerful tool to study biological systems. The interactome networks display relationships between biomolecules, and expression data show contexts of gene expression changes. The patterns of gene expression could be inferred for functional links between biomolecules under the studied conditions.

Most of the network strategies have been employed for describing only static and partial snapshots of biological systems at different times and conditions. However, it is known that the biological systems are dynamic. Currently, several approaches are being developed, moving toward the strategy to study the dynamic of cellular system. The concept is known as differential network [77]. The strategy tries to capture dynamic re-wiring of cellular stages. For example, nodes could contain information of the changes, for example, different intensity of gene expression, and only edges showing changes between different stages could be drawn. A number of algorithms and tools for differential molecular networks have been developed [78, 79]. The studies via differential networks provide a promising way to study biological systems. A strategy that provides more details to explain biological system would lead to a greater understanding of the living cell.

References

1. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57
2. Subramanian A et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43): 15545–15550
3. Mitra K et al (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14(10):719–732
4. Markowetz F (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. PLoS Comput Biol 6(2):e1000655
5. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
6. Vidal M, Cusick ME, Barabasi AL (2011) Interactome networks and human disease. Cell 144(6):986–998
7. Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340(6230):245–246
8. Rolland T et al (2014) A proteome-scale map of the human interactome network. Cell 159(5): 1212–1226
9. Ito T et al (2000) Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A 97(3):1143–1147
10. Giot L et al (2003) A protein interaction map of Drosophila melanogaster. Science 302(5651): 1727–1736
11. Li S et al (2004) A map of the interactome network of the metazoan C. elegans. Science 303(5657):540–543
12. Walhout AJ, Vidal M (2001) Protein interaction maps for model organisms. Nat Rev Mol Cell Biol 2(1):55–62
13. Rigaut G et al (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17(10):1030–1032
14. Collins SR et al (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 6(3):439–450
15. Kuhner S et al (2009) Proteome organization in a genome-reduced bacterium. Science 326(5957):1235–1240
16. Seebacher J, Gavin AC (2011) SnapShot: protein-protein interaction networks. Cell 144(6): 1000, 1000 e1
17. Bader GD, Betel D, Hogue CW (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31(1):248–250
18. Chatr-Aryamontri A et al (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43(Database issue):D470–D478
19. Han K et al (2004) HPID: the human protein interaction database. Bioinformatics 20(15): 2466–2470
20. Keshava Prasad TS et al (2009) Human protein reference database—2009 update. Nucleic Acids Res 37(Database issue):D767–D772
21. Licata L et al (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40(Database issue):D857–D861
22. Murali T et al (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. Nucleic Acids Res 39(Database issue):D736–D743
23. Orchard S et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(Database issue):D358–D363
24. Salwinski L et al (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32(Database issue):D449–D451

25. Pagel P et al (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21(6):832–834
26. Szklarczyk D et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452
27. Ouwerkerk PB, Meijer AH (2001) Yeast one-hybrid screening for DNA-protein interactions. Curr Protoc Mol Biol Chap 12:Unit 12. 12
28. Nelson JD, Denisenko O, Bomsztyk K (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat Protoc 1(1):179–185
29. Newman RH et al (2013) Construction of human activity-based phosphorylation networks. Mol Syst Biol 9:655
30. Hu J et al (2014) PhosphoNetworks: a database for human phosphorylation networks. Bioinformatics 30(1):141–142
31. Hornbeck PV et al (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 43(Database issue):D512–D520
32. Sethupathy P, Corda B, Hatzigeorgiou AG (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. RNA 12(2):192–197
33. Lapointe CP et al (2015) Protein-RNA networks revealed through covalent RNA marks. Nat Methods 12(12):1163–1170
34. Hume MA et al (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res 43(Database issue):D117–D122
35. Mathelier A et al (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44(D1):D110–D115
36. Matys V et al (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31(1):374–378
37. Dinkel H et al (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res 39(Database issue):D261–D267
38. Gnad F, Gunawardena J, Mann M (2011) PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res 39(Database issue):D253–D260
39. Miller ML et al (2008) Linear motif atlas for phosphorylation-dependent signaling. Sci Signal 1(35):ra2
40. Papadopoulos GL et al (2009) The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 37(Database issue):D155–D158
41. Xiao F et al (2009) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37(Database issue):D105–D110
42. Garcia DM et al (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 18(10):1139–1146
43. Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42(Database issue):D68–D73
44. Wong N, Wang X (2015) miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res 43(Database issue):D146–D152
45. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18(3): 326–332
46. Jeong H et al (2000) The large-scale organization of metabolic networks. Nature 407(6804): 651–654
47. Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320
48. Caspi R et al (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 44(D1):D471–D480
49. Kanehisa M et al (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462

50. Schellenberger J et al (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11:213
51. Whitaker JW et al (2009) metaTIGER: a metabolic evolution resource. Nucleic Acids Res 37(Database issue):D531–D538
52. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:Article17
53. Klammer M et al (2010) Identifying differentially regulated subnetworks from phosphoproteomic data. BMC Bioinformatics 11:351
54. Cho DY, Kim YA, Przytycka TM (2012) Chapter 5: network biology approach to complex diseases. PLoS Comput Biol 8(12), e1002820
55. Hartwell LH et al (1999) From molecular to modular cell biology. Nature 402(6761 Suppl): C47–C52
56. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439): 509–512
57. Jeong H et al (2001) Lethality and centrality in protein networks. Nature 411(6833):41–42
58. Fraser HB et al (2002) Evolutionary rate in the protein interaction network. Science 296(5568): 750–752
59. Yu H et al (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322(5898):104–110
60. Han JD et al (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995):88–93
61. Kostka D, Spang R (2004) Finding disease specific alterations in the co-expression of genes. Bioinformatics 20(Suppl 1):i194–i199
62. Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12(1):56–68
63. Chuang HY et al (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140
64. Nacu S et al (2007) Gene expression network analysis and applications to immunology. Bioinformatics 23(7):850–858
65. Ideker T et al (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(Suppl 1):S233–S240
66. Backes C et al (2012) An integer linear programming approach for finding deregulated subgraphs in regulatory networks. Nucleic Acids Res 40(6), e43
67. Dittrich MT et al (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24(13):i223–i231
68. Klebanov L et al (2007) Statistical methods and microarray data. Nat Biotechnol 25(1):25–26, author reply 26–7
69. Shannon P et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
70. Ulitsky I et al (2010) DEGAS: de novo discovery of dysregulated pathways in human diseases. PLoS One 5(10), e13367
71. Chowdhury SA, Koyuturk M (2010) Identification of coordinately dysregulated subnetworks in complex phenotypes. Pac Symp Biocomput 133–144. doi:10.1142/9789814295291_0016
72. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559
73. de la Fuente A (2010) From ‘differential expression’ to ‘differential networking’—identification of dysfunctional regulatory networks in diseases. Trends Genet 26:326–333
74. Hudson NJ, Reverter A, Dalrymple BP (2009) A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. PLoS Comput Biol 5(5):e1000382
75. Carter SL et al (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20(14):2242–2250
76. Mani KM et al (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol 4:169

77. Ideker T, Krogan NJ (2012) Differential network biology. Mol Syst Biol 8:565
78. Van Landeghem S et al (2016) Diffany: an ontology-driven framework to infer, visualise and analyse differential molecular networks. BMC Bioinformatics 17:18
79. Ruan D, Young A, Montana G (2015) Differential analysis of biological networks. BMC Bioinformatics 16:327

Adv Biochem Eng Biotechnol (2017) 160: 33–56
DOI: 10.1007/10_2016_41

Published online: 28 July 2016

Rigid-Docking Approaches to Explore Protein–Protein Interaction Space

Yuri Matsuzaki, Nobuyuki Uchikoga, Masahito Ohue, and Yutaka Akiyama

Abstract Protein–protein interactions play core roles in living cells, especially in the regulatory systems. As information on proteins has rapidly accumulated on publicly available databases, much effort has been made to obtain a better picture of protein–protein interaction networks using protein tertiary structure data. Predicting relevant interacting partners from their tertiary structure is a challenging task and computer science methods have the potential to assist with this. Protein–protein rigid docking has been utilized by several projects, docking-based approaches having the advantages that they can suggest binding poses of predicted binding partners which would help in understanding the interaction mechanisms and that comparing docking results of both non-binders and binders can lead to understanding the specificity of protein–protein interactions from structural viewpoints. In this review we focus on explaining current computational prediction methods to predict pairwise direct protein–protein interactions that form protein complexes.

Y. Matsuzaki (✉)
Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Tokyo, Japan
e-mail: matsuzaki@acls.titech.ac.jp

N. Uchikoga
Department of Physics, Faculty of Science and Engineering, Chuo University, Tokyo, Japan

M. Ohue
Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, Japan

Y. Akiyama
Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, Japan

Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Tokyo, Japan
e-mail: akiyama@c.titech.ac.jp

Keywords Protein docking, Protein–protein interaction, Supercomputing

Contents

1 Introduction 34
1.1 Protein–Protein Interaction Network 34
1.2 Computational Methods to Predict Pairwise Direct PPIs 35
1.3 Exploiting Protein Docking for PPI Prediction 36
2 Computational Protein–Protein Docking 37
3 Computational PPI Prediction 40
3.1 PPI Prediction Using Protein Docking 41
3.2 PPI Prediction Without Docking 42
4 Profile Methods in Post-Docking Processes 43
5 Implementation of Docking Software on Supercomputing Environments 45
6 Conclusion 47
References 47

1 Introduction

1.1 Protein–Protein Interaction Network

Protein–protein interactions (PPI) play crucial roles in living cells such as signal transduction and regulation of metabolic pathways. Information on proteins has rapidly accumulated on publicly available databases. There are more than 10 million non-redundant protein sequences in the database UniProt [1]. Protein Data Bank (PDB) (http://www.rscb.org) [2] stores 105,849 structural data of proteins corresponding to 29,824 protein clusters, each of which has 95% sequence identity (accessed 25 November 2015).

Proteins exert their function by interacting with other molecules. For protein–protein interactions there are combinatorial numbers of possible interactors. This information is available from public databases such as BioGRID (56,086 gene products, 415,624 non-redundant, physical interactions) [3], DIP (28,215 proteins, 80,286 interactions) [4], HPRD (30,047 proteins, 41,327 interactions, accessed 25 November 2015) [5], IntAct (89,430 interactors, 564,831 interactions) [5, 6], and MINT (35,553 proteins, 241,458 interactions) [7] (note: all database statistics were obtained on 25 November 2015). There are continuous efforts to compile and analyze the interactome of each species. For example, Schwikowski et al. [8] assembled a *Saccharomyces cerevisiae* direct protein–protein interactome from published interactions and constructed a large protein network of 2,358 interactions for 1,548 proteins. A number of studies reported assessments of model organism PPIs such as yeast and humans [9] [10]. Considering the numerous interactions among multiple proteins, our current knowledge of PPI networks is nowhere near complete, with many novel possible interactions not yet discovered.

A better picture of PPI networks has many potential applications. PPI networks can provide insights into mechanisms governing various cellular processes, and PPI networks have also been used to estimate important modules related to diseases and lethality [11].

PPI networks are also investigated as inter-species networks, for example, virus–host protein interactions. Franzosa and Xia analyzed available virus–host protein complex interface structures and showed that inter-species PPI networks have distinct structural, functional, and evolutionary principles from the within-host PPI network [12]. A combined network of host–pathogen PPIs was analyzed by Rachita and Nagarajaram to show that viruses not only tend to target bottlenecks, hubs, and rich clubs of host PPI networks but also make use of peripheral nodes of host networks by bridging them to larger host network components, thus realizing virus-specific use of host machinery [13]. Such inter-species PPI network analyses would contribute to the understanding of disease-causing mechanisms produced by a variety of pathogens in addition to viruses.

In general, a protein–protein interaction network represents two types of protein organizations: protein complexes and functional modules. Functional modules are the group of proteins that contribute to a specific cellular process. Proteins in the same functional group can be in the different location, and timing of expression may differ. Because of the lack of pairwise protein interaction data, discrimination of protein complex and functional modules is a difficult task. In this review we focus on pairwise direct protein-protein interactions that form protein complexes.

1.2 Computational Methods to Predict Pairwise Direct PPIs

Computational approaches to predict pairwise direct PPIs have been developed using various types and levels of information [14]. Major sources of information include sequence homology, gene ontology, and gene co-expression. Machine-learning approaches utilize protein features such as amino acid sequence-based features and physicochemical features as input. Structural information was also shown to be useful to improve accuracy of the prediction [15]. The two approaches mentioned are examples of powerful methods that (1) use a known interaction surface structure as ‘template’ and construct a model of interaction by superimposing query protein pair structures and then (2) use the information of how well the model fits the known interaction surfaces to evaluate the potential of interaction. The PRISM protocol evaluates the well-fitting pairs by consecutive soft docking and predicts the input pair’s binding possibility by using the docking score [16]. Another method called PrePPI combines the superimposed model evaluations to other non-structural features such as co-expression and functional and evolutionary similarities by Bayesian classification and evaluates the possibility of PPIs [17].

1.3 Exploiting Protein Docking for PPI Prediction

Although these heuristic methods show good prediction power, other emerging approaches using de novo docking to predict whether input of two proteins have the potential to interact, which does not directly use information of known complex structures, may have notable advantages. A protein–protein docking approach searches the entire surface of each of the target protein pair structures for presumable binding sites and then it outputs docking models with the docking score, usually evaluated by shape complementarity and physicochemical features for each model. Possibilities of interaction of the input of two proteins are evaluated by several aspects, such as the docking score of the top-rated model and docking score distribution [18–20].

An advantage of a docking-based method over a template-based method is that it searches entire surfaces of target proteins, and thus it has the potential to discover novel PPIs with currently unknown interacting surfaces.

A second advantage is that we can obtain and utilize the information of so-called 'decoys' (false docking poses) generated by docking calculations. A typical use of docking decoys may be for binding site prediction [21, 22]. Assuming that the high-scoring decoys are seen frequently near the true binding site, the intensity of high-scoring decoys is incorporated to predict binding sites for protein interactions. Information regarding decoys has been exploited for several purposes. Wass et al. suggested that decoy interaction surfaces might provide information regarding binding partners [19]. They successfully discriminated binders from non-binders by using high averaged docking scores of 20,000 best scoring models for prediction. They discussed whether it might reflect a concept of the binding process of two proteins; the binding is initiated by the formation of nonspecific complexes followed by rearrangements of them to more stable and specific interactions. Based on this concept, binders may yield high docking scores not only for the final binding surfaces but also for the non-binding surfaces. Torchara et al. adopted an approach related to this idea to distinguish the near-native complex structures among generated docking decoys [23]. They assumed the funnel-like interaction energy distribution for the binding of two proteins and built a post-docking evaluation model to find near native poses using a Markov model, which states that transition probabilities are defined by docking score differences.

On a more general note, the protein docking community has been adding criteria of docking score optimization to improve correlation to binding affinities. Traditionally, the docking score was not necessarily reflective of the actual binding energy. It was used to rank the docking decoys in each docking process. Docking methods were evaluated based on whether the nearby native complex structure was obtained in high ranked decoy sets. However, a recent study performed on multiple docking and re-ranking methods and to reevaluate high scored decoys based on more detailed calculations than in the initial docking showed that, although no scoring methods had strong correlation with experimentally determined binding affinities, some methods were capable of categorizing each protein pair to three categories of binding strength (high, medium, and low) [24]. This study led to the proposal of a new dataset of protein pairs and their binding affinity values as a protein affinity benchmark, which was later integrated with the most widely known protein docking benchmark dataset [25]. Another recent study proposed a regression model of input protein pair features and their binding affinity that does not require protein tertiary structures [26], which is useful for large-scale predictions.

High-scoring models generated by a protein docking calculation can be a useful data source, as a sampling of probable transient binding modes and functional, 'correct' binding modes would help elucidate the protein–protein interaction mechanism and specificity. Developments of such analysis methods are still in their early stages.

Docking-based approaches have several limitations. First, the target protein's tertiary structure is required, which would narrow down the target from the whole interactome space. Some studies tried to overcome this by employing homology modeling [27]. However, this approach cannot be applied to intrinsically disordered regions, which are considered to be important for the specificity of PPIs. The second difficulty is that it involves expensive calculation costs compared to other approaches without docking. For this reason, less time-consuming rigid-docking approaches that do not consider protein flexibility have been used for this purpose. Moreover, docking tools used for large-scale PPI analyses have been developed to run effectively in massive parallel computing environments [28, 29].

Even with these limitations, given that there are already widely available protein tertiary structure data, it seems to be an interesting direction of the PPI research field to exploit protein docking for large-scale PPI studies. In this chapter we mainly focus on the application of rigid docking in the context of PPI network predictions.

2 Computational Protein–Protein Docking

Protein–protein interactions can provide valuable insights for understanding the principles of biological systems and for elucidating the causes of incurable diseases. Although many structures of interacting proteins have been determined by X-ray crystallography and NMR spectroscopy, there are still many protein complexes undetermined experimentally because of cost and experimental limitations. Protein–protein docking is a computational method for predicting the structure of a protein complex from known component structures and is a powerful approach that can result in otherwise unattainable discoveries.

In the field of protein complex structure prediction with computational protein–protein docking, users have a variety of choices of methods and must consider the trade-off of computing time and the prediction power. Furthermore, users must consider whether using known crystalline complex structures as templates (template-based method) or not (non-template-based, template-free, or de novo methods) is best for their analyses. InterPreTS [30], 3D-Partner [31], HOMCOS [32], and Interactome3D [33] are the template-based prediction tools whose templates are built using the complete surface area of known complex structures. Input protein structures are then examined by searching structural similarity with all the data in the template dataset. On the other hand, PRISM [16, 34, 35] uses only the interface structure to build a template dataset. By contrast, in the de novo prediction methods, which do not require template data, users start with the simulated 3D structures of the two unbound component proteins. Assuming that the complex formed has limited conformational changes, the target two-protein structures are regarded as rigid bodies, and a 3D rotational and 3D translational search (6D search) is performed over all possible associations. Then a re-ranking of the resultant complexes may be undertaken, possibly using computationally more intensive calculations. Conformational flexibility may be introduced into the algorithm to refine the few remaining candidates when there are only a limited number of complexes to consider. A method that explicitly introduces structural flexibility also exists and is called soft docking. Typical tools and publications on protein docking tools are shown in Table 1.

Although various protein–protein docking methods have been developed, de novo docking is the most mainstream one, and the initial sampling of conformations by the rigid-body search, which is one of the de novo docking methods, is employed

Table 1 Representative computational protein–protein docking (complex structure modeling) tools

Template-based docking
Homology-based modeling:
InterPreTS [30], 3D-Partner [31], HOMCOS [32], Interactome3D [33]
Using interface structure template:
PRISM [16, 34, 35]
De novo docking
Exhaustive search with FFT (fast Fourier transform):
ZDOCK [36–39], MEGADOCK [28, 29, 40, 41], PIPER/ClusPro [42–44], FTDock [45], SDOCK [46], GRAMMX [47, 48], MolFit [49, 50], F^2DOCK [51], DOT [52, 53], ASPDock [54]
Exhaustive search with spherical Fourier transform:
Hex [55–57], FRODOCK [58]
Local matching search:
LZerD [59, 60], PatchDock [61], CS [62], shDock [63], SP-Dock [64]
Monte Carlo search:
HADDOCK [65, 66], RosettaDock [67–70]
Randomized search with swarm intelligence:
SwarmDock [71, 72]
Post-docking analysis
Structure refinement:
FiberDock [73], EigenHex [74], RDOCK [75], FireDock [76]
Docking pose rescoring:
ZRANK [77, 78], SIPPER [79], SPIDER [80], pyDock [81, 82], DARS [83], PIE [84, 85], GB-Rerank [86]¸ BACH-SixthSense [87]
Other reranking methods:
PISA [88], CyClus [89], CONSRANK [90, 91], IFP [92], DockRank [93]
Integrated server:
CCharPPI [94]

in most cases as an important process. A number of algorithms and many different scoring functions have been developed in the last 20 years, as recently reviewed by Eisenstein and Katchalski [95], Ritchie [96], Janin [97], Vakser [98], Vajda et al. [99], and Huang [100]. The scoring functions for protein–protein conformation searches and the methods for rescoring candidate structures generated by the initial docking have been reported by Moal et al. [101] and Vajda et al. [99]. Moal et al. also performed a large-scale comparative study [102] that serves as a powerful informative guide to choose suitable methods. The various template-based methods have been reviewed by Szilagyi and Zhang [103]. Details of the comparison of template-based and de novo docking have been mentioned by Vreven et al. in 2014 [104]. The applicability of the template-based methods has been discussed by Kundrotas et al. [105] and Negroni et al. [106].

Figure 1 illustrates the de novo docking procedure using the fast Fourier transform (FFT)-based exhaustive search algorithm. In this method, as in FFT-based docking software such as ZDOCK, MEGADOCK, and PIPER, the protein structure is projected onto a 3D grid space $\mathbf{N}^3$, and the scoring function is calculated by discrete Fourier transform (DFT) and inverse discrete Fourier transform (DFT^{-1}) using the correlation of two discrete functions (protein grids), as follows:

$$S(\mathbf{t}) = \sum_{\mathbf{v}\in\mathbf{N}^3} R(\mathbf{v}) \times L(\mathbf{v}+\mathbf{t}) = \mathrm{DFT}^{-1}\left[\mathrm{DFT}[R(\mathbf{v})]^* \times \mathrm{DFT}[L(\mathbf{v})]\right]$$

where $R(\mathbf{v})$ and $L(\mathbf{v})$ are the discrete functions of the receptor (R) and ligand (L) proteins, respectively, $\mathbf{v} = (l, m, n)$ is a coordinate in the 3D grid space $\mathbf{N}^3$, and $\mathbf{t} = (\alpha, \beta, \gamma)$ is the parallel translation vector of the ligand protein. The asterisk operator

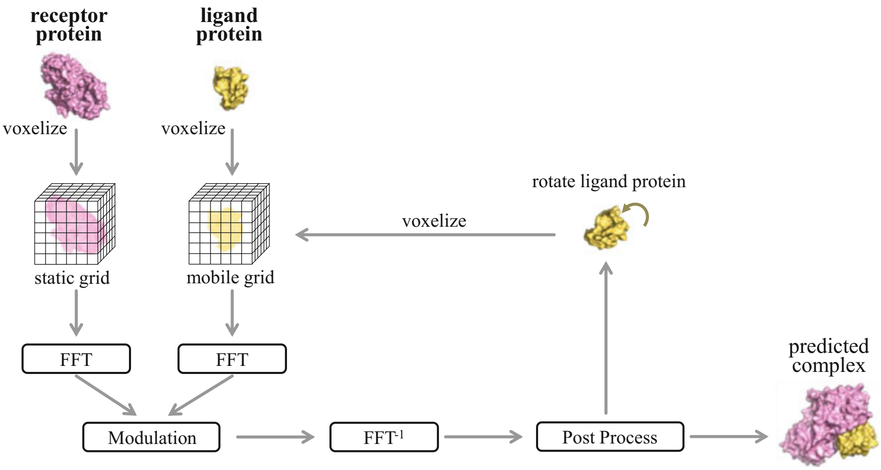


Fig. 1 Typical de novo protein–protein docking procedure using the FFT-based exhaustive search algorithm

$*$ indicates the complex conjugate of a complex number. To execute directly the simple convolution sums in $S(\mathbf{t})$ per ligand rotation pattern, $O(N^6)$ calculations are required; however, this is reduced to $O(N^3 \log N)$ using the FFT as DFT.

The discrete functions R and L usually take into account multiple effects, such as shape complementarity, electrostatic interaction, and desolvation free energy (e.g., ZDOCK [36, 37], PIPER [43], and MEGADOCK [40]). The total scoring function is the weighted sum of the partial scoring functions, according to the following example:

$$\begin{aligned}
S_{\text{total}}(\mathbf{t}) &= w_{\text{shape}} S_{\text{shape}}(\mathbf{t}) + w_{\text{elec}} S_{\text{elec}}(\mathbf{t}) + w_{\text{desol}} S_{\text{desol}}(\mathbf{t}) \\
S_{\text{shape}}(\mathbf{t}) &= \text{DFT}^{-1}\left[\text{DFT}\left[R_{\text{shape}}(\mathbf{v})\right]^* \times \text{DFT}\left[L_{\text{shape}}(\mathbf{v})\right]\right] \\
S_{\text{elec}}(\mathbf{t}) &= \text{DFT}^{-1}\left[\text{DFT}[R_{\text{elec}}(\mathbf{v})]^* \times \text{DFT}[L_{\text{elec}}(\mathbf{v})]\right] \\
S_{\text{desol}}(\mathbf{t}) &= \text{DFT}^{-1}\left[\text{DFT}[R_{\text{desol}}(\mathbf{v})]^* \times \text{DFT}[L_{\text{desol}}(\mathbf{v})]\right]
\end{aligned}$$

In this example, the total scoring function is calculated based on three correlation functions. In actuality, the desolvation free energy function S_{desol} also often comprises multiple correlation functions. For example, ZDOCK uses six correlation functions and PIPER uses nine for the calculation of S_{desol}. In general, computational time required for docking increases with the number of correlation functions. To enable faster calculation, MEGADOCK employs a score function that requires only one correlation function by compressing three terms (shape complementarity, electrostatic interaction, and desolvation free energy) into one correlation function. The total scoring function is represented by the functions as follows:

$$\begin{aligned}
R(\mathbf{v}) &= R_{\text{rPSC}}(\mathbf{v}) + w_{\text{RDE}} R_{\text{RDE}}(\mathbf{v}) + i R_{\text{elec}}(\mathbf{v}) \\
L(\mathbf{v}) &= L_{\text{rPSC\&RDE}}(\mathbf{v}) - i w_{\text{elec}} L_{\text{elec}}(\mathbf{v}) \\
S_{\text{total}}(\mathbf{t}) &= S_{\text{rPSC}}(\mathbf{t}) + w_{\text{RDE}} S_{\text{RDE}}(\mathbf{t}) + w_{\text{elec}} S_{\text{elec}}(\mathbf{t}) \\
&= \Re\left[\text{DFT}^{-1}[\text{DFT}[R(\mathbf{v})]^* \times \text{DFT}[L(\mathbf{v})]]\right]
\end{aligned}$$

where S_{rPSC} is a shape complementarity term, S_{RDE} is a desolvation free energy term, and S_{elec} is an electrostatic term. S_{total} consists of one correlation function.

3 Computational PPI Prediction

PPI prediction is used to predict the binding partner protein of one protein (and thus to predict two physically interacting pairs of proteins). In general, computational methods for PPI prediction fall into two categories, one with and one without protein docking. Methods using docking conduct docking calculations of input of two protein tertiary structures and evaluate how probable is the two input proteins binding. Methods without docking are represented by those with supervised machine learning. In many cases they learn and build a prediction model by input of a variety of protein features and output the possibility of the input proteins' ability for physical interaction. A docking-based model is usually more computationally intensive and limited to application to proteins whose tertiary structures are solved. However, they can provide not only the prediction of PPI possibility but also the probable binding poses that can help further discussions on the biological meaning of the interaction. This context is similar to that in the cheminformatics and medicinal chemistry fields [107–109]. In structure-based virtual screening, the compounds interacting with a target protein are predicted using a protein–ligand docking method; in ligand-based virtual screening they are predicted using supervised machine learning.

3.1 PPI Prediction Using Protein Docking

The PPI prediction problem can be defined as that of finding the binding partner protein that obtained optimal binding free energy ΔG. Thus, the naïve prediction method uses docking results to identify a protein pair with the better docking score than a threshold as interacting and one with a worse score as not interacting. However, docking score functions and conformational search spaces are coarse-grained, and docking scores are biased by their size and shape. Therefore, accurate prediction simply using raw docking scores is difficult.

So far, some PPI prediction methods use rigid-body docking techniques and are proposed and applied to real biological problems. The first such approaches were reported by Tsukamoto and Yoshikawa et al. [110–113] and Matsuzaki et al. [18] in 2008–2009. All of them used ZDOCK's docking scores [37]. Yoshikawa et al. and Matsuzaki et al. mainly used clustering of high-scoring conformations. The standardized docking score of representative structure by the clustered structures has been shown to improve the PPI prediction accuracy. Sacquin-Mora et al. also tackled this problem around the same time [114]. They considered a set of six complexes, and found that the correct interaction partners could be identified from 12 proteins if the residues forming the interface are known.

Yoshikawa et al. proposed a new scoring system [115] based on statistical analysis of interaction affinity score distributions sampled by their protein functions in 2010. Wass et al. used the rigid-body docking software Hex to predict interaction partners in 2011 [19, 116]. The complexes can be identified from the decoy dockings for approximately 50% of the complexes. In 2012, Ohue et al. presented a more accurate method with re-ranking techniques [40, 41] and proposed high-performance computing implementation called MEGADOCK to conduct a large-scale PPI prediction [28, 29]. With this software, predictions of PPIs in the human apoptosis pathway [117] and the bacterial chemotaxis signaling pathway [18, 118] were conducted. For prediction of bacterial chemotaxis-related PPIs, 101 protein structures corresponding to 13 proteins with structural variations were examined by exhaustive docking. It is a small well-known pathway and comprised of relatively stable protein binding pairs such as the receptor complex and transient protein, and binding occurs during phosphate transfer reactions. In 2013 Lopes et al. demonstrated a large-scale analysis of PPIs based on protein docking and showed that binding site predictions resulting from evolutionary sequence analysis are possible and realizable on the 168 proteins of the ZLAB Benchmark 2.0 [119]. They evaluated the quality of the interaction signal and the contribution of docking information compared to evolutionary information, showing that the combination of the two improves partner identification. Zhang et al. considered both the binding affinity and features of the binding energy landscape to distinguish binding pairs from non-binding pairs. The lowest docking score, the average Z-score, and convergence of the low-score solutions by SDOCK [46] were incorporated in their analysis [20]. Their method was used to screen for proteins that bind to tumor necrosis factor-α (TNFα). Out of 67 candidates, 16 proteins were validated by biochemical experiments (surface plasmon resonance binding assay), and 2 of these proteins showed significant binding affinity to TNFα. Ongoing studies include enrichment of the epidermal growth factor receptor (EGFR) pathway by predicting novel PPIs with non-small cell lung cancer related proteins, for which the predicted binding pairs are filtered by examining the expression data correlation for cancer cells (in preparation).

Although there are some interesting results obtained by computational predictions, the validation and discussion of possible roles of novel PPIs is crucial to broaden our knowledge of PPIs. For this purpose, we need to pursue collaborations between bioinformaticians and biologists. To expose predicted PPIs to the broader scientific community, web servers providing predicted PPIs and probable binding pose data might be a useful tool.

3.2 *PPI Prediction Without Docking*

To predict PPIs, not only de novo docking-based methods but also template-based methods based on similarity searches against known crystal complex structures can be used. Examples of well-known template-based PPI prediction methods are 3D-partner [31] and HOMCOS [32], which are based on whole sequence homology and structural similarity (called dimeric threading), PRISM [34, 35], which is based on interface template structures, and PrePPI [17, 120], which is based on structural matching and other experimental results such as co-expression, functional similarity, and evolutionary information. These methods can become more powerful as known interaction information and complex structures accumulate.

PRISM has been applied to a large number of pathways: human apoptosis pathway [121], MAPK signaling pathway [122], Toll-like receptor pathway [123–125], interleukin-1 initiated signaling pathway [126], and the interleukin-10 interaction network [127].

PRISM and PrePPI predict binding partners and predict the complex structure. By contrast, when a user just needs to predict whether a pair of proteins interact or not, it is also possible to utilize homology search tools such as BLAST and machine-learning techniques. If interacting proteins a and b in organism s_1 are found, their orthologs a' and b' in another organism s_2 also interact. Such conserved interaction proteins are called interologs [128–131]. Support vector machine (SVM) [132] is used in many machine-leaning techniques-based studies [133–140]. Prediction web servers are also available, for example, SPPS [141], which is based on the method by Shen et al. [135], and PRED_PPI [142], which is based on the method by Guo et al. [137]. Trigrams of amino acid sequences thus have vectors of $20^3 = 8{,}000$ dimensions and of the compressed amino acid alphabet, wherein similar amino acids are grouped into seven categories, as reported by Shen et al. [135]. Thus, in this study, the vectors have $7^3 = 343$ dimensions and are used as feature vectors for supervised learning. Pairwise kernel [136, 140] and S-kernel [135] are used as a kernel function of SVM in addition to Gaussian kernel (also known as the radial basis function kernel) [139].

Such machine learning-based predictions contain some difficulties on prediction performance assessment of PPI predictions. One is the quality of training data used for the supervised learning. Because experimentally determined PPIs data can have false positives and false negatives, especially those obtained by high-throughput methods, one needs to choose reliable experimental data carefully if one needs to build a highly precise prediction model [143]. Another difficulty lies in obtaining negative samples (experimental information that shows a particular pair is ‘not’ interacting). High-quality negative PPI data are equally important for learning and validation processes. The Negatome database [144, 145] has some data of such negative samples, although this is still not enough. At present, pseudo negative pairs (those not found in the positive PPI database that could be selected as candidates for pseudo negative pairs) are mainly used in machine learning. It is known that the choice of non-interactors in the training set affects the evaluation of accuracy of PPI predictions, and many of the reported PPI prediction accuracies tend to be overestimated [146]. This fact makes PPI prediction performance evaluation difficult, and there has been considerable debate on this matter [147, 148].

4 Profile Methods in Post-Docking Processes

In the previous sections we reviewed some prediction methods and software for protein–protein interactions (PPI), with the main focus being on protein docking. In this section we discuss an example of post-docking analysis using interaction profiles and future perspectives to apply it to PPI predictions. Protein docking methods are useful and generally used to search near-native complex structures of an input of two proteins known to interact. Rigid-body docking software generates thousands of protein complex poses, or decoys, including not only true but also many false positives, which are eliminated after post-docking analysis (see ‘post-docking analysis’ in Table 1). There are some post-docking works for predicting protein interaction surfaces by algorithms combined with multiple scoring functions [102, 149].

From other viewpoints, a set of decoys can be used as a set of interface samples to approach analysis of PPI mechanisms. In the post-docking process, cluster analysis is typically used for classifying decoys with various parameters, such as 3D structures of root mean square deviation (RMSD), interaction properties, and interaction fingerprints. RMSD is generally used for measuring similarities with the native structure, especially for evaluating accuracy of predicted complexes with the native crystal structures. On the other hand, interaction properties are used for estimating interaction scores with electrostatic, hydrophobic, and desolvation interactions. Fingerprint methods represent the residues included in the interface area of each decoy's conformations in a compact data structure. They are useful for investigating protein interaction interfaces, which are often combined with interaction properties [150, 151].

The rigid-body docking process does not consider molecular conformation changes. However, after obtaining multiple structure data for a target protein, often called ensemble conformations, multiple rigid-body docking processes may be able to overcome this issue. In this case, it is necessary to perform post-docking analysis for sets of decoys. When analyzing a set of decoys generated by multiple dockings using different structure data, it is difficult to compare the decoy structures by RMSD, whose values depend on alignment locations using 3D structures. The interaction profile method is introduced in the post-docking process when the protein interaction surfaces are the main focus of an investigation. Profiles are composed of interaction amino acid residues. Similarities between profiles are evaluated much more easily than when using 3D structures, such as RMSD, because there is no need to align their 3D structures. When the interaction profile is defined as an interaction amino acid pattern, similarity is calculated by, for example, the Tanimoto index. In cluster analysis, post-docking analysis with the profile method worked better than that with RMSD because classified groups with smaller energy score deviations are obtained by the profile method [152].

The rigid-body docking process explores docking space. However, in some cases, near-native structures cannot be found in a set of decoys generated by the rigid-body process because the exploration space is not enough. Such problems are found even in bound state cases. This serious problem of rigid-body docking could be solved by observing the interaction surfaces of decoys. Using the interaction profile with interacting amino acids in the Re-docking scheme [92], several candidates of the correct interface surfaces can be obtained after cluster analysis. Each surface is made by assembling interaction fingerprints of the decoys classified in the same decoy group. If the initial docking process could not explore enough docking space, near-native interactions were not found, even in areas including native interactions. Then docking processes are performed in each surface iteratively, indicating that multiple docking processes explored larger docking spaces. The re-docking scheme successfully obtained near-native structures in the cases of no near-natives after an initial docking process [92].

We have been describing interaction profile methods from the viewpoint of applying them to post-docking analysis to discriminate near-native conformations from false positive conformations. Interaction profiles can also be useful to investigate proteomes and PPI networks. On the docking-based PPI prediction (predicting interacting pairs of protein), conventional methods use post-docking analysis and rely heavily on docking scores. However, the docking score is still insufficient to represent binding energy itself [24], and it is difficult to pinpoint the most favorable binding pose only from docking scores. Thus, for predicting the PPIs, information regarding multiple decoys (usually high-scoring decoys) are examined. Interface profiles are useful for this analysis process because they make comparisons of decoys easier and emphasize the similarity of interface residue composition, which is important to discuss functionality of the interactions.

Another interesting application of using interface profiles is to analyze data obtained by dockings of one ‘receptor’ protein to multiple ‘interactor candidate’ proteins. In this case, interaction profiles are built for the receptor protein. Statistical comparison of the interaction fingerprints distribution of high-scoring decoys generated by dockings against different candidate binders would provide valuable data to investigate different types of protein contacts, for example specific and non-specific and transient and permanent.

Moreover, because the interaction profile is based on amino acid sequences, it is possible that this method can be applied to genome and proteome studies. There are findings that suggest related genes or domains are located relatively close in the genomes. For example, genes categorized into gene fusion, or Rosetta Stone are involved in protein interactions [153–156]. Analysis of decoy sets with interaction profiles can be connected to sequence-based analysis, for example, mapping of frequently interacting sites to the genome sequence using profiles.

In conclusion, interaction fingerprints allow fast and interface-focused analysis capable of analyzing large-scale data obtained by exhaustive dockings. It could be an approach to understand mechanisms of PPIs in the context of genome structures and evolutions.

5 Implementation of Docking Software on Supercomputing Environments

In this section we discuss the computational techniques that enable large numbers of exhaustive rigid dockings that feasibly cover the interactome scale. Analyses of part of an interactome may require large numbers of dockings. For example, to identify the drug-induced pathway of epidermal growth factor receptor (EGFR) signaling, about 200 proteins needed to be examined. In our preliminary survey on the EGFR pathway and related proteins data, we identified about 2,000 structures corresponding to these proteins. Therefore, the PPI network prediction system needs to handle about 2,000 × 2,000 combinations of protein structures.

To solve such large-scale problems, a highly efficient computing system is necessary. Developments of massively parallel supercomputing environments have continued to grow over the past decade. Some top ranked supercomputers have shown a peak performance of 27 petaflops (Titan, Oak Ridge National Laboratory, USA) and 11 petaflops (K computer, RIKEN, Advanced Institute of Computer Science (AICS), Japan) in November 2015 (http://top500.org).

An example of the docking software designed for usage on a supercomputer is MEGADOCK [28, 29]. It enables the performance of mega-scale numbers of protein–protein rigid dockings at once on massively parallel supercomputing systems. It is incorporated in a novel scoring function 'real Pairwise Shape Complementarity' (rPSC) in an FFT-based rigid docking scheme. rPSC represents the surface shape complementarities, the electrostatic interactions, and the desolvation free energy in a single complex number. Thus, the calculation using rPSC requires only one FFT calculation for a docking process. It makes each docking faster than conventional software and has multiple correlation functions that require multiple FFT calculations. Second, to conduct a large number of docking calculations effectively in parallel computing environments, MEGADOCK employs a hybrid parallelization (MPI/OpenMP) technique where a number of docking processes are distributed among the nodes by MPI, with each docking process also being calculated in parallel by threads by OpenMP within one node. The current version of MEGADOCK also implements parallelization on Graphics Processing Unit (GPU) [157] and Many Integrated Core (MIC) [158]. Users can choose a suitable implementation depending on their computing environments.

This data parallelizing system was scalable as shown by measurements carried out on two supercomputing environments. On the K computer, RIKEN AICS, Japan, a strong scaling efficiency of 91% on 82,944 nodes (when compared to the calculation time of the same problem on 41,472 nodes) was observed. It was also shown to have sufficient parallelization scalability on TSUBAME 2.5, Tokyo Institute of Technology, Japan, for both parallelization using MPI/OpenMP and on GPUs [29].

Studies aiming at incorporating protein structure and docking to the whole-genome scale are now actively conducted by researchers using supercomputers. Some pioneering studies have been performed by taking advantage of this large computational power. For example, Mosca et al. presented all the docking models generated by dockings of over 3,000 protein–protein pairs of *S. cerevisiae*. They used 217 proteins with experimentally determined crystal structures and 1,023 proteins structure models using homology modeling among roughly 6,000 putative *S. cerevisiae* proteins [27]. For larger scale dockings, along with MEGADOCK, a docking software Hex is well-known for its fast calculation and capable of parallelized calculation on GPUs [56].

6 Conclusion

The protein–protein interaction (PPI) network is a rich data source for systems biology analyses and is useful for various purposes such as inter-species PPI network comparison, drug target detection, and understanding the molecular mechanisms underlying specific cellular functions. However, current knowledge of PPI networks is quite incomplete, and thus prediction of novel PPIs by efficient computational methods is an important task. In this review we have described current computational methods to predict pairwise direct PPIs (binder predictions), with the main focus on the methods based on rigid-docking and post-processing. Although there is no perfect method of PPI prediction, both in the de novo docking-based and other heuristic methods, some have shown promising prediction performance. They have been applied to the discovery of novel PPIs related to important biological pathways such as mammalian signal transduction.

In this review we have explained the advantages and disadvantages of docking-based methods and current approaches to overcome the disadvantages. We have also emphasized that the docking decoys information previously unexplored thoroughly as false positive conformations can potentially be a valuable data source for analyzing PPIs from the viewpoint of interaction paths or interaction mechanisms.

Acknowledgements This work was partly supported by KAKENHI (grant number 15K00407, 24240044, 19300102, 16K00388), a Grant-in-Aid for Research and Development of The Next-Generation Integrated Life Simulation Software, all from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). An application MEGADOCK mentioned in this chapter was developed using the TSUBAME supercomputer system at the Global Scientific Information and Computing Center, Tokyo Institute of Technology, K computer at RIKEN, Japan, through the HPCI System Research Projects (Project ID: hp120131, hp140173).

References

1. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43 (Database issue):D204–D212. doi:10.1093/nar/gku989
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235
3. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43(Database issue):D470–D478. doi:10.1093/nar/gku1204
4. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32(Database issue):D449–D451. doi:10.1093/nar/gkh086
5. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA,

Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human protein reference database – 2009 update. Nucleic Acids Res 37(Database issue):D767–D772. doi:10.1093/nar/gkn892

6. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project – IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(Database issue):D358–D363. doi:10.1093/nar/gkt1115
7. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40(Database issue):D857–D861. doi:10.1093/nar/gkr930
8. Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nat Biotechnol 18:1257–1261. doi:10.1038/82360
9. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7:120. doi:10.1186/gb-2006-7-11-120
10. Hakes L, Pinney JW, Robertson DL, Lovell SC (2008) Protein-protein interaction networks and biology – what's the connection? Nat Biotechnol 26:69–72. doi:10.1038/nbt0108-69
11. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42. doi:10.1038/35075138
12. Franzosa EA, Xia Y (2011) Structural principles within the human-virus protein-protein interaction network. Proc Natl Acad Sci U S A 108:10538–10543. doi:10.1073/pnas.1101440108
13. Rachita HR, Nagarajaram HA (2014) Viral proteins that bridge unconnected proteins and components in the human PPI network. Mol Biosyst 10:2448–2458. doi:10.1039/c4mb00219a
14. Rao VS, Srinivas K, Sujini GN, Kumar GN (2014) Protein-protein interaction detection: methods and analysis. Int J Proteomics 2014:147648. doi:10.1155/2014/147648
15. Hue M, Riffle M, Vert JP, Noble WS (2010) Large-scale prediction of protein-protein interactions from structures. BMC Bioinformatics 11:144. doi:10.1186/1471-2105-11-144
16. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A (2005) PRISM: protein interactions by structural matching. Nucleic Acids Res 33:W331–W336. doi:10.1093/nar/gki585
17. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B (2013) PrePPI: a structure-informed database of protein-protein interactions. Nucleic Acids Res 41(Database issue):D828–D833. doi:10.1093/nar/gks1231
18. Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y (2009) *In silico* screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis. J Bioinform Comput Biol 7:991–1012
19. Wass MN, Fuentes G, Pons C, Pazos F, Valencia A (2011) Towards the prediction of protein interaction partners using physical docking. Mol Syst Biol 7:469. doi:10.1038/msb.2011.3
20. Zhang C, Tang B, Wang Q, Lai L (2014) Discovery of binding proteins for a protein target using protein-protein docking-based virtual screening. Proteins 82:2472–2482. doi:10.1002/prot.24611
21. Martin J, Lavery R (2012) Arbitrary protein-protein docking targets biologically relevant interfaces. BMC Biophys 5:7. doi:10.1186/2046-1682-5-7
22. Hwang H, Vreven T, Weng Z (2014) Binding interface prediction by combining protein-protein docking results. Proteins 82:57–66. doi:10.1002/prot.24354
23. Torchala M, Moal IH, Chaleil RA, Agius R, Bates PA (2013) A Markov-chain model description of binding funnels to enhance the ranking of docked solutions. Proteins 81:2143–2149. doi:10.1002/prot.24369

24. Kastritis PL, Bonvin AM (2010) Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 9:2216–2225. doi:10.1021/pr9009854
25. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, Bonvin AM, Weng Z (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 427:3031–3041. doi:10.1016/j.jmb.2015.07.016
26. Yugandhar K, Gromiha MM (2014) Protein-protein binding affinity prediction from amino acid sequence. Bioinformatics 30:3583–3589. doi:10.1093/bioinformatics/btu580
27. Mosca R, Pons C, Fernández-recio J, Aloy P (2009) Pushing structural information into the yeast interactome by high-throughput protein docking experiments. PLoS Comput Biol 5: e1000490. doi:10.1371/journal.pcbi.1000490
28. Matsuzaki Y, Uchikoga N, Ohue M, Shimoda T, Sato T, Ishida T, Akiyama Y (2013) MEGADOCK 3.0: a high-performance protein-protein interaction prediction software using hybrid parallel computing for petascale supercomputing environments. Source Code Biol Med 8:18. doi:10.1186/1751-0473-8-18
29. Ohue M, Shimoda T, Suzuki S, Matsuzaki Y, Ishida T, Akiyama Y (2014) MEGADOCK 4.0: an ultra-high-performance protein-protein docking software for heterogeneous supercomputers. Bioinformatics 30:3281–3283. doi:10.1093/bioinformatics/btu532
30. Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 19:161–162. doi:10.1093/bioinformatics/19.1.161
31. Cockell SJ, Oliva B, Jackson RM (2007) Structure-based evaluation of in silico predictions of protein protein interactions using comparative docking. Bioinformatics 23:573–581. doi:10.1093/bioinformatics/btl661
32. Fukuhara N, Kawabata T (2008) HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. Nucleic Acids Res 36:W185–W189. doi:10.1093/nar/gkn218
33. Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10:47–53. doi:10.1038/nmeth.2289
34. Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc 6:1341–1354. doi:10.1038/nprot.2011.367
35. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A (2014) PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res 42:W285–W289. doi:10.1093/nar/gku397
36. Chen R, Li L, Weng Z (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80–87. doi:10.1002/prot.10389
37. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z (2007) Integrating statistical pair potentials into protein complex prediction. Proteins 69:511–520. doi:10.1002/prot.21502
38. Pierce BG, Hourai Y, Weng Z (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One 6:e24657. doi:10.1371/journal.pone.0024657
39. Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 30:1771–1773. doi:10.1093/bioinformatics/btu097
40. Ohue M, Matsuzaki Y, Ishida T, Akiyama Y (2012) Improvement of the protein protein docking prediction by introducing a simple hydrophobic interaction model: an application to interaction pathway analysis. Lect Notes Comput Sci 7632:178–187. doi:10.1007/978-3-642-34123-6_16
41. Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y (2014) MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data. Protein Pept Lett 21:766–778. doi:10.2174/09298665113209990050
42. Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2003) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20:45–50. doi:10.1093/bioinformatics/btg371

43. Kozakov D, Brenke R, Comeau SR, Vajda S (2006) PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 65:392–406. doi:10.1002/prot.21117
44. Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, Vajda S (2013) How good is automated protein docking? Proteins 81:2159–2166. doi:10.1002/prot.24403
45. Gabb HA, Jackson RM, Sternberg MJE (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272:106–120. doi:10.1006/jmbi.1997.1203
46. Zhang C, Lai L (2011) SDOCK: a global protein-protein docking program using stepwise force-field potentials. J Comput Chem 32:2598–2612. doi:10.1002/jcc.21839
47. Tovchigrechko A, Vakser IA (2005) Development and testing of an automated approach to protein docking. Proteins 60:296–301. doi:10.1002/prot.20573
48. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:W310–W314. doi:10.1093/nar/gkl206
49. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 89:2195–2199
50. Ben-Zeev E, Eisenstein M (2003) Weighted geometric docking: incorporating external information in the rotation-translation scan. Proteins 52:24–27. doi:10.1002/prot.10391
51. Bajaj C, Chowdhury R, Siddavanahalli V (2011) F^{2}Dock: fast Fourier protein-protein docking. IEEE/ACM Trans Comput Biol Bioinform 8:45–58. doi:10.1109/TCBB.2009.57
52. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001) Protein docking using continuum electrostatics and geometric fit. Protein Eng Des Sel 14:105–113. doi:10.1093/protein/14.2.105
53. Roberts VA, Thompson EE, Pique ME, Perez MS, Ten Eyck LF (2013) DOT2: macromolecular docking with improved biophysical models. J Comput Chem 34:1743–1758. doi:10.1002/jcc.23304
54. Li L, Guo D, Huang Y, Liu S, Xiao Y (2011) ASPDock: protein-protein docking algorithm using atomic solvation parameters model. BMC Bioinformatics 12:36. doi:10.1186/1471-2105-12-36
55. Ritchie DW, Kemp GJL (2000) Protein docking using spherical polar Fourier correlations. Proteins Struct Funct Genet 39:178–194. doi:10.1002/(SICI)1097-0134(20000501)39:2<178::AID-PROT8>3.0.CO;2-6
56. Ritchie DW, Venkatraman V (2010) Ultra-fast FFT protein docking on graphics processors. Bioinformatics 26:2398–2405. doi:10.1093/bioinformatics/btq444
57. Macindoe G, Mavridis L, Venkatraman V, Devignes M-D, Ritchie DW (2010) HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 38: W445–W449. doi:10.1093/nar/gkq311
58. Garzon JI, Lopez-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, Chacon P (2009) FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics 25:2544–2551. doi:10.1093/bioinformatics/btp447
59. Venkatraman V, Yang YD, Sael L, Kihara D (2009) Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics 10:407. doi:10.1186/1471-2105-10-407
60. Esquivel-Rodríguez J, Yang YD, Kihara D (2012) Multi-LZerD: multiple protein docking for asymmetric complexes. Proteins Struct Funct Bioinf 80(7):1818–1833. doi:10.1002/prot.24079
61. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33:W363–W367. doi:10.1093/nar/gki481
62. Shentu Z, Al Hasan M, Bystroff C, Zaki MJ (2007) Context shapes: efficient complementary shape matching for protein-protein docking. Proteins 70:1056–1073. doi:10.1002/prot.21600
63. Gu S, Koehl P, Hass J, Amenta N (2012) Surface-histogram: a new shape descriptor for protein-protein docking. Proteins 80:221–238. doi:10.1002/prot.23192

64. Axenopoulos A, Daras P, Papadopoulos GE, Houstis EN (2013) SP-Dock: protein-protein docking using shape and physicochemical complementarity. IEEE/ACM Trans Comput Biol Bioinf 10:135–150. doi:10.1109/TCBB.2012.149
65. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein – protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737. doi:10.1021/ja026939x
66. de Vries SJ, van Dijk M, Bonvin AMJJ (2010) The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 5:883–897. doi:10.1038/nprot.2010.32
67. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 331:281–299. doi:10.1016/S0022-2836(03)00670-3
68. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36:W233–W238. doi:10.1093/nar/gkn216
69. Chaudhury S, Gray JJ (2008) Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. J Mol Biol 381:1068–1087. doi:10.1016/j.jmb.2008.05.042
70. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, Kuroda D, Xu J, Weitzner BD, Renfrew PD, Sripakdeevong P, Borgo B, Havranek JJ, Kuhlman B, Kortemme T, Bonneau R, Gray JJ, Das R (2013) Serverification of molecular modeling applications: the Rosetta online server that includes everyone (ROSIE). PLoS One 8:e63906. doi:10.1371/journal.pone.0063906
71. Moal IH, Bates PA (2010) SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci 11:3623–3648. doi:10.3390/ijms11103623
72. Torchala M, Moal IH, Chaleil RAG, Fernandez-Recio J, Bates PA (2013) SwarmDock: a server for flexible protein-protein docking. Bioinformatics 29:807–809. doi:10.1093/bioinformatics/btt038
73. Mashiach E, Nussinov R, Wolfson HJ (2010) FiberDock: flexible induced-fit backbone refinement in molecular docking. Proteins 78:1503–1519. doi:10.1002/prot.22668
74. Venkatraman V, Ritchie DW (2012) Flexible protein docking refinement using pose-dependent normal mode analysis. Proteins 80:2262–2274. doi:10.1002/prot.24115
75. Li L, Chen R, Weng Z (2003) RDOCK: refinement of rigid-body protein docking predictions. Proteins 53:693–707. doi:10.1002/prot.10460
76. Andrusier N, Nussinov R, Wolfson HJ (2007) FireDock: fast interaction refinement in molecular docking. Proteins 69:139–159. doi:10.1002/prot.21495
77. Pierce B, Weng Z (2007) ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 67:1078–1086. doi:10.1002/prot
78. Pierce B, Weng Z (2008) A combination of rescoring and refinement significantly improves protein docking performance. Proteins 72:270–279. doi:10.1002/prot.21920
79. Pons C, Talavera D, de la Cruz X, Orozco M, Fernandez-Recio J (2011) Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. J Chem Inf Model 51:370–377. doi:10.1021/ci100353e
80. Khashan R, Zheng W, Tropsha A (2012) Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. Proteins Struct Funct Bioinf 80:2207–2217. doi:10.1002/prot.24110
81. Cheng TM-K, Blundell TL, Fernandez-Recio J (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins Struct Funct Bioinf 68:503–515. doi:10.1002/prot.21419
82. Jiménez-García B, Pons C, Fernández-Recio J (2013) pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics 29:1698–1699. doi:10.1093/bioinformatics/btt262
83. Chuang G-Y, Kozakov D, Brenke R, Comeau SR, Vajda S (2008) DARS (decoys as the reference state) potentials for protein-protein docking. Biophys J 95:4217–4227. doi:10.1529/biophysj.108.135814

84. Ravikant DVS, Elber R (2010) PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins 78:400–419. doi:10.1002/prot.22550
85. Viswanath S, Ravikant DVS, Elber R (2013) Improving ranking of models for protein complexes with side chain modeling and atomic potentials. Proteins 81:592–606. doi:10.1002/prot.24214
86. Chowdhury R, Rasheed M, Keidel D, Moussalem M, Olson A, Sanner M, Bajaj C (2013) Protein-protein docking with F^{2}Dock 2.0 and GB-rerank. PLoS One 8:e51307. doi:10.1371/journal.pone.0051307
87. Sarti E, Granata D, Seno F, Trovato A, Laio A (2015) Native fold and docking pose discrimination by the same residue-based scoring function. Proteins 83:621–630. doi:10.1002/prot.24764
88. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797. doi:10.1016/j.jmb.2007.05.022
89. Omori S, Kitao A (2013) CyClus: a fast, comprehensive cylindrical interface approximation clustering/reranking method for rigid-body protein-protein docking decoys. Proteins 81:1005–1016. doi:10.1002/prot.2425
90. Oliva R, Vangone A, Cavallo L (2013) Ranking multiple docking solutions based on the conservation of inter-residue contacts. Proteins 81:1571–1584. doi:10.1002/prot.24314
91. Chermak E, Petta A, Serra L, Vangone A, Scarano V, Cavallo L, Oliva R (2015) CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts. Bioinformatics 31:1481–1483. doi:10.1093/bioinformatics/btu837
92. Uchikoga N, Matsuzaki Y, Ohue M, Hirokawa T, Akiyama Y (2013) Re-docking scheme for generating near-native protein complexes by assembling residue interaction fingerprints. PLoS One 8:e69365. doi:10.1371/journal.pone.0069365
93. Xue LC, Jordan RA, Yasser E-M, Dobbs D, Honavar V (2014) DockRank: ranking docked conformations using partner-specific sequence homology-based protein interface prediction. Proteins 82:250–267. doi:10.1002/prot.24370
94. Moal IH, Jimenez-Garcia B, Fernandez-Recio J (2015) CCharPPI web server: computational characterization of protein-protein interactions from structure. Bioinformatics 31:123–125. doi:10.1093/bioinformatics/btu594
95. Eisenstein M, Katchalski-Katzir E (2004) On proteins, grids, correlations, and docking. C R Biol 327:409–420. doi:10.1016/j.crvi.2004.03.006
96. Ritchie DW (2008) Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 9:1–15. doi:10.2174/138920308783565741
97. Janin J (2010) Protein–protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 6:2351. doi:10.1039/c005060c
98. Vakser IA (2013) Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol 23:198–205. doi:10.1016/j.sbi.2012.12.003
99. Vajda S, Hall DR, Kozakov D (2013) Sampling and scoring: a marriage made in heaven. Proteins 81:1874–1884. doi:10.1002/prot.24343
100. Huang S-Y (2014) Search strategies and evaluation in protein–protein docking: principles, advances and challenges. Drug Discov Today 19:1081–1096. doi:10.1016/j.drudis.2014.02.005
101. Moal IH, Moretti R, Baker D, Fernández-Recio J (2013) Scoring functions for protein-protein interactions. Curr Opin Struct Biol 23:862–867. doi:10.1016/j.sbi.2013.06.017
102. Moal IH, Torchala M, Bates PA, Fernández-Recio J (2013) The scoring of poses in protein-protein docking: current capabilities and future directions. BMC Bioinformatics 14:286. doi:10.1186/1471-2105-14-286
103. Szilagyi A, Zhang Y (2014) Template-based structure modeling of protein–protein interactions. Curr Opin Struct Biol 24:10–23. doi:10.1016/j.sbi.2013.11.005
104. Vreven T, Hwang H, Pierce BG, Weng Z (2014) Evaluating template-based and template-free protein-protein complex structure prediction. Brief Bioinform 15:169–176. doi:10.1093/bib/bbt047

105. Kundrotas PJ, Zhu Z, Janin J, Vakser IA (2012) Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A 109:9438–9441. doi:10.1073/pnas.1200678109
106. Negroni J, Mosca R, Aloy P (2014) Assessing the applicability of template-based protein docking in the twilight zone. Structure 22:1356–1362. doi:10.1016/j.str.2014.07.009
107. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 52:867–881. doi:10.1021/ci200528d
108. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935–949. doi:10.1038/nrd1549
109. McInnes C (2007) Virtual screening strategies in drug discovery. Curr Opin Chem Biol 11:494–502. doi:10.1016/j.cbpa.2007.08.033
110. Yoshikawa T, Tsukamoto K, Hourai Y, Fukui K (2008) Parameter tuning and evaluation of an affinity prediction using protein-protein docking. In: Proc 10th WSEAS Int Conf Math Methods Comput Tech Electr Eng, 312–317
111. Tsukamoto K, Yoshikawa T, Hourai Y, Fukui K, Akiyama Y (2008) Development of an affinity evaluation and prediction system by using the shape complementarity characteristic between proteins. J Bioinform Comput Biol 6:1133–1156
112. Yoshikawa T, Tsukamoto K, Hourai Y, Fukui K (2009) Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins. J Chem Inf Model 49:693–703
113. Tsukamoto K, Yoshikawa T, Yokota K, Hourai Y, Fukui K (2009) The development of an affinity evaluation and prediction system by using protein-protein docking simulations and parameter tuning. Adv Appl Bioinform Chem 2:1–15
114. Sacquin-Mora S, Carbone A, Lavery R (2008) Identification of protein interaction partners and protein–protein interaction sites. J Mol Biol 382:1276–1289. doi:10.1016/j.jmb.2008.08.002
115. Yoshikawa T, Seno S, Takenaka Y, Matsuda H (2010) Improved prediction method for protein interactions using both structural and functional characteristics of proteins. IPSJ Trans Bioinf 3:10–23. doi:10.2197/ipsjtbio.3.10
116. Wass MN, David A, Sternberg MJE (2011) Challenges for the prediction of macromolecular interactions. Curr Opin Struct Biol 21:382–390. doi:10.1016/j.sbi.2011.03.013
117. Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y (2013) Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods. BMC Proc 7:S6. doi:10.1186/1753-6561-7-S7-S6
118. Matsuzaki Y, Ohue M, Uchikoga N, Akiyama Y (2014) Protein-protein interaction network prediction by using rigid-body docking tools: application to bacterial chemotaxis. Protein Pept Lett 21:790–798
119. Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A (2013) Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput Biol 9:e1003369. doi:10.1371/journal.pcbi.1003369
120. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B (2012) Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature 490:556–560. doi:10.1038/nature11503
121. Acuner Ozbabacan SE, Keskin O, Nussinov R, Gursoy A (2012) Enriching the human apoptosis pathway by predicting the structures of protein-protein complexes. J Struct Biol 179:338–346. doi:10.1016/j.jsb.2012.02.002
122. Kuzu G, Keskin O, Gursoy A, Nussinov R (2012) Constructing structural networks of signaling pathways on the proteome scale. Curr Opin Struct Biol 22:367–377. doi:10.1016/j.sbi.2012.04.004
123. Guven Maiorov E, Keskin O, Gursoy A, Nussinov R (2013) The structural network of inflammation and cancer: merits and challenges. Semin Cancer Biol 23:243–251. doi:10.1016/j.semcancer.2013.05.003

124. Guven-Maiorov E, Keskin O, Gursoy A, Nussinov R (2015) A structural view of negative regulation of the toll-like receptor-mediated inflammatory pathway. Biophys J 109:1214–1226. doi:10.1016/j.bpj.2015.06.048
125. Guven-Maiorov E, Keskin O, Gursoy A, VanWaes C, Chen Z, Tsai C-J, Nussinov R (2015) The architecture of the TIR domain signalosome in the toll-like receptor-4 signaling pathway. Sci Rep 5:13128. doi:10.1038/srep13128
126. Acuner-Ozbabacan E, Engin B, Guven-Maiorov E, Kuzu G, Muratcioglu S, Baspinar A, Chen Z, Van Waes C, Gursoy A, Keskin O, Nussinov R (2014) The structural network of Interleukin-10 and its implications in inflammation and cancer. BMC Genomics 15:S2. doi:10.1186/1471-2164-15-S4-S2
127. Acuner Ozbabacan SE, Gursoy A, Nussinov R, Keskin O (2014) The structural pathway of interleukin 1 (IL-1) initiated signaling reveals mechanisms of oncogenic mutations and SNPs in inflammation and cancer. PLoS Comput Biol 10:e1003470. doi:10.1371/journal.pcbi.1003470
128. Gallone G, Simpson TI, Armstrong JD, Jarman AP (2011) Bio::Homology::InterologWalk – a Perl module to build putative protein-protein interaction networks through interolog mapping. BMC Bioinformatics 12:289. doi:10.1186/1471-2105-12-289
129. Rezende AM, Folador EL, Resende DDM, Ruiz JC (2012) Computational prediction of protein-protein interactions in Leishmania predicted proteomes. PLoS One 7:e51304. doi:10.1371/journal.pone.0051304
130. Folador EL, Hassan SS, Lemke N, Barh D, Silva A, Ferreira RS, Azevedo V (2014) An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. Integr Biol (Camb) 6:1080–1087. doi:10.1039/c4ib00136b
131. Murakami Y, Mizuguchi K (2014) Homology-based prediction of interactions between proteins using averaged one-dependence estimators. BMC Bioinformatics 15:213. doi:10.1186/1471-2105-15-213
132. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. doi:10.1007/BF00994018
133. Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein-protein interactions. Bioinformatics 21(Suppl 1):i38–i46. doi:10.1093/bioinformatics/bti1016
134. Martin S, Roe D, Faulon J-L (2005) Predicting protein-protein interactions using signature products. Bioinformatics 21:218–226. doi:10.1093/bioinformatics/bth483
135. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H (2007) Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci U S A 104:4337–4341. doi:10.1073/pnas.0607879104
136. Vert J-P, Qiu J, Noble WS (2007) A new pairwise kernel for biological network inference with support vector machines. BMC Bioinformatics 8:S8. doi:10.1186/1471-2105-8-S10-S8
137. Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res 36:3025–3030. doi:10.1093/nar/gkn159
138. Park Y (2009) Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. BMC Bioinformatics 10:419. doi:10.1186/1471-2105-10-419
139. Zhao X-W, Ma Z-Q, Yin M-H (2012) Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition. Protein Pept Lett 19:492–500. doi:10.2174/092986612800191080
140. Zhang S-W, Hao L-Y, Zhang T-H (2014) Prediction of protein–protein interaction with pairwise kernel support vector machine. Int J Mol Sci 15:3220–3233. doi:10.3390/ijms15023220
141. Liu X, Liu B, Huang Z, Shi T, Chen Y, Zhang J (2012) SPPS: a sequence-based method for predicting probability of protein-protein interaction partners. PLoS One 7:e30938. doi:10.1371/journal.pone.0030938

142. Guo Y, Li M, Pu X, Li G, Guang X, Xiong W, Li J (2010) PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. BMC Res Notes 3:145. doi:10.1186/1756-0500-3-145
143. Shi M-G, Xia J-F, Li X-L, Huang D-S (2010) Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids 38:891–899. doi:10.1007/s00726-009-0295-y
144. Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A (2010) The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res 38:D540–D544. doi:10.1093/nar/gkp1026
145. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Res 42:D396–D400. doi:10.1093/nar/gkt1079
146. Yu J, Guo M, Needham CJ, Huang Y, Cai L, Westhead DR (2010) Simple sequence-based kernels do not predict protein-protein interactions. Bioinformatics 26:2610–2614. doi:10.1093/bioinformatics/btq483
147. Park Y, Marcotte EM (2011) Revisiting the negative example sampling problem for predicting protein-protein interactions. Bioinformatics 27:3024–3028. doi:10.1093/bioinformatics/btr514
148. Park Y, Marcotte EM (2012) Flaws in evaluation schemes for pair-input computational predictions. Nat Methods 9:1134–1136. doi:10.1038/nmeth.2259
149. de Vries SJ, Bonvin AMJJ (2011) CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS One 6:e17695. doi:10.1371/journal.pone.0017695
150. Marcou G, Rognan D (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model 47:195–207. doi:10.1021/ci600342e
151. Deng Z, Chuaqui C, Singh J (2004) Structural Interaction Fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. J Med Chem 47:337–344. doi:10.1021/jm030331x
152. Uchikoga N, Hirokawa T (2010) Analysis of protein-protein docking decoys using interaction fingerprints: application to the reconstruction of CaM-ligand complexes. BMC Bioinformatics 11:236. doi:10.1186/1471-2105-11-236
153. Enright AJ, Iliopulous I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90. doi:10.1038/47056
154. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285:751–753. doi:10.1126/science.285.5428.751
155. Marcotte CJ, Marcotte EM (2002) Predicting functional linkages from gene fusions with confidence. Appl Bioinformatics 1:93–100
156. Yanai I, Derti A, DeLisi C (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A 98:7940–7945. doi:10.1073/pnas.141236298
157. Shimoda T, Ishida T, Suzuki S, Ohue M, Akiyama Y (2013) MEGADOCK-GPU: acceleration of protein-protein docking calculation on GPUs. In: Proc. Int. Conf. Bioinformatics, Comput. Biol. Biomed. Informatics – BCB'13. ACM Press, New York, pp 883–889. doi:10.1145/2506583.2506693
158. Shimoda T, Suzuki S, Ohue M, Ishida T, Akiyama Y (2015) Protein-protein docking on hardware accelerators: comparison of GPU and MIC architectures. BMC Syst Biol 9:S6. doi:10.1186/1752-0509-9-S1-S6

Adv Biochem Eng Biotechnol (2017) 160: 57–74
DOI: 10.1007/10_2016_40

Published online: 8 December 2016

Protein–Protein Interface and Disease: Perspective from Biomolecular Networks

Guang Hu, Fei Xiao, Yuqian Li, Yuan Li, and Wanwipa Vongsangnak

Abstract Protein–protein interactions are involved in many important biological processes and molecular mechanisms of disease association. Structural studies of interfacial residues in protein complexes provide information on protein–protein interactions. Characterizing protein–protein interfaces, including binding sites and allosteric changes, thus pose an imminent challenge. With special focus on protein complexes, approaches based on network theory are proposed to meet this challenge. In this review we pay attention to protein–protein interfaces from the perspective of biomolecular networks and their roles in disease. We first describe the different roles of protein complexes in disease through several structural aspects of interfaces. We then discuss some recent advances in predicting hot spots and communication pathway analysis in terms of amino acid networks. Finally, we highlight possible future aspects of this area with respect to both methodology development and applications for disease treatment.

G. Hu (✉) and Y. Li
Center for Systems Biology, School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China
e-mail: huguang@suda.edu.cn

F. Xiao
School of Basic Medicine and Biological Sciences, Medical College of Soochow University, Suzhou 215123, China

Y. Li
School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

W. Vongsangnak (✉)
Department of Zoology, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

Computational Biomodelling Laboratory for Agricultural Science and Technology (CBLAST), Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
e-mail: wanwipa.v@ku.ac.th

Keywords Amino acid network, Biomolecular network, Communication pathway, Disease, Hot spots, Protein–protein interaction

Contents

1 Introduction 58
2 Protein Complexes and Impact on Disease Association 60
2.1 G Protein–Coupled Receptors 61
2.2 Toroidal Proteins 61
2.3 p53 Tumor Suppressor Protein 62
2.4 Heat Shock Protein 90 62
3 Network Approaches for Hot Spots Identification 63
3.1 Graph Theoretical Analysis 63
3.2 Machine Learning Algorithms Based on Network Parameters 64
3.3 Network Connectivity of Interface Residues 64
3.4 Amino Acid Networks Based on Contact Energy 65
4 Communication Pathways Through Protein–Protein Interfaces 66
4.1 GPCR Dimerization Pathways 66
4.2 Hsp90-Client Protein Pathways 67
4.3 Pathways Across β Clamp Dimer 68
5 Perspectives 69
References 70

1 Introduction

Proteins are the most important biological macromolecules within living organisms. However, proteins rarely act alone and perform a vast array of biological functions in collaboration with other molecules, not only proteins but also DNA and RNA. Indeed, protein–protein interactions form the molecular basis of signaling and metabolic pathways, which are affected in multiple human diseases such as Creutzfeld–Jacob disease, Alzheimer's disease, and cancer [1–3]. As increasing numbers of structures for protein–protein complexes have been determined, the interactions between different protein chains within them provide detailed structural information for large-scale protein–protein interactions [4]. In particular, the analysis of protein–protein interfaces has enormous potential for understanding the molecular mechanism of diseases. For example, disease-associated mutation is always located at protein–protein interfaces of complexes [5]. In some cases, diseases are also related to allosteric changes affected by pathways because of the alterations in protein–protein interfaces [6, 7]. Thus, the means for targeting hot spots in protein interfaces and allosteric pathways through protein interfaces are becoming essential tools in drug discovery. The detection and modulation of protein–protein interfaces can also help to predict the drug side-effects [8].

Protein–protein interfaces are defined based on three different measures, including arithmetic distances between residue pairs, accessibility surface area, and Voronoi polyhedra [9]. Accordingly, current methods characterizing interfaces are mainly based on sequence data and some geometrical and physicochemical parameters, such as interface size, shape complementarity, hydrophobicity, and secondary structure on complex formation [10]. Apart from these methods, dynamical models were also used to investigate other interesting properties of interfacial residues, such as their fluctuation dynamics [11] and druggability [12]. Dynamical analysis shows that interfacial residues are more conserved and they have higher packing density than other surface residues [13]. The interfaces are considered as mostly 'undruggable' because of their large, flat, and featureless properties. However, the hot spots are key residues in the interfaces that contribute to most of the binding free energy, often forming central regions of the interface and thus possibly binding with small molecules as the drug targets [14]. The hot spots at protein–protein interfaces also proved to be disease-associated non-synonymous SNPs [15], whose mutation could dysregulate interactions. Interestingly, the network-based method was proposed to analyze interface properties of cancer-related proteins [16], which opened a door to study the structures of protein–protein interfaces from the perspective of biomolecular network.

Indeed, the method of representing biomolecular structures as networks is ever increasingly employed to investigate protein structures and functions [17, 18]. These networks could be named 'protein structure networks (PSNs),' 'protein contact networks (PCNs),' 'amino acid networks (AANs),' or 'residue interaction networks (RINs).' Here, we prefer to use AANs, which can be constructed from the Cartesian coordinates or the ensemble of protein structures. Each node represents a residue or a C_α atom, and each edge can be unweighted, just based on cut-off, or weighted, such as the Van der Waals contact score. Unlike other computational methods, AANs can describe protein structures and functions from the global prospective in terms of different topological parameters [19]. The clustering coefficient C_i is the normalized number of edges between the first neighbors of the vertex i by dividing it through the maximal number of such edges, which describe the hierarchical structure of proteins. The characteristic path length, L, is defined as the average shortest paths through which the two concerned nodes are connected by the smallest number of intermediate nodes. The analysis of C_i and L has revealed that proteins display small-world behavior [20]. The betweenness and the closeness are two more important parameters for descriptive network centrality. The betweenness centrality B_k of a node k is the number of times that a node is included in the shortest path between each pair of nodes, normalized by the total number of pairs [21]. The betweenness centrality of a node reflects the amount of control that this node exerts over the interactions of other nodes in the network. The closeness centrality C_k of a node k is the reciprocal of the average shortest path length, which is a measure of how quickly information spreads from a given node to other reachable nodes in the network [22].

AANs have utilized their small-world properties and centrality measures in the study of protein folding, protein stability and the prediction of functionally important residue [23]. There are several publications that demonstrate state-of-the-art of ANNs, from applications in protein allostery [24, 25] to therapeutic drug discovery [26–28]. More recently, AAN studies have been extended to protein complexes [29], especially focusing on their interactions including protein–protein interfaces and protein–DNA/RNA interfaces [30, 31]. Figure 1 shows an example of AANs

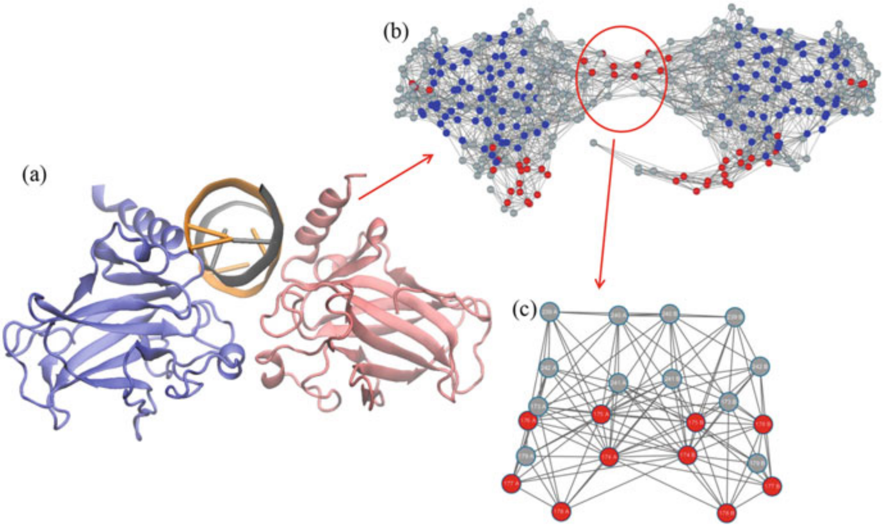


Fig. 1 Different structures of p53 dimer (PDB code: 3EXJ). (**a**) Cartoon representation of p53 dimer bound with DNA. (**b**) AAN representation of p53 dimer. (**c**) Sub-network of the p53 dimer interface. The AAN was generated by RINalyzer [32] based on the cut-off of 7 Å, in which *red nodes* represent helix structures, *blue nodes* represent sheet structures, and *gray nodes* represent loops. In the sub-network, node representations also include chain identifiers

for p53 protein complex. The network study of interfacial residues is of particular importance, not only because interfacial residues show special modular topological structures and sequence evolution but also because they participate in the interaction and have intrinsic communication ability [33].

In this chapter we focus on the protein–protein interfaces in different protein complexes from the perspective of biomolecular network and their roles in disease. We first describe the role of protein complexes in disease, taking advantage of the structural aspects of interfaces. Next, we review some network methods for predicting hot spots, and provide examples of using network theory for the communication pathway analysis through protein interfaces. Finally, we highlight possible future aspects of this area, such as mapping the structure of protein interface into disease-related protein–protein interaction networks.

2 Protein Complexes and Impact on Disease Association

Protein complexes are polymerized from chains or monomers, providing structural data to study their oligomerization mechanism and protein–protein interaction, which might associate with a wide spectrum of diseases and offer potentially therapeutic targets [34, 35]. Although there are many protein complexes associated with diseases, the application of biomolecular networks in the study of relations between protein–protein interfaces and disease is just beginning. In this section, some disease-related protein complexes which have been investigated by our group and others are listed, including G-protein-coupled receptors (GPCRs), toroidal proteins, the p53 tumor suppressor protein, and heat shock protein 90 (Hsp 90).

2.1 G Protein–Coupled Receptors

GPCRs form the largest superfamily of signal transduction membrane proteins [36]. A GPCR monomer consists of seven-transmembrane helices (H1–H7) which are connected by three intracellular and extracellular loops. It is known that many GPCRs exit as oligomers, and their oligomeric state plays a crucial role in many essential physiological processes as diverse as neurotransmission, cellular metabolism, cellular secretion, cell growth, immune defense, and cell differentiation. GPCRs provide about half of the total targets for existing drugs, and are involved in many illnesses such as retinal diseases and Alzheimer's disease [37].

AANs have been used to investigate structural communication of GPCRs, including GTPase, rhodopsin, β_2- and β_1-adrenergic receptors (ARs), and A_{2A} adenosine receptor ($A_{2A}R$). The network analysis of dynamical ensemble of GTPase has shown that the observed slight reduction of the RGS9-catalyzed GTPase activity of transducin depends on both perturbed communication between RGS9 and GTP binding site and inter-protein communication involving the nucleotide [38]. These results show that interactions both within and between proteins play key roles in the functions of GPCRs. Thus, mutations in protein–protein interactions of GPCRs can lead to many diseases.

2.2 Toroidal Proteins

Toroidal proteins are a family of proteins with donut-shaped or ring-shaped proteins, which are quite common forms for enzymes [39]. Toroidal proteins are known as oligomers assembled from two or more protein chains, forming a central hole that embraces DNA molecules inside or binds RNA molecules at the outside of the ring. These particular topological structures have the advantage of generating multiple identical binding sites for DNA or RNA. They have topologically different quaternary structures but share similar protein–protein interfaces.

On the other hand, toroidal proteins are also involved in various diseases. Proliferating Cell Nuclear Antigen (PCNA) is the most common toroidal trimer, which not only provides a scaffold for DNA replication but is also involved in DNA repair and cell cycle control. The direct inhibition of PCNA interacting with other proteins has been considered as an important step to treat diseases, including prostate cancer [40] and breast cancer [41]. p97 is another toroidal protein with a homohexameric structure [42], and chaperonins are toroidal proteins consisting of 7, 8, or 9 subunits, which assist in other proteins for their correct folding [43]. Their tertiary structures are formed by the help of ATP hydrolysis, and leading to the disease upon disturbing this process. For examples, p97 has been found implicated in Huntington's disease, Machado–Joseph disease, leukemia, and various cancers, chaperonins 60 are involved in arthritis, atherosclerosis, and prion diseases, and group 2 chaperonins are involved in neurodegenerative disorders, cardiovascular diseases, and cancer.

2.3 p53 Tumor Suppressor Protein

In the cell system, p53 regulates cell cycle progression and apoptosis, acting as a tumor suppressor by preventing DNA damage and oncogene activation [44]. The p53 tumor suppressor protein is a transcription factor which exists as a dimer or a tetramer. However, when binding with DNA, the p53 tetramer is more stable. Some diseases such as Alzheimer's disease and cancer can not only be caused by the fail replication of bound DNA, but also be associated with a conformational change or misfolding of the p53 tumor suppressor. The p53 tumor suppressor can therefore be considered as a key protein in a disease-related protein–protein interaction network. In this network, p53 interacts with partner proteins through different types of interfaces. Different diseases have different interactions caused by mutations of p53 at interfaces [45]. The identification of mutated hot spots at p53 interfaces and their mutated effects on interactions are important for p53 gene therapy.

2.4 Heat Shock Protein 90

Heat Shock Protein 90 (Hsp90) is also a chaperone protein, but does not belong to toroidal topology. Hsp90 chaperones demonstrate a common structure as homodimers, including *N*-terminal domain, middle domain, and *C*-terminal domain, named Hsp90-NTD, Hsp90-MD, and Hsp90-CTD, respectively [46]. These three domains carry out their own functions, binding ATP and client proteins and participating in dimerization, respectively. Hsp90 exits in different conformations that aid binding with ATP and other substrates, and has emerged as an important therapeutic target by focusing on the signal communication pathway caused by conformational change. The interactions between Hsp90 and client proteins are associated with all six hallmarks of cancer [47]. Thus, network approaches are used to perform the systems investigation of Hsp90–client proteins interactions, which could provide advance understanding of the molecular chaperone mechanisms underlying cancer at network level.

3 Network Approaches for Hot Spots Identification

Hot spots are key residues in protein–protein binding regions, providing a mechanical insight into interfaces. It has been found that hot spots always relate to disease-related mutations and drug binding sites [48]. AANs and their network topology properties can be used to predict hot spots, including graph theoretical analysis, machine learning algorithms based on network parameters, and network connectivity of interface residues. Knowledge of hot spots is extremely important for the understanding of molecular mechanisms of diseases. In the following we describe different network approaches for hot spots identification.

3.1 Graph Theoretical Analysis

To the best of our knowledge, the first type of AAN was proposed by Vishveshwara and co-workers [49], in which edges are defined based on strength of interaction I_{ij} between residues i and j. In this way, a protein can be represented as a 'Laplacian matrix' to encode connectivity information among residues. From this matrix, we can obtain their eigenvectors and eigenvalues as graph spectral information on AANs. The side chain interactions are captured at different threshold values of interaction strength cut-off $\mathrm{I}_{\mathrm{min}}$. The first application of the graph spectral approach is to detect a variety of side-chain clusters.

Such analysis has also been carried out on protein interfaces, where the eigenvalues and eigenvectors of the Laplacian matrix can identify hot spots. Applying this approach to the α–α dimer interface of a kind of RNA polymerase can indicate hot regions, including nine residues (i.e., Phe8, Leu31, Glu32, Phe35, Thr38T, Leu39, Ile46, Ser50, and Gln227), in which Phe35 and Ile46 are two hot spots [50]. The graph spectral theory was further applied to a larger dataset of homodimers to predict interface hot spots. In comparison with dimeric proteins [51], the legume lectin family provides the higher oligomeric protein models for analyzing the role of interface residues. AANs for galectins, pentraxins, calnexin, calreticulin, and rhesus rotavirus Vp4 sialic-acid-binding domains were constructed, and the following analysis was carried out in terms of amino acid clusters with special emphasis on protein–protein interfaces [52–54]. These studies showed hot spots associated with highly connected residues at the interface, also called interfacial hubs.

3.2 Machine Learning Algorithms Based on Network Parameters

Network parameters of AANs can be used to predict key residues in proteins. Another kind of un-weighted AANs just based on cut-offs was used to investigate a set of 48 dimers [21] and a set of 18 protein–protein complexes [54, 55]. The small-world network was further applied to find hot spots of protein complexes by ranking clustering coefficients and betweenness. It was found that highly central residues in the network were most probably hot spots, and betweenness showed particularly high accuracy in predicting key interface residues in dimerization [21]. By further rewiring of the small-world networks, they have generated cluster structures of central residues at protein–protein interfaces [55]. AANs can also be weighted by energy function derived from knowledge-based potentials. If we construct some minimum cut trees from such AANs, one can predict hot spots just using the most simple network parameters, degree, directly. This means that the most connected node in the minimum cut tree for protein complexes corresponded to key residues in protein interfaces [56].

Machine learning algorithms can be used to combine and train these network features to improve further the prediction performance. Support vector machines (SVMs) and neural networks are strongly associated with machine learning in bioinformatics [57]. Although these methods are less used to predict hot spots, we still found a relevant work. In 2014, Li et al. [58] proposed an SVM model that includes network parameters such as degree, closeness, and betweenness in both bound and unbound proteins for the prediction of hot spots. A satisfactory accuracy (ACC) value of 79.0% and a Matthew's correlation coefficient (MCC) value of 0.470 were obtained for independent hot spots sets.

3.3 Network Connectivity of Interface Residues

Gemini constitutes a series of programs and databases based on network theory to investigate network connectivity of interface residues in oligomeric proteins [59]. It contains four components: GeminiDistances, GeminiRegions, GeminiGraph, and GeminiData. First, GeminiDistances selects residues of the protein–protein interface based on minimal interactions. Then, GeminiRegions divides the protein–protein interfaces into different regions consisting of elementary interaction networks between residues of two adjacent monomers. Finally, GeminiGraph constructs a network for these interface regions, and thus a bi-colored graph for the protein–protein interface is obtained. The results obtained from GeminiDistances, GeminiRegions, and GeminiGraph are stored in GeminiData. Accordingly, Gemini offers quite a useful method to characterize protein interfaces quantitatively by using network theory.

Gemini has been applied to the analysis of hot spots of β-strand interfaces from a set of protein oligomers [60] and the tumor suppressor p53 tetramer [61]. The interface networks for β-strand interfaces contain two types of sub-networks. The backbone (*BB*) and side chain (*SC*) networks involve interactions between main chain atoms and between side chain atoms, respectively. Hot spots in the *BB* network are mostly hydrophobic residues, whereas in the *SC* network there are charged residues such as Arg and Glu. The charge distribution of hot spots helps the assembly through the intermolecular β-strands, which might be more sensitive to the disease mutations. Gemini networks were built for both p53 WT and mutant structures to explore the changes of networks upon single mutation. The G334V mutant is accompanied by the rewiring of the WT network, which not only leads to the dissociation of the p53 tetramer but also has a strong global effect on the network. Thus, Glu 334 was predicted as a hot spot and its mutation associated with cancer.

3.4 Amino Acid Networks Based on Contact Energy

Amino acid contact energy networks (AACENs) are the newest AANs [62], in which nodes are represented as residues and edges are established when environment-dependent residue–residue contact energies are less than zero. The AACENs was first proposed to study protein evolution. In comparison with other AANs, an AACEN use a different definition whose connections are based on contact energy. On the other hand, hot spots are defined as critical interfacial residues that contribute most to the binding energy. According to the similarity of the two definitions, we have suggested that the AACEN may provide a straightforward way to detect hot spots. By extracting sub-networks only including interface residues between different chains, it may provide a simple but straightforward method to identify hot spots of protein complexes.

More recently, we have extended AACENs to study the oligomerization of toroidal proteins [63]. Using the hot spots predicted by HotRegion as a reference data set, the performance values of sub-network nodes are S 66%, C 74%, P 62%, and A 71% for sensitivity, specificity, precision, and accuracy, respectively. In the case of the β clamp dimer (Fig. 2), the sub-network of the interface between chain A and chain B contains 13 nodes, and 7 out of 9 hot spots have been predicted successfully. Among them, Phe106 has the largest degree and mutation of this residue mostly leads to the destabilization of the interface.

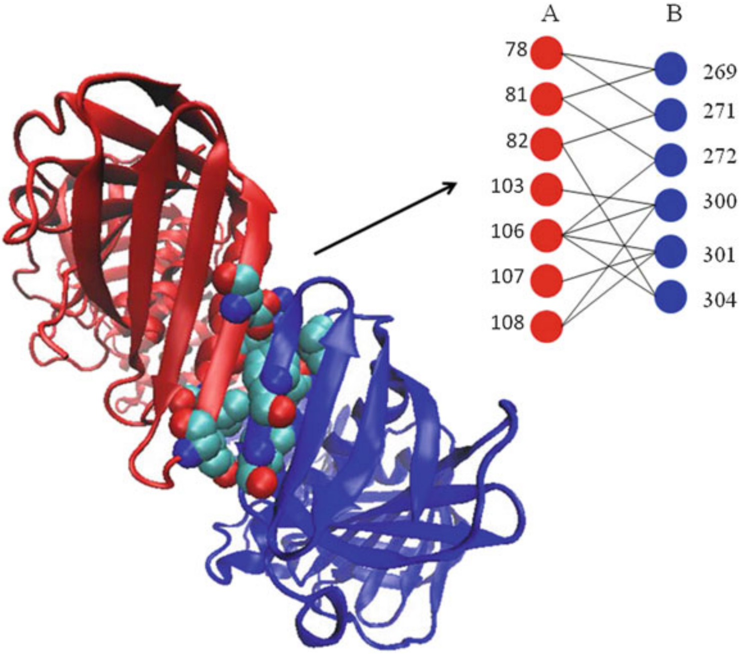


Fig. 2 Hot spots in the AB interface of the β clamp (PDB code: 2POL) predicted by AACENs, which are shown as Van der Waals representations in three-dimensional structure and as nodes in the sub-network (adopted from [63])

4 Communication Pathways Through Protein–Protein Interfaces

Allostery is a common phenomenon in protein complexes whereby a perturbation by an effector at one site of the monomer leads to a functional change at another through protein–protein interfaces. Therefore, the investigation of communication pathways between different protein subunits is important to understand the molecular basis of disease caused by allostery regulation [64]. As described below, we revisit communication pathways of three respective proteins in Sect. 2, including GPCR dimers, Hsp90 complexes, and a toroidal protein. It should be noted that p53 participates in protein–protein interaction networks for signal transduction, and thus includes higher level communications which is discussed from various perspectives.

4.1 GPCR Dimerization Pathways

GPCRs are allosteric proteins whose biological functions are regulated by allosteric communication between the extracellular and intracellular poles of the helix bundle, or across monomers in the dimerization and oligomerization states. The AAN representation of structural ensembles from molecular dynamics (MD) simulation has proved to be a meaningful way to enumerate these pathways in GPCRs [65]. The dimerization has an effect on the structural communication within the monomer. Network analysis found that the helix 1 (H1) plays an important role in $A_{2A}R$ dimerization [66], and the conservation of helices 1, 2, 6, and 7 between the two poles of the helix bundle. In addition, allosteric communications across three types of dimerization interfaces were investigated. The topology of pathways across H1–H1/H2–H2 and H1–H4/H2–H2 dimer interfaces remain, but the number of hub-involving links and paths increase significantly from the monomer. D2R forms higher-order oligomers, and H1, H4, and H8 from different inter-monomer interfaces [67]. Of course, the higher-order oligomerization has different impacts on such communication. The information flow between different D2R monomers is mediated by H1–H1, H1–H2, and H8–H8 contacts. Indeed, Table 5.1 in [68] has listed dimeric and oligomeric interfaces involved in various GPCRs. Thus, the communication analysis of these dimer interfaces can explain how oligomerization affects the functional dynamics of GPCRs.

As an AIDS- and HIV-1-related GPCR protein, human CXC chemokine receptor type 4 (CXCR4) provides a new structural model to investigate dimerization. In this special case, sequence information has been introduced to combine with the network to investigate the structural communication in CXCR4 [69]. First, statistical coupling analysis was used to quantify pairwise correlation of amino acid evolution. Second, MD was performed to obtain structural information on the dimer interface. Then the co-evolutionary relationship was to weight networks for CXCR4 dimer. This combined network analysis has found that three helixes, TMs 3, 4, and 5, are considered as co-evolution sectors and play key roles in the communication through the dimer interface.

4.2 Hsp90-Client Protein Pathways

Hsp90 molecular chaperone represents a typical paradigm that signals transduction pathways control functionally important biological processes and relate to serious human cancer. The Hsp90-CTD region, for example, participates in the Hsp90 dimerization, and involves the inter-domain communication pathways from the nucleotide binding site to the allosteric binding site. The interaction between Hsp90 and the diverse array of client proteins (p53 and oncogenic kinases) is the principle of these molecular mechanisms. AANs combined fluctuation dynamics have given novel insight of their functional dynamics and allosteric communications for Hsp90-client protein complexes [47]. AANs of Hsp90 interaction networks show small-world organization, and thus the centrality analysis enables the quantitative modeling of signal propagation mechanisms. Such analyses have been performed on Hsp90-p53, Hsp90-p23, Hsp90-Aha1, Hsp90-Cdc37, Hsp90-Sgt1, and Hsp90-Sgt1-Rar1 complexes [70–73]. The interfaces of these Hsp90 complexes were identified by network parameters, including clusters, hubs, cliques, and communities. These local topological features give clues to key residues for mediating allosteric communication pathways. In general, two conserved residues with high centrality are found to appear most on pathways. These studies showed that AANs enable not only global topological analysis but also communication analysis of Hsp90–client proteins interactions.

4.3 Pathways Across β Clamp Dimer

The β clamp is a PCNA-like toroidal dimer. We have previously used the weighted AAN based on the Van der Waals contact score [74] to explore structural communications encoded by global topology of the β clamp [75]. First, the communication pathway across the dimer interface in the β clamp can be identified from AANs using the Floyd–Warshall algorithm. Six shortest paths have been predicted through the interface between chain A and chain B: (1) A: Pro71 (α_1) → A: Lys74 → B: Ile272 → B: Arg269 (α_2), as show in Fig. 3; (2) A: Gly81 → B: Arg269 → B: Ile 272; (3) A: Gly81 → A: Ile78 → B: 273 → B: Glu300; (4) A: Leu82 → B: Arg269 → B: Ile272; (5) A: Leu82 → A: Phe106 → B: Leu273 → B: Glu300; (6) B: Gly81 → B: Asp77 → B: Lys74 → A: Glu300 → A: Gln299. These pathways

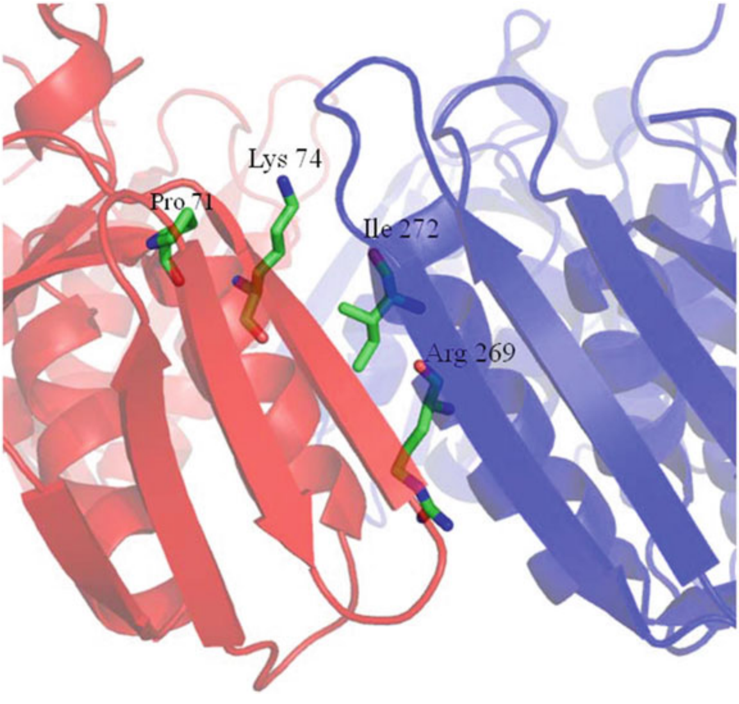


Fig. 3 Communication pathway at the AB interface of the β clamp obtained by the shortest path analysis of AAN based on Van der Waals contact score (adopted from [75])

through the AB interface could mediate long-range communication between two monomers. Second, z-scores based on betweenness and closeness were used to evaluate the importance of each residue along these pathways. In our case, for example, Lys74 in chain A and Ile272 in chain B are two interface hot spots for the first pathway.

5 Perspectives

Protein–protein interactions form the molecular basis of very many diseases at the level of systems biology, providing putative new targets for drug discovery. However, protein–protein interaction networks are always constructed based on the protein associations, in which one protein could connect with many other proteins. Such kinds of networks miss structural details of individual protein and their binding information, and thus they are too abstract to reflect biological reality. Fortunately, the growing number of structures for protein complexes gives enough structural information on how proteins interact on a genome-wide scale. If protein–protein interactions include binding information from protein three-dimensional structures, the related networks move from static and abstract representations to physical and dynamical interactions, which are more reasonable to identify individual proteins as drug targets in a protein–protein interaction network. Therefore, how integrating protein interfaces into protein–protein interaction networks is becoming a major challenge in future development [76]. Indeed, this object belong to the goals of structural systems biology [77], which was proposed about 10 years ago but its development has not been as rapid as expected. This concept is summarized in Fig. 4, in which p53 interacts with a multitude of protein partners and their hot spots information can been integrated into a protein–protein interaction network. Further systems biological analyses of the protein–protein interaction network, such as mapping gene expression patterns or GeneGo pathway analysis, can predict effective disease-associated biomarkers [79].

As a final perspective, the following two breakthroughs in methods might have potential applications to refresh this field:

1. PRISM is a structural matching-based method for predicting protein–protein interactions [80]. The main idea of this approach is that two proteins may interact if their unbound surface pairs can find a known protein interface with similar structure. To complement this method, a database with 22,604 unique interface structures has been built to provide a rich resource for template-based docking [81]. This method has also recently been formalized as a web server [82].
2. Another method for predicting PPIs based on structures of protein complexes is called PrePPI [83]. It uses an algorithm based on Bayesian statistics to combine structural and non-structural interaction clues, which give fine details of the interaction between proteins and their interacting partners.

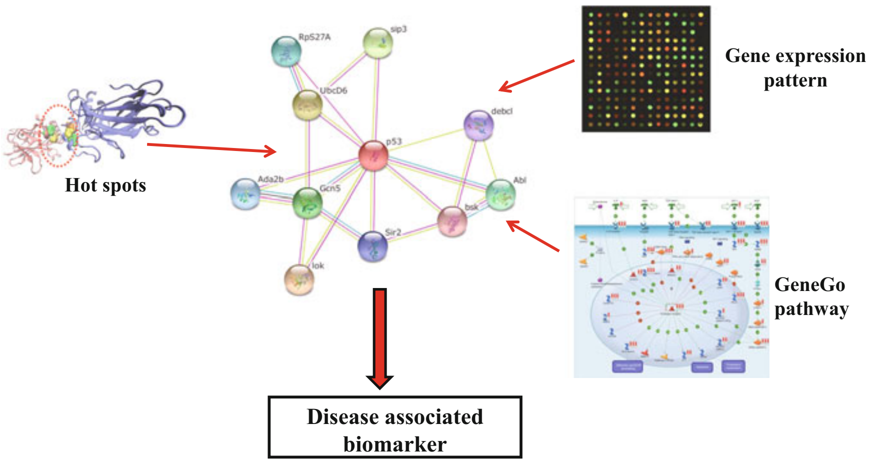


Fig. 4 Integration of hot spots of p53 dimer and gene expression patterns into protein–protein interaction network and the GeneGo pathway analysis for predicting effective disease-associated biomarkers. String server [78] was used for modeling protein–protein interactions in the p53 centered network

Of course, this research area is just starting. There are many open questions both in methodology development and applications in disease treatment. We only list two of them for the coming years. Protein–protein interactions at interfaces are dynamic, so how to develop and apply computational methods for investigating these dynamical features remains a challenge. The elastic network model is efficient for high-throughput investigations of protein dynamics and thus represents a promising approach [84]. So far, there are many databases for protein–protein interactions related to diseases [85]. Therefore, knowing how to annotate these interactions by interface information, such as that on binding sites and signal transduction, would improve our understanding of molecular mechanisms of diseases.

Acknowledgements This work was supported by the National Natural Science Foundation of China (21203131, 31200989, and 61401068), the Natural Science Foundation of the Jiangsu Higher Education Institutions (12KJB180014), and the China Postdoctoral Science Foundation (2016M590495), Preproposal Research Fund (PRF4/2558 and PRF-PII/59), Faculty of Science, Kasetsart University.

References

1. Gonzalez MW, Kann MG (2012) Chapter 4: protein interactions and disease. PLoS Comput Biol 8(12):11
2. Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 8(5):333–346

3. Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, Golshani A (2008) Computational methods for predicting protein-protein interactions. Adv Biochem Eng Biotechnol 110:247–267
4. Cavga AD, Karahan N, Keskin O, Gursoy A (2015) Taming oncogenic signaling at protein interfaces: challenges and opportunities. Curr Top Med Chem 15(20):2005–2018
5. Nero TL, Morton CJ, Holien JK, Wielens J, Parker MW (2014) Oncogenic protein interfaces: small molecules, big challenges. Nat Rev Cancer 14(4):248–262
6. Nussinov R, Tsai CJ (2013) Allostery in disease and in drug discovery. Cell 153(2):293–305
7. Persico M, Di Dato A, Orteca N, Fattorusso C, Novellino E, Andreoli M, Ferlini C (2015) From protein communication to drug discovery. Curr Top Med Chem 15(20):2019–2031
8. Engin HB, Keskin O, Nussinov R, Gursoy A (2012) A strategy based on protein-protein interface motifs may help in identifying drug off-targets. J Chem Inf Model 52(8):2273–2286
9. Lesieur C (2014) The assembly of protein oligomers — old stories and new perspectives with graph theory. In: DCL: INTECH (ed) Oligomerization of chemical and biological compounds. doi:10.5772/58576
10. Fernandez-Recio J (2011) Prediction of protein binding sites and hot spots. Wiley Interdiscip Rev Comput Mol Sci 1(5):680–698
11. Zen A, Micheletti C, Keskin O, Nussinov R (2010) Comparing interfacial dynamics in protein-protein complexes: an elastic network approach. BMC Struct Biol 10:13
12. Ulucan O, Eyrisch S, Helms V (2012) Druggability of dynamic protein-protein interfaces. Curr Pharm Des 18(30):4599–4606
13. Lin JJ, Lin ZL, Hwang JK, Huang TT (2015) On the packing density of the unbound protein-protein interaction interface and its implications in dynamics. BMC Bioinformatics 16(Suppl 1):S7
14. Cukuroglu E, Engin HB, Gursoy A, Keskin O (2014) Hot spots in protein-protein interfaces: towards drug discovery. Prog Biophys Mol Biol 116(2-3):165–173
15. David A, Razali R, Wass MN, Sternberg MJE (2012) Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat 33(2):359–363
16. Kar G, Gursoy A, Keskin O (2009) Human cancer protein-protein interaction network: a structural perspective. PLoS Comput Biol 5(12):e1000601
17. Brinda KV, Vishveshwara S (2005) A network representation of protein structures: implications for protein stability. Biophys J 89(6):4159–4170
18. Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, Csermely P (2007) Network analysis of protein dynamics. FEBS Lett 581(15):2776–2782
19. Doncheva NT, Assenov Y, Domingues FS, Albrecht M (2012) Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc 7 (4):670–685
20. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393 (6684):440–442
21. del Sol A, O'Meara P (2005) Small-world network approach to identify key residues in protein-protein interaction. Proteins Struct Funct Bioinf 58(3):672–682
22. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344 (4):1135–1146
23. Yan WY, Zhou JH, Sun MM, Chen JJ, Hu G, Shen BR (2014) The construction of an amino acid network for understanding protein structure and function. Amino Acids 46(6):1419–1439
24. Vuillon L, Lesieur C (2015) From local to global changes in proteins: a network view. Curr Opin Struct Biol 31:1–8
25. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
26. Viswanathan K, Shriver Z, Babcock GJ (2015) Amino acid interaction networks provide a new lens for therapeutic antibody discovery and anti-viral drug optimization. Curr Opin Virol 11:122–129

27. Csermely P, Korcsmaros T, Kiss HJ, London G, Nussinov R (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther 138(3):333–408
28. Liang Z, Hu G (2016) Protein structure network-based drug design. Mini Rev Med Chem 16(16):1330–1343
29. Zhang XW, Perica T, Teichmann SA (2013) Evolution of protein structures and interactions from the perspective of residue contact networks. Curr Opin Struct Biol 23(6):954–963
30. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA: protein complexes. Proc Natl Acad Sci U S A 106(16):6620–6625
31. Sathyapriya R, Vijayabaskar MS, Vishveshwara S (2008) Insights into protein-DNA interactions through structure network analysis. PLoS Comput Biol 4(9):e1000170
32. Doncheva NT, Klein K, Domingues FS, Albrecht M (2011) Analyzing and visualizing residue networks of protein structures. Trends Biochem Sci 36(4):179–182
33. Reichmann D, Rahat O, Albeck S, Meged R, Dym O, Schreiber G (2005) The modular architecture of protein-protein binding interfaces. Proc Natl Acad Sci U S A 102(1):57–62
34. Hayden EY, Teplow DB (2013) Amyloid beta-protein oligomers and Alzheimer's disease. Alzheimers Res Ther 5(6):60
35. Gabizon R, Friedler A (2014) Allosteric modulation of protein oligomerization: an emerging approach to drug design. Front Chem 2:9
36. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK et al (2007) High-resolution crystal structure of an engineered human beta(2)-adrenergic G protein-coupled receptor. Science 318(5854):1258–1265
37. Mariani S, Dell'Orco D, Felline A, Raimondi F, Fanelli F (2013) Network and atomistic simulations unveil the structural determinants of mutations linked to retinal diseases. PLoS Comput Biol 9(8):14
38. Raimondi F, Felline A, Portella G, Orozco M, Fanelli F (2013) Light on the structural communication in Ras GTPases. J Biomol Struct Dyn 31(2):142–157
39. Hu G, Michielssens S, Moors SLC, Ceulemans A (2011) Normal mode analysis of Trp RNA binding attenuation protein: structure and collective motions. J Chem Inf Model 51 (9):2361–2371
40. Zhao H, Lo YH, Ma L, Waltz SE, Gray JK, Hung MC, Wang SC (2011) Targeting tyrosine phosphorylation of PCNA inhibits prostate cancer growth. Mol Cancer Ther 10(1):29–36
41. Zhao H, Chen MS, Lo YH, Waltz SE, Wang J, Ho PC, Vasiliauskas J, Plattner R, Wang YL, Wang SC (2014) The Ron receptor tyrosine kinase activates c-Abl to promote cell proliferation through tyrosine phosphorylation of PCNA in breast cancer. Oncogene 33(11):1429–1437
42. Chapman E, Fry AN, Kang MJ (2011) The complexities of p97 function in health and disease. Mol Biosyst 7(3):700–710
43. Ranford JC, Henderson B (2002) Chaperonins in disease: mechanisms, models, and treatments. J Clin Pathol Mol Pathol 55(4):209–213
44. Malecka KA, Ho WC, Marmorstein R (2009) Crystal structure of a p53 core tetramer bound to DNA. Oncogene 28(3):325–333
45. Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R (2009) Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. Mol Biosyst 5(12):1770–1778
46. Ali MMU, Roe SM, Vaughan CK, Meyer P, Panaretou B, Piper PW, Prodromou C, Pearl LH (2006) Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. Nature 440(7087):1013–1017
47. Verkhivker GM (2014) Computational studies of allosteric regulation in the Hsp90 molecular chaperone: from functional dynamics and protein structure networks to allosteric communications and targeted anti-cancer modulators. Isr J Chem 54(8-9):1052–1064
48. Ma BY, Nussinov R (2014) Druggable orthosteric and allosteric hot spots to target protein-protein interactions. Curr Pharm Des 20(8):1293–1301

49. Kanna N, Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 292(2):441–464
50. Kannan N, Chander P, Ghosh P, Vishveshwara S, Chatterji D (2001) Stabilizing interactions in the dimer interface of alpha-subunit in Escherichia coli RNA polymerase: a graph spectral and point mutation study. Protein Sci 10(1):46–54
51. Brinda KV, Kannan N, Vishveshwara S (2002) Analysis of homodimeric protein interfaces by graph-spectral methods. Protein Eng 15(4):265–277
52. Brinda KV, Mitra N, Surolia A, Vishveshwara S (2004) Determinants of quaternary association in legume lectins. Protein Sci 13(7):1735–1749
53. Brinda KV, Surolia A, Vishveshwara S (2005) Insights into the quaternary association of proteins through structure graphs: a case study of lectins. Biochem J 391:1–15
54. Brinda KV, Vishveshwara S (2005) Oligomeric protein structure networks: insights into protein-protein interactions. BMC Bioinformatics 6
55. del Sol A, Fujihashi H, O'Meara P (2005) Topology of small-world networks of protein-protein complex structures. Bioinformatics 21(8):1311–1315
56. Tuncbag N, Salman FS, Keskin O, Gursoy A (2010) Analysis and network representation of hotspots in protein interfaces using minimum cut trees. Proteins Struct Funct Bioinf 78 (10):2283–2294
57. Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananza R, Santafe G, Perez A, Robles V (2006) Machine learning in bioinformatics. Brief Bioinform 7(1):86–112
58. Ye L, Kuang QF, Jiang L, Luo JS, Jiang YP, Ding ZL, Li YZ, Li ML (2014) Prediction of hot spots residues in protein-protein interface using network feature and microenvironment feature. Chemom Intell Lab Syst 131:16–21
59. Feverati G, Lesieur C (2010) Oligomeric interfaces under the lens: gemini. PLoS One 5(3):15
60. Feverati G, Achoch M, Zrimi J, Vuillon L, Lesieur C (2012) Beta-strand interfaces of non-dimeric protein oligomers are characterized by scattered charged residue patterns. PLoS One 7(4):e32558
61. Feverati G, Achoch M, Vuillon L, Lesieur C (2014) Intermolecular beta-strand networks avoid hub residues and favor low interconnectedness: a potential protection mechanism against chain dissociation upon mutation. PLoS One 9(4):e94745
62. Yan WY, Sun MM, Hu G, Zhou JH, Zhang WY, Chen JJ, Chen B, Shen BR (2014) Amino acid contact energy networks impact protein structure and evolution. J Theor Biol 355:95–104
63. Yan WY, Hu G, Shen BR (2016) Network analysis of protein structures: the comparison of three topologies. Curr Bioinformatics 11(4):480–489. doi:10.2174/1574893611666160602124707
64. Feher VA, Durrant JD, Van Wart AT, Amaro RE (2014) Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol 25:98–103
65. Fanelli F, Felline A, Raimondi F (2013) Network analysis to uncover the structural communication in GPCRs. In: Receptor-receptor interactions, vol 117. pp 43–61
66. Fanelli F, Felline A (2011) Dimerization and ligand binding affect the structure network of A (2A) adenosine receptor. Biochim Biophys Acta 1808(5):1256–1266
67. Fanelli F, Mauri M, Capra V, Raimondi F, Guzzi F, Ambrosio M, Rovati G, Parenti M (2011) Light on the structure of thromboxane A(2) receptor heterodimers. Cell Mol Life Sci 68 (18):3109–3120
68. Fanelli F, Seeber M, Felline A, Casciari D, Raimondi F (2013) Quaternary structure predictions and structural communication features of GPCR dimers. In: Oligomerization in health and disease, vol 117. pp 105–142
69. Nichols SE, Hernandez CX, Wang Y, McCammon JA (2013) Structure-based network analysis of an evolved G protein-coupled receptor homodimer interface. Protein Sci 22(6):745–754
70. Blacklock K, Verkhivker GM (2013) Experimentally guided structural modeling and dynamics analysis of Hsp90-p53 interactions: allosteric regulation of the Hsp90 chaperone by a client protein. J Chem Inf Model 53(11):2962–2978
71. Blacklock K, Verkhivker GM (2014) Allosteric regulation of the Hsp90 dynamics and stability by client recruiter cochaperones: protein structure network modeling. PLoS One 9(1):e86547

72. Blacklock K, Verkhivker GM (2014) Computational modeling of allosteric regulation in the Hsp90 chaperones: a statistical ensemble analysis of protein structure networks and allosteric communications. PLoS Comput Biol 10(6):e1003679
73. Tse A, Verkhivker GM (2015) Molecular dynamics simulations and structural network analysis of c-Abl and c-Src kinase core proteins: capturing allosteric mechanisms and communication pathways from residue centrality. J Chem Inf Model 55(8):1645–1662
74. Martin AJM, Vidotto M, Boscariol F, Di Domenico T, Walsh I, Tosatto SCE (2011) RING: networking interacting residues, evolutionary information and energetics in protein structures. Bioinformatics 27(14):2003–2005
75. Hu G, Yan WY, Zhou JH, Shen BR (2014) Residue interaction network analysis of Dronpa and a DNA clamp. J Theor Biol 348:55–64
76. Kar G, Kuzu G, Keskin O, Gursoy A (2012) Protein-protein interfaces integrated into interaction networks: implications on drug design. Curr Pharm Des 18(30):4697–4705
77. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7(3):188–197
78. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41(Database issue):D808–D815
79. Li Y, Vongsangnak W, Chen L, Shen B (2014) Integrative analysis reveals disease-associated genes and biomarkers for prostate cancer progression. BMC Med Genomics 7(Suppl 1):S3
80. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A (2005) PRISM: protein interactions by structural matching. Nucleic Acids Res 33:W331–W336
81. Cukuroglu E, Gursoy A, Nussinov R, Keskin O (2014) Non-redundant unique interface structures as templates for modeling protein interactions. PLoS One 9(1):e86738
82. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A (2014) PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res 42(W1):W285–W289
83. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T et al (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490(7421):556–560
84. Bahar I, Cheng MH, Lee JY, Kaya C, Zhang S (2015) Structure-encoded global motions and their role in mediating protein-substrate interactions. Biophys J 109(6):1101–1109
85. Hwang YC, Lin CF, Valladares O, Malamon J, Kuksa PP, Zheng Q, Gregory BD, Wang LS (2015) HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics 31(8):1290–1292

Adv Biochem Eng Biotechnol (2017) 160: 75–102
DOI: 10.1007/10_2016_42

Published online: 26 October 2016

Cyanobacterial Biofuels: Strategies and Developments on Network and Modeling

Amornpan Klanchui, Nachon Raethong, Peerada Prommeenate, Wanwipa Vongsangnak, and Asawin Meechai

Abstract Cyanobacteria, the phototrophic microorganisms, have attracted much attention recently as a promising source for environmentally sustainable biofuels production. However, barriers for commercial markets of cyanobacteria-based biofuels concern the economic feasibility. Miscellaneous strategies for improving the production performance of cyanobacteria have thus been developed. Among these, the simple ad hoc strategies resulting in failure to optimize fully cell growth coupled with desired product yield are explored. With the advancement of genomics and systems biology, a new paradigm toward systems metabolic engineering has been recognized. In particular, a genome-scale metabolic network reconstruction and modeling is a crucial systems-based tool for whole-cell-wide investigation

A. Klanchui
Biological Engineering Program, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

N. Raethong
Interdisciplinary Graduate Program in Bioscience, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

P. Prommeenate
Biochemical Engineering and Pilot Plant Research and Development (BEC) Unit, National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, King Mongkut's University of Technology Thonburi, Bangkok 10150, Thailand

W. Vongsangnak
Department of Zoology, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

Computational Biomodelling Laboratory for Agricultural Science and Technology (CBLAST), Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

A. Meechai (✉)
Department of Chemical Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
e-mail: asawin.mee@kmutt.ac.th

and prediction. In this review, the cyanobacterial genome-scale metabolic models, which offer a system-level understanding of cyanobacterial metabolism, are described. The main process of metabolic network reconstruction and modeling of cyanobacteria are summarized. Strategies and developments on genome-scale network and modeling through the systems metabolic engineering approach are advanced and employed for efficient cyanobacterial-based biofuels production.

Keywords Biofuels, Cyanobacteria, Genome-scale metabolic model, Metabolic network reconstruction, Systems metabolic engineering

Contents

1 Introduction 76
2 Miscellaneous Strategies Toward Enhancement of Cyanobacterial Biofuels Production 78
2.1 Traditional Strategies 78
2.2 Advanced Strategies 80
3 Development of Cyanobacterial Metabolic Network and Modeling 82
3.1 Construction of Cyanobacterial Genome-Scale Metabolic Model 83
3.2 Modeling Aided Strategies for Cyanobacterial Biofuels Production 88
4 Conclusions and Perspectives 93
References 94

1 Introduction

The rapid consumption of global energy has caused an environmental crisis and fossil fuel depletion. In turn, the need for sustainable biofuel production and development has become clear. Gradually, photosynthetic organisms are being promoted as they can recycle the greenhouse gas emitted from daily activities into a usable form of energy known as biofuel [1, 2]. Biofuel is defined as gaseous, liquid, or solid fuels produced directly or indirectly from organic matter. The first generation of biofuel is basically derived from food crops such as oil-palm, soybean, corn, and sugarcane. However, the production process leads to potential stress involving issues of land, water, and food scarcity [3, 4]. To overcome the increasing controversy in terms of 'food vs. fuel', the second generation of biofuel has developed using non-food lignocellulosic materials, which include agricultural residues, municipal and industrial wastes, and grasses. Unfortunately, this generation appears unsustainable because it requires high energy intensive conversion processes leading to increased CO_2 emission [5]. Further, the third and fourth generations involve more advanced technologies. These generations produce algae-based biofuels, which have great potential to capture and reduce CO_2 in the global atmosphere [6]. The third generation of biofuel involves improvement of biomass yield via the cultivation process and the fourth generation aims to use metabolic engineering and post-genomics tools for enhancing algae-to-biofuels production [7].

Considering algae, they use natural sunlight, CO_2, water, and nutrients to make their own biomass and hence carbon-based biofuels. The mechanism of photosynthesis in algae is similar to that in higher plants. However, algae have a distinctive growth yield, efficient CO_2 fixation, and less land requirements compared to terrestrial crops [8–10]. Algae are comprised of two major groups, namely multicellular macroalgae (e.g., seaweeds) and unicellular microalgae. One example of microalgae, often called cyanobacteria (blue-green algae), is known to produce a crucial amount of oxygen (around 30%) on Earth [11]. They can grow in a variety of habitats including aquatic and terrestrial environments [12]. Their capacity for oxygen production through a unique carbon-concentrating mechanism is appealing for enhancing photosynthetic CO_2 fixation in crops [13, 14]. Compared to the other algae, the cultivation and transformation systems of cyanobacteria have progressed and been extensively developed [15–18]. Regarding the trends and progress of the fourth biofuel generations through genetic engineering, cyanobacteria have recently been studied as the potential cell factory for supporting energy needs with economic and environmental sustainability.

Despite their great potential, the big challenge is how to increase productivity and lower cost in order to compete economically with fossil fuels. To address these obstacles, strain isolation and improvement, cultivation optimization, nutrient utilization, and downstream processing have been developed [19, 20]. However, several technical bottlenecks exist through using only traditional biological techniques resulting in trial-and-error solutions [7]. With recent advances in genome sequencing and the emergence of systems metabolic engineering, development of genome-scale cellular networks and modeling serve as a key tool for understanding the genotype–phenotype relationship. The genome-scale metabolic models (GEMs) are derived from genome information, biochemical characterization, and multilevel omics data. GEMs, thereby, offer insight into cellular function and organization at a molecular level. Moreover, they also provide a technological framework to accelerate the modification of existing pathways and the creation of new pathways to obtain desired products [21–23]. Hence, employment of GEMs through systems metabolic engineering provides a promising technology for strain design strategies for cyanobacterial biofuels production.

In the first section the different strategies using traditional and advanced approaches toward enhancement of cyanobacterial biofuels production are described. Emphasizing the advanced approach, the development of cyanobacterial metabolic network and modeling is later discussed. In the last section the challenges and directions for cyanobacteria improvement as versatile cellular factories for biofuels through the systems metabolic engineering strategies are discussed.

2 Miscellaneous Strategies Toward Enhancement of Cyanobacterial Biofuels Production

Cyanobacteria are promising sources for renewably produced fuel because of their key advantages, such as fast growth, high photosynthetic efficiency, genetic tractability, and genome availability. Cyanobacterial cells exhibit certain properties that are able to directly produce and secrete various important biochemical and biofuel feedstocks, for example isobutyraldehyde [24], isobutanol [24], 2,3-butanediol [25], 1-butanol [26], 2-methyl-1-butanol [27], acetone [28], ethylene [29], and fatty acids [30]. Moreover, the biomass itself is also considered to be a suitable raw material for sustainable biofuels production [15]. Despite their great potential, the big challenge is how to increase biofuels productivity and compete with the low cost of fossil fuels. To overcome the limitations and challenges, different strategies developed to improve cyanobacterial biofuels production, categorized as traditional (e.g., cultivation process and genetic modification) and advanced (e.g., systems metabolic engineering) strategies are described. To present biofuels production with miscellaneous strategies under different growth conditions, five different types of biofuels, namely biodiesel, bioethanol, biogas, biohydrogen, and bioelectricity are selected for discussion as listed in Table 1.

2.1 *Traditional Strategies*

The process of cyanobacteria cultivation demands favorable environmental conditions including light, nutrient, salinity, temperature, and pH. An adjustment of these factors could impact their biomass composition [52]. Several pieces of research showed that the improvement of cyanobacterial biomass by cultivation techniques has been used for cyanobacterial bioethanol and biogas production as listed in Table 1. Markou et al. [53] reported that nutrient limitation, especially phosphorus, is one of the most influential factors enhancing glycogen accumulation in *Arthrospira platensis*. Subsequently, Aikawa et al. [54] demonstrated efficient bioethanol production using a direct conversion method of this glycogen-enriched cyanobacterium. Greater amounts of lipid or carbohydrate content (up to 60–65% of dry weight) were also observed in other cyanobacteria under stress conditions [52, 55]. Growth environments and media can also enhance the productivity of biohydrogen, for example. *Cyanothece* sp. ATCC 51142 [41] and *Anabaena cylindrica* [44]. Moreover, alteration of growth conditions was also applied in the production of biodiesel and bioelectricity (Table 1). The technique for cell cultivation under stress conditions may become an interesting strategy to generate potential biofuels feedstocks. However, a serious concern for the cultivation of cyanobacteria under stress is growth rate restriction, which results in a dramatic decrease in total biomass production.

Table 1 List of miscellaneous strategies for improving cyanobacterial biofuels productivity under different growth conditions

Biofuels	Cyanobacterial strains	Strategy for biofuels production	Biofuel productivity	References
Biodiesel	*Synechocystis* sp. PCC 6803	Genetic modification/Δ phaAB, sll1951, Δslr2001-slr2002, Δslr1710, Δslr2132	197 ± 14 mg/L	[30]
	Synechococcus elongatus PCC 7942	Genetic modification/tesA and Δaas, Promoter trc	80 ± 10 mg/ DCW	[31]
	Arthrospira platensis	Cultivated in nitrogen deprivation	8% increased	[32]
Bioethanol	*Synechocystis* sp. PCC 6803	Metabolic engineering targeted pdc and slr1192; ΔphaAB, Promoter rbc	5.50 g/L	[16]
	Synechococcus elongatus PCC 7942	Synthetic metabolic pathway	26.5 mg/L	[33]
	Synechococcus sp. PCC 7002	Cultivated in nitrate limitation	30 g/L	[34]
	Synechococcus elongatus PCC 7942	Genetic modification/pdc and adhII, Promoter rbcLS	54 nmol/L/day	[35]
	Arthrospira platensis	Cultivated in stress condition	1.08 g/L/day	[36]
	Arthrospira platensis	Mutagenesis	20% increased	[37]
Biohydrogen	*Synechocystis* sp. PCC 6803	Genetic modification/Δ narB, ΔnirA	186 nmol/mg Chl-a/h	[38]
	Synechococcus sp. PCC 7002	Genetic modification/ΔldhA	14.1 mol/day/ 1017 cells	[39]
	Synechococcus elongatus PCC 7942	Genetic modification/ hydEF, hydG, hydA, Promoter psbA1, lac	2.8 μmol/h/mg Chl-a	[40]
	Cyanothece sp. ATCC 51142	Cultivated in continuous light	300 μmol H_2/ mg Chl-a/h	[41]
	Arthrospira maxima	Batch culture	400 μmol/L/h	[42]
	Arthrospira platensis	Anaerobic in the dark	1 μmol H_2/ 12 h/mg	[43]
	Anabaena cylindrical	Cultivated in light limitation	30 ml H_2/L/h	[44]
Biogas	*Arthrospira maxima*	Biomass feedstock	0.4 L/day	[45]
		Biomass feedstock	350 ml CH_4 gVS^{-1}	[46]
	Arthrospira platensis	Biomass feedstock	293 ml CH_4 gVS^{-1}	[47]
		Biomass feedstock	203 ml CH_4 gVS^{-1}	[36]

(continued)

Table 1 (continued)

Biofuels	Cyanobacterial strains	Strategy for biofuels production	Biofuel productivity	References
Bioelectricity	*Synechocystis* sp. PCC 6803	CO_2 limitation and excess light	5 mA/m^2	[48]
	Synechococcus elongatus PCC 7942	Cultivated in light	0.3–0.4 W/m^2	[49]
	Anabaena variabilis M-2	Cultivated in anaerobic	0.4 V	[50]
	Taihu Lake cyanobacteria	SMFC in acidic fermentation broth	72 mW/m^2	[51]

Recently, genetic engineering has been employed for improving biofuel production. Conventional gene modification is a direct change of genetic material of interest without consideration of other biological elements. Attempts to increase biofuel content by means of genetic change were presented in the reviews by Rosgaard et al. [56]. A number of exogenous gene transfer methods have been investigated and progressed in cyanobacteria, mainly including natural transformation, electroporation, conjugation, and particle guns [57–59]. The continuous development of these genetic tools are used to generate transformants for enhancing cyanobacterial biofuels, namely biodiesel, bioethanol, biogas, biohydrogen, and bioelectricity production, such as *Synechocystis* sp. PCC 6803 [30, 38], *Synechococcus elongatus* PCC 7942 [31, 35], and *Synechococcus* sp. PCC 7002 [39, 60] as also seen in Table 1. Nonetheless, there are some cyanobacterial trains, particularly the important commercial genus *Arthrospira*, which are resistant to common genetic modification techniques [61, 62]. Therefore, a genetic control system for the desired pathways of cyanobacteria is needed for establishment of transformation protocols suitable for specific desired products and certain different types of species.

2.2 Advanced Strategies

The emergence of metabolic engineering has enabled the improvement of cyanobacterial strains for bioethanol production. With the recent wealth of genomics and post genomics data available, there is growing interest in integrating the conventional metabolic engineering and multi-level omics data in the discipline of systems biology. With the paradigm shift, metabolic engineering has evolved into systems metabolic engineering with a systems-level understanding of the cellular function [63, 64]. Systems metabolic engineering, which deals with analysis, design, and synthesis with integrative systems and synthetic biology, was employed to make strain engineering efficient for cyanobacterial biofuels production [65–67]. Therefore, systems metabolic engineering offers a conceptual and technological framework to speed the optimization of cyanobacterial biofuels production as depicted in Fig. 1. It is clear that this powerful framework is applied based on the multi-level omics data and computational model known as GEMs.

In terms of genomics data, there are 301 cyanobacteria genomes deposited from the NCBI database (August 2015). Of these, 95 species are reported for completed genomes (http://www.ncbi.nlm.nih.gov/bioproject). Based on genomic information, the genome-wide comparative analysis was performed to identify the genetic conservation and the variation, together with metabolic diversity [68, 69]. This resulted in feasible designed cyanobacterial strains with the potential to improve biofuels production. There has also been research on transcriptomics conducted to understand the global response of cyanobacteria metabolism [70, 71]. Using

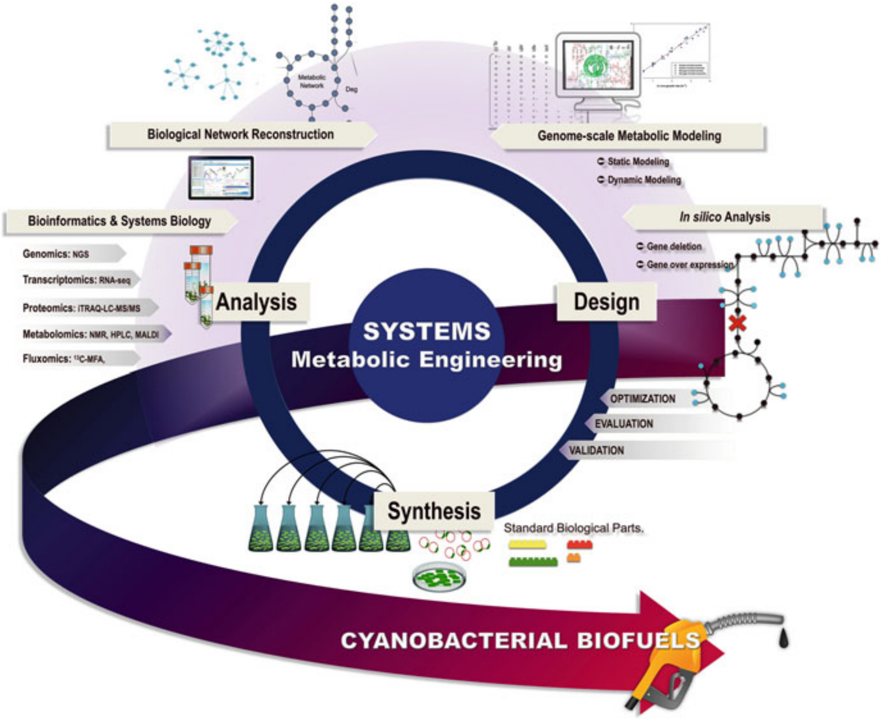


Fig. 1 Diagram illustrating the revolution of systems metabolic engineering for strain design strategies toward improving cyanobacterial biofuels production. GEMs are reconstructed based on integrative multi-level omics data. The cyanobacterial model is then applied to analyze the metabolic capabilities in the discipline of systems biology. Once the metabolic engineering targets are identified, the results of analysis are subsequently employed to guide in the cyanobacterial strains design. Finally, the cyanobacterial strains are synthesized according to the in silico analysis through synthetic biology. The cyanobacterial strains leave the cycle whenever they show desired phenotypes fitting with the requirement of industrial biotechnology

high-resolution DNA microarrays, the transcriptomics study of engineered *Synechocystis* sp. PCC 6803 strain cultured under continuous biofuel/ethanol production in fully automated photobioreactors (PBRs) was investigated. The transcriptomics results afterward provided in-depth characterization of the cellular response of long-term ethanol production in *Synechocystis* sp. PCC 6803 [72]. Using transcriptomics RNA-sequencing with the gene knockout technique, the possible target genes involved in ethanol tolerance in *Synechocystis* sp. PCC 6803 were presented [65]. This indicates that transcriptomics can be used to reveal key expressed genes associated with bioethanol production in cyanobacteria.

With regard to proteomics, it has been performed to elucidate the cellular responses to biofuel stress in cyanobacteria, which proves challenging for the host's tolerance to the toxic biofuels [73, 74]. There are several conditions using the proteomic approach for studying expressed protein in the cells treated with ethanol [75], butanol [76], hexane [77], salt [78], and subjected to N-starvation [79]. The report presented by Pei et al. [80] showed that the constructed protein network identified a core set of proteins that commonly respond to both biofuel stress and environmental perturbation. This result was previously classified as a core transcriptional response (CTR) by Sigh et al. [81]. Hence, the results from multi-level omics studies provide more understanding in the cellular process and underlined molecular keys for making sustainable fuels from cyanobacteria.

With regard to GEMs, large-scale metabolic networks and models serve as the crucial tools for systems biology. They represent information infrastructures describing the whole relationship of gene-protein-reaction in cells. They offer not only powerful analytical tools for quantitative, structural, and design analysis of cellular metabolism, but also provide a computational framework to integrate high-throughput datasets [82]. For this reason, GEMs have been extensively utilized in systems metabolic engineering strategies [21, 83]. Successful applications of GEMs for improving the desired products have been widely reported for several industrial microorganisms, particularly, *Escherichia coli* [84]. To date, an overall increasing numbers of cyanobacterial models have been published as shown schematically in Fig. 2. This dramatic progress enables researchers to gain insights into metabolic capacity and develop strategies for manipulation of cyanobacterial metabolism and regulation. However, the comparative characteristics of each cyanobacterial metabolic model are different in aspects of the total number of genes, reactions, metabolites, and cellular compartments (Table 2). In the following sections, the descriptions of how to develop and utilize cyanobacterial models as a major tool for understanding whole-cell-wide metabolism toward biofuels production are highlighted.

3 Development of Cyanobacterial Metabolic Network and Modeling

Cyanobacterial metabolism is the central biochemical machinery that enables biological structures to sustain their cellular functions. A better understanding of the metabolism of cyanobacteria plays a pivotal role in optimizing cell growth and

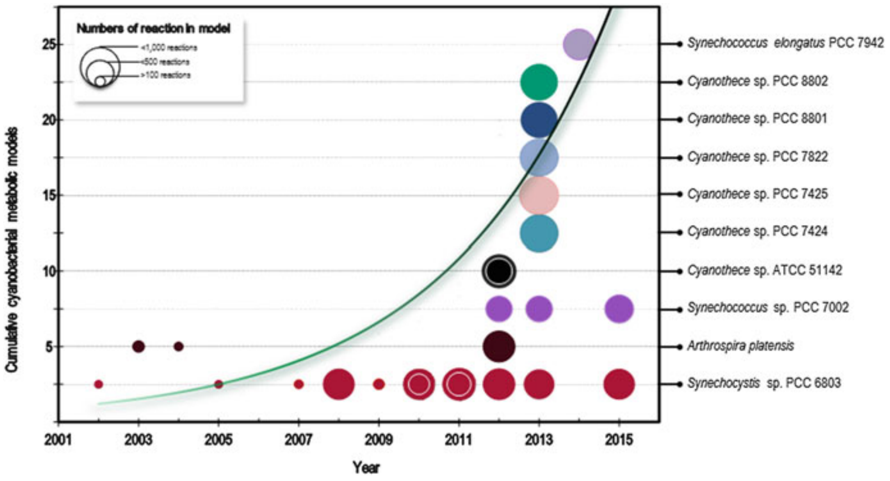


Fig. 2 Historical timeline and increase in the number of developed cyanobacterial metabolic models. Each *circle* represents an individual cyanobacterial model along with the different circle sizes indicating the numbers of reaction in the model. In particular, cyanobacterial strains are discriminated in the respective *lines* and marked with *different colors*

biofuel yields. However, unlike the heterotrophs that utilize single organic compounds as sources of carbon and energy, cyanobacteria make their own food by utilizing light for energy and inorganic carbon, for example CO_2 or bicarbonate, as a carbon source. Furthermore, cyanobacteria also perform a circadian cycle with daily light. With this double complex and specific mechanism of cyanobacteria, the GEMs have been developed for a system-level understanding of cyanobacterial metabolism. In this section we review the reconstruction of genome-scale metabolic network and modeling of cyanobacteria. A description of simulation approaches that have been used for gaining insights into cyanobacterial metabolism is provided. Strain design strategies revealed by cyanobacterial GEMs are also proposed.

3.1 Construction of Cyanobacterial Genome-Scale Metabolic Model

A typical construction process of cyanobacterial genome-scale metabolic model is divided into four main steps [107]: (1) reconstruction of a draft cyanobacterial metabolic network, (2) refinement of cyanobacterial network, (3) conversion of network to cyanobacterial model, and (4) evaluation of cyanobacterial model. For further details, the four main steps are described below and in Fig. 3.

Step 1: Reconstruction of a draft cyanobacterial metabolic network. The aim is to obtain every possible candidate for metabolic reactions and pathways without

Table 2 List of characteristics of developed cyanobacterial metabolic models

Cyanobacterial strains	Year	Model type	Gene	Reaction	Metabolite	Compartment	References
Synechocystis sp. PCC 6803	2002	S	NA	29	23	2	[85]
	2005	S	NA	70	46	2	[86]
	2007	S	78	56	63	2	[87]
	2009	S	NA	90	56	2	[88]
	2009	G	633	831	704	2	[89]
	2010	S	337	380	291	2	[90]
	2010	G	669	882	790	2	[91]
	2011	G	811	956	911	2	[92]
	2011	G	393	493	465	2	[93]
	2012	G	678	863	795	4	[94]
	2012	G	731	1,156	996	2	[95]
	2013	G	677	759	601	6	[96]
	2015	G	677	814	601	6	[97]
Arthrospira platensis	2003	S	NA	121	134	2	[98]
	2004	S	NA	22	17	2	[99]
Arthrospira platensis C1	2012	G	692	688	658	2	[100]
Synechococcus sp. PCC 7002	2012	G	611	552	542	2	[101]
	2013	G	708	648	581	2	[102]
	2015	G	706	649	542	2	[103]
Cyanothece sp. ATCC 51142	2012	G	806	667	587	2	[104]
	2012	G	773	946	811	5	[95]
Cyanothece sp. PCC 7424	2013	G	792	1,242	1,368	5	[105]
Cyanothece sp. PCC 7425	2013	G	731	1,306	1,368	5	[105]
Cyanothece sp. PCC 7822	2013	G	826	1,258	1,368	5	[105]
Cyanothece sp. PCC 8801	2013	G	752	1,172	1,368	5	[105]
Cyanothece sp. PCC 8802	2013	G	755	1,161	1,368	5	[105]
Synechococcus elongatus PCC 7942	2014	G	715	851	838	2	[106]

Model types S and G stand for small-scale and genome-scale metabolic models, respectively NA stands for not assigned

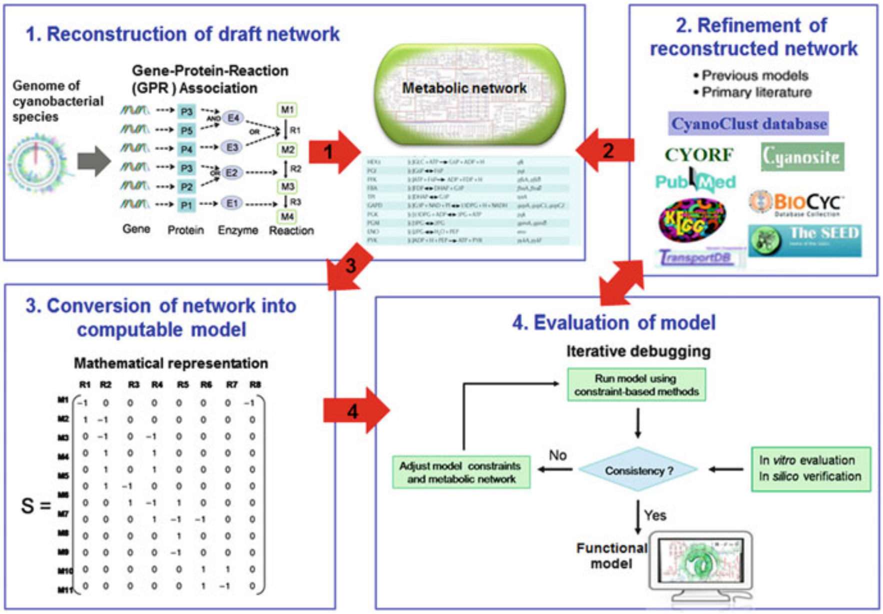


Fig. 3 Diagram illustrating four main steps for construction of a cyanobacterial genome-scale metabolic model

considering the complete reconstruction. This step starts with genome annotation. The connections between identified metabolic genes encoding enzymes and their corresponding biochemical reactions are subsequently determined. Gene-protein-reaction (GPR) associations of the network are obtained. Then all gathered GPR relationships are assembled to generate a draft network of cyanobacteria. It should be noted that the draft network may contain incorrect or missing assignments of species-specific metabolic processes because of incomplete annotation and database information. To provide information for creation of a draft cyanobacterial metabolic network, the databases and tools listed in Table 3 are usually used.

Step 2: Refinement of cyanobacterial network. After obtaining the draft cyanobacterial metabolic network it is important to perform manual curation to achieve a high-quality metabolic network. In general, each pathway within the network has to be curated in a canonical manner, from the clear pathway assignment to the ambiguous one. Certain information should be checked during performing manual correction, for example information about the reaction including reaction name, substrate and cofactor usage, balanced reaction stoichiometry, directionality, reaction identifier of the reference database, spontaneous reaction, demand and sink reaction, type of transportation, exchange reaction, ATP-maintenance reaction, and the biomass formation equation. In addition, information about the metabolite including metabolite name and abbreviation, neutral and charged formula, charge value, and metabolite identifier as well as information

Table 3 List of different databases and tools for creating cyanobacterial genome-scale metabolic model

Databases/tools	FTP links	Description	References
MetaCyc	http://metacyc.org/	A highly curated metabolic database	[108]
BioCyc	http://biocyc.org/	Organism-specific database	[108]
KEGG	http://www.genome.jp/kegg/	Database resource for cell metabolism	[109]
Reactome	http://www.reactome.org	A database of biological processes	[110]
UniProt	http://www.uniprot.org/	Protein database	[111]
CyanoBase	http://genome.microbedb.jp/cyanobase	Cyanobacterial genome database	[112]
CyanoClust	http://cyanoclust.c.u-tokyo.ac.jp	Cyanobacterial homolog proteins database	[113]
Cyanobacterial KnowledgeBase	http://nfmc.res.in/ckb/index.html	Cyanobacterial genome and proteome database	[114]
CyanoCOG	http://www2.sbi.kmutt.ac.th/orthoCOG/cyanoCOGnew/home	Cyanobacterial orthologous proteins database	[115]
CyanoPhyChe	http://bif.uohyd.ac.in/cpc	A database for Physico-chemical properties of cyanobacterial proteins	[116]
SpirPro	http://spirpro.sbi.kmutt.ac.th	Spirulina proteome database	[115]
CyanOmics	http://lag.ihb.ac.cn/cyanomics	Omics database of *Synechococcus* sp. PCC 7002	[117]
CyanoEXpress	http://cyanoexpress.sysbiolab.eu	Transcriptome database of *Synechocystis* sp. PCC 6803	[118]
ProPortal	http://proportal.mit.edu/	Cyanobacterium Prochlorococcus database	[119]
BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Functional annotation tool	[120]
Pfam	http://pfam.xfam.org/	Database and tool for protein families	[121]
PUBMED	http://www.ncbi.nlm.nih.gov/pubmed/	Literature database	[122]
Pathway Tools	http://bioinformatics.ai.sri.com/ptools/	A comprehensive symbolic systems biology software	[123]
KASS KEGG	http://www.genome.jp/tools/kaas/	KEGG automatic annotation server	[124]
Model SEED	http://www.theseed.org/wiki/Main_Page	Annotation and reconstruction tool for microbial genomes	[125]
SuBliMinal Toolbox	http://www.mcisb.org/subliminal/	Metabolic network reconstruction tool	[126]
RAVEN Toolbox	http://biomet-toolbox.org/index.php?page=downtools-raven	Reconstruction, analysis and visualization of metabolic networks	[127]

(continued)

Table 3 (continued)

Databases/tools	FTP links	Description	References
GEMSiRV	http://sb.nhri.org.tw/GEMSiRV	Reconstruction, visualization, and simulation of metabolic networks	[128]
OptFlux	http://www.optflux.org/	Metabolic modeling tool	[129]
MicrobesFlux	http://tanglab.engineering.wustl.edu/static/MicrobesFlux.html	Metabolic modeling tool	[130]
MetaNET	http://metanet.osdd.net	Metabolic modeling tool	[131]
MOST	http://most.ccib.rutgers.edu/	Metabolic modeling and strain design	[132]
COBRA Toolbox	https://opencobra.github.io/	Metabolic modeling tool	[133]
DRUM		Metabolic modeling tool for Non-Balanced Growth	[134]
DFBAlab	http://yoric.mit.edu/dfbalab	MATLAB code for dynamic flux balance analysis	[135]

of enzyme and reaction localization cellular compartments involving in the determination of for each subsystem are refined Information on missing functions/reactions may be obtained from experiments and metabolic pathway databases. Information on growth requirements is also concerned. Other issues encountered during metabolic network reconstruction are listed by Feist et al. [136]. The end of the manual refinement process results in a refined network reflecting strain specific physiology.

Step 3: Conversion of network to cyanobacterial model. Conversion of the reconstructed network of cyanobacteria is transformed to a mathematically consistent form known as the stoichiometric matrix, S ($m \times n$) (Fig. 3) [137]. This matrix is a rectangular array of stoichiometric coefficients, which are the number written in front of metabolites involved in the particular chemical reaction, arranged in m rows and n columns. Considering a simple network, m rows correspond to the number of compound species and n columns represent the number of reactions. The intersection of row and column in S expresses the relative quantity of metabolites taking part in such a reaction. After generation of S, automatically achieved using tools, a constraint-based simulation approach is applied to access the function of the reconstructed metabolic network [23]. This modeling approach provides a static model built upon principles of biological systems with physical and chemical laws. Imposition of constraints usually includes the connection of metabolites within the given system, thermodynamics (reaction reversibility), and upper and lower bounds of individual reaction fluxes. With the pseudo-steady assumption, cellular metabolites must be produced and consumed in a mass-balanced conservation with short timescales. The equation of system-wide metabolism should be written as $S \cdot v = 0$, where v is vector of conversion rate of reaction fluxes (mole. unit biomass^{-1}.h^{-1}). Flux balance analysis (FBA) is the most widely used technique for investigating metabolic state and balances on a large-scale network. In the FBA, the objective function (Z) is set to obtain a single optimal flux distribution inside the edge of the solution space. The objective function can be minimized or maximized. Using linear programming, the values of the objective function, typically the biomass constituents ($Z = v_{\text{biomass}}$), and other metabolic fluxes can be calculated [137].

Step 4: Evaluation of cyanobacterial model. This stage often involves testing and correcting. After the created GEMs are simulated, the amount of biomass growth is shown. Despite this, network gaps are common errors found when modeling large-scale networks. This result presents no growth observation. The analysis and closure of gaps in pathways becomes an intractable task. Thus, gap filling algorithms are necessary. Other modeling errors linked to incorrect reaction constraints were exchanging substance across compartments and no consumption and production of metabolites. The analysis and solution in GEMs development has been extensively reviewed [107]. To debug modeling problems, all improved data are manually added to the refined metabolic network (step 2) and then repeated in steps 3 and 4 (Fig. 3). Once the model shows biomass flux prediction, the model validation can eventually be performed using independent published experiments. These validations may result in filling phenotypic gaps and complementing additional biological information to the model. The reconstruction process can be iteratively performed until simulated data are in agreement with the experimental data and are consistent with the physiology of the cyanobacteria observed.

3.2 Modeling Aided Strategies for Cyanobacterial Biofuels Production

During 2002–2007 the metabolic network reconstruction of cyanobacteria was initially performed on a small scale (Fig. 2 and Table 2). The reconstruction was based on either inferring the enzyme information from biochemical knowledge or adapting the network of known organisms, including the metabolic models of *Synechocystis* sp. PCC 6803 [85–87] and *A. platensis* [98]. After the availability of genome sequences, genome-scale metabolic networks were created by applying the reconstruction process shown in Fig. 3. The first GEM of cyanobacteria was *Synechocystis* sp. PCC 6803 which was built and published by Fu et al. [89]. To date, more than 15 GEMs of cyanobacterial species have been developed and studied for different purposes, including *Synechocystis* sp. PCC 6803 [91–93, 95–97, 138], *A. platensis* [100], *Synechococcus elongatus* PCC 7942 [106], *Synechococcus* sp. PCC 7002 [101–103], and six *Cyanothece* strains, *Cyanothece* sp. ATCC 51142 [95, 104, 105], *Cyanothece* sp. PCC 7424, 7425, 7822, 8801, and 8802 [105] (Table 2). The release of these GEMs provides opportunity to gain new biological knowledge and assist design strategies for biofuels production.

3.2.1 Overview of the Cyanobacterial GEMs

The published GEMs share some general principles of the network pathway involved in central carbon metabolisms, namely glycolysis, citric acid cycle, and pentose phosphate pathways and photosynthesis as well as nitrogen assimilation. These cellular processes are the main components to serve the incorporation of inorganic substrates, which produce precursor metabolites and the energy for cellular functions. These core metabolic pathways are similar to the model of heterotrophic microorganisms, such as *E. coli*. However, the major difference of metabolic content in each model depends on the degree of details in cellular compartments, photosynthesis, pathway for secondary metabolites, the definition and organization within the reconstructed network, and biomass constitution.

The presence of cellular compartments in the reconstructed network has been identified based on cyanobacterial cell structure dependent on the purpose of the model. Most of the published models have at least two compartments, namely the cytosol and the extracellular space (Table 2). The increase in the number of compartments is related to the detail of photosynthesis. Nogales et al. [138] built GEM comprising four different cellular compartments of *Synechocystis* sp. PCC 6803, namely extracellular, periplasm, cytoplasm, and thylakoid. This work aims to reveal key photosynthetic processes in mechanistic detail under various lights. The photosynthesis and respiration were situated in the thylakoid membrane of this in silico model. The photosynthetic linear electron flow (LEF), photosystem I (700 nm) and photosystem II (at 680 nm), and alternate electron flow pathways (AEF) accounted for balancing the ATP/NADPH ratio [102, 104]. Nevertheless, some simple models present photophosphorylation as a single reaction, where harvested photons convert H_2O into ATP and NADP [89]. Another extensive study into cellular compartments was made by Knoop et al. [96]. They developed GEM of *Synechocystis* sp. PCC 6803 consisting of six cellular compartments, cytosol, thylakoid membrane, thylakoid lumen, plasma membrane, periplasmic space, carboxysomes, and extracellular space.

The GEMs of cyanobacteria also present a different degree of detail in secondary metabolites biosynthesis pathway. It seems to be that the most present secondary metabolite in the model is chlorophyll a [91, 98, 139]. However, researchers have attempted to incorporate the other secondary metabolites into a developing model of *Synechocystis* sp. PCC 6803 [96, 138]. Further, the degree of difference of cyanobacterial network was observed in the context of definition and organization of metabolic network within the model. For example, the reactions were duplicated when several alternative enzymes were involved [96], and some models used Boolean gene-protein-reaction rules to merge such reactions into a single one. Some metabolic networks also include the isomeric forms of metabolites, such as alpha D-glucose and beta-D-glucose. The difference in organization of the reconstructed metabolic network causes a difficulty for comparisons of multiple cyanobacterial species. Formulation of the biomass-constituting equation of the model was carried out using literature data and analytical measurements [96]. It is an approximation of the cellular macromolecule composition, such as proteins, carbohydrates, lipids, DNA, RNA, cell wall components, cofactors, and secondary metabolites identified within the cell. It has been demonstrated that cellular composition directly responds to environmental conditions and the cellular genotype [52]. As the details regarding the reactions and the metabolites in reconstructed metabolic networks are strongly related to the detail in biomass-constituting equation, the set of macromolecules constituting biomass may have an impact on accurate predictions of product production [140]. Therefore, organism-specific biomass- constituting equations are necessary, but many broad approximations can possibly be made.

3.2.2 Cyanobacterial GEM Simulations

Even though the models have been constructed in great detail for the different purposes of each study, similar simulations were performed by way of a static approach under three metabolic modes, namely autotrophy, heterotrophy, and mixotrophy. This allows a better understanding of which cellular processes enable cyanobacteria to live in a broad variety of environmental conditions. Most of the studies have applied Flux Balance Analysis (FBA) [141] with maximized biomass formation as an objective function to simulate the model. Additionally, Flux Variability Analysis (FVA) [142] is usually used to determine the maximum and minimum flux values found in FBA and to identify the blocked or non-essential reactions of the metabolic models. However, the balanced growth assumption of FBA cannot handle some bioprocesses, such as the circadian cycle where some internal metabolites are accumulated and consumed under day/night cycles. Thus, the investigation of metabolic state under these dynamic conditions was proposed for growth under day/night cycles [90]. In the following we review various simulation studies of the cyanobacterial GEMs under the three principle metabolic modes by using static modeling and the dynamic modeling approaches. A system-level understanding of cyanobacterial metabolism supported by GEMs allows systems metabolic engineering strategies to optimize biofuel production.

3.2.3 Autotrophy

Photoautotrophic condition is characterized by simulation where energy and carbon sources come from light and CO_2, respectively. All GEMs are demonstrated by simulating autotrophic growth to represent their basic metabolic capabilities of cellular systems. The availability of light intensity is represented by photon flux. Generally, a constant number of photons used per photosystem are assumed [86, 87, 89]. In fact, this quantity can be affected by the physiological status of the cellular harvesting processes where the energy of photons is captured and transferred via pigment to the proteins of the reaction centers. Moreover, the absorption of a photon depends on a particular wavelength, which can fluctuate throughout the day. Simulation results of light influence upon the cell growth suggesting a high impact on qualitative flux distribution was observed under excessive light and limiting carbon sources [138]. To meet energy and carbon cellular requirements, AEF was developed to restore ATP/NADPH ratio [138] and act as energy valves in the case of excessive light [102, 104]. AEF is represented as either cyclic of electron flow (ferredoxin plastoquinone reductase), the NADPH dehydrogenase complexes, plastoquinone oxidase, cytochrome oxidase, or the Mehler and hydrogenase reactions [143]. Typically, the ATP/NADPH ratio was set in a genome-scale metabolic model since the fixation of one molecule of CO_2 through the Calvin cycle in cyanobacteria required 3 ATP and 2 NADPH [89]. Another interesting simulation was performed in terms of the effect of the light quality on the autotrophic metabolism. However, an attempt to develop the detail of light harvesting mechanisms was performed in green algae, *Chlamydomonas* [144]. This would be the first step in challenging the light harvesting process model in cyanobacterial GEMs. As main cellular metabolism is related to photosynthesis, it is necessary to expand the explicit action of photon harvesting, chlorophyll fluorescence, photoinhibition, photoacclimation, and non-photochemical quenching.

3.2.4 Heterotrophy

Although cyanobacteria are photoautotrophs, some of them have the capability for heterotrophic growth in the dark, supported by an organic carbon source. However, how cyanobacteria regulate this heterotrophic activity still remains largely unknown. Metabolic flux predictions under heterotrophic conditions are of considerable interest for almost all cyanobacterial GEMs. The significant differences between autotrophy and heterotrophy are carbon source types. Glucose, acetate, and glycogen are the major sources of carbon for in silico simulation. Based on a given set of constraints on the exchange rate of nitrate, phosphate, sulfur, and possible external parameters, the maximal growth yield as well as flux distribution can be obtained. Under heterotrophic growth, simulation results showed that the reaction fluxes in the glycolysis pathway move forward to synthesize precursor metabolites for downstream pathways. Comparing the metabolic state between autotrophy and heterotrophy, the degree of active reactions are different in the central carbon pathways (glycolysis, TCA cycle, Calvin cycle, and pentose phosphate pathway) whereas the synthesis of amino acids, lipids, DNA, and RNA does not vary significantly [100, 145]. The main energy source was produced from glycolysis, the TCA cycle, and the oxidative pentose phosphate pathway instead of the photosynthesis system and AEF. Highly active fluxes of the oxidative pentose phosphate process indicated that NADPH was the key metabolite under heterotrophic growth. Moreover, results showed that dark respiration utilized carbon at approximately 40% [91].

3.2.5 Mixotrophy

Mixotrophy is a combination of metabolic modes where an organism can obtain its energy from light, carbon dioxide, and sugars. This is also a common phenomenon in cyanobacteria, in particular response from environments under light- or nutrient-limitation [146]. Understanding the ability to switch between autotrophy and mixotrophy is now being recognized and challenged [147]. Computational analysis revealed that the metabolic state of the cell system under mixotrophic conditions is varied between autotrophy and heterotrophy [100]. However, the results also depend on the ratio of light and organic carbon set for the simulation [93, 104]. In addition, Knoop et al. [90] showed that flux distribution resulting from simulation during mixotrophic growth agrees well with experimental analysis. An illustration of an autotrophic, heterotrophic, and mixotrophic central metabolism flux map of the GEM of *Synechocysti*s sp. PCC 6803 is provided by Baroukh et al. [148].

3.2.6 Day/Night Cycle

Although the FBA could fulfill basic insights into the three metabolic modes in cyanobacteria, it cannot be used to capture the transition of metabolic process of the cellular system. Knoop et al. [90] attempted to simulate a day/night cycle through FBA by decreasing the light intensity and increasing the carbon source at the same time. The results showed a shift in metabolic state between autotrophy- and heterotrophy-like conditions. Currently, a framework for providing dynamic metabolic modeling has been developed (Table 3) such as Dynamic Flux Balance Analysis (DFBA) [149]. However, DFBA is also based on the assumption that there is no intracellular accumulation of compounds to circumvent the large accumulation of particular metabolites during the day and their consumption during the night. Knoop et al. [90] employed DFBA to compute dynamic metabolic fluxes for a full diurnal cycle. In simulation, a different biomass composition corresponding to the metabolism shifted from a night-time, heterotrophic metabolism, to a day-time, autotrophic metabolism, was used for predicting all metabolic flux dynamics. Another modeling framework named DRUM [134] was developed based on Elementary Flux Mode analysis. This technique splits the full network into sub-networks and allows the accumulation and generation of connected metabolites in each sub-network. The software was used to demonstrate the metabolic flux of lipid and carbohydrate accumulation under the diurnal cycle of *Tisochrysis lutea*. In the context of cyanobacterial biofuel production, dynamic metabolic modeling is crucial in understanding the outdoor dynamics of cell metabolism.

3.2.7 Modeling Toward Strain Design Strategies for Biofuel Production

To investigate the cyanobacterial metabolism provided by GEMs, Erdrich et al. [150] used the GEM of *Synechocystis* sp. PCC 6803 to identify and characterize systematically suitable strain design strategies for ethanol and isobutanol synthesis. This work demonstrated that cyanobacterial GEM has been developed and applied as a vital tool for rational metabolic engineering to improve biofuel production. The team utilized the original GEM of *Synechocystis* sp. PCC 6803 published by Knoop et al. [96]. The alternate electron transport pathways and reactions of both ethanol and isobutanol synthesis pathway and transportation were added. They used two different computational methods, CASOP [151] and Constrained Minimal Cut Sets [152], to identify intervention strategies that enforce coupled biomass and high-yield product synthesis under phototrophic growth concerning the diurnal rhythm of cyanobacteria. This research revealed the suitable knockout gene set target routes to reduce the ratio of ATP/NADPH in the photosynthetic electron transport chain. Here, proof-of-concept in using cyanobacterial GEM toward biofuel production has been established. However, further development of tailored modeling approaches is of crucial importance for gaining insight to cyanobacterial metabolism and supporting strain design strategies.

4 Conclusions and Perspectives

With the amount of fossil fuels continually decreasing, biofuels could soon become vital sources of the world’s energy. Much research has been carried out seeking ways to produce biofuels that are economically feasible. Not until recently did cyanobacteria become leading candidates as excellent sources for biofuels production because they can simply take free energy from sunlight and atmospheric carbon dioxide and subsequently convert them into valuable fuels, namely biodiesel, bioethanol, biohydrogen, biogas, and bioelectricity. Cyanobacteria can also be used as cell factories for production of bio-based chemicals, that is, short chain alcohols. Although not yet feasible in term of production cost, we believe that cyanobacteria have far greater advantages over other organisms in many ways and they are worthy of investing more research efforts. Systems metabolic engineering (Fig. 4) offers great promise for rational strain improvement of cyanobacteria with desirable phenotype, overproduction of biofuels in this case. This approach, considered as a means for the fourth generation of biofuels production, includes working in a cycle of analysis, design, and synthesis steps. It can be realized by the employment of cyanobacterial genomes to reconstruct a precise network which is then converted to a GEM. Along with multi-level omics information, a cyanobacterial GEM helps facilitate systematic analysis of biofuel pathways and in silico prediction of metabolic capabilities of various designed strains. Next, some selected designs guided by the model are subject to further strain construction in the

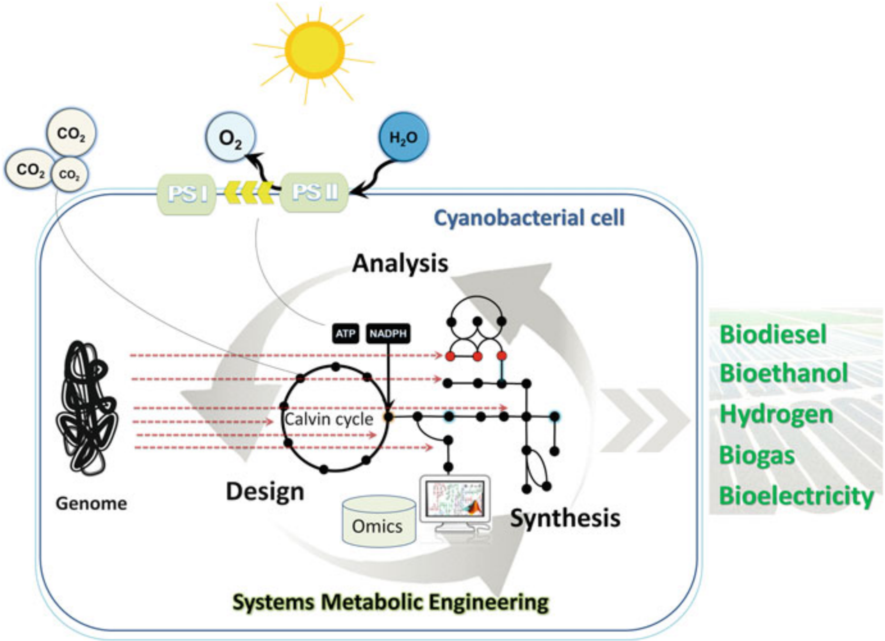


Fig. 4 The ideal engineered fuel cell factory driven by network and computational modeling for fourth generation biofuel production

laboratory, aiming to obtain super-cyanobacterial strains. Though the future of cyanobacteria can be bright, as witnessed by the availability of many cyanobacterial GEMs, there are still some challenges needing to be addressed, especially molecular biology tools for genetic construction of engineered strains. Nevertheless, with recent efforts in synthetic biology field, we believe that efficient genetic tools for simple construction of genetically engineered cyanobacteria can soon be made available.

Acknowledgements The authors acknowledge the financial support provided by King Mongkut's University of Technology Thonburi through the 'KMUTT 55th Anniversary Commemorative Fund', the National Center for Genetic Engineering and Biotechnology (BIOTEC), NSTDA, Thailand (P-11-01089), Science Achievement Scholarship of Thailand (SAST), The Thailand Research Fund (TRG5880245), and Preproposal Research Fund (PRF4/2558 and PRF-PII/59) and Department of Zoology, Faculty of Science, Kasetsart University.

References

1. Sivakumar G, Vail DR, Xu JF, Burner DM, Lay JO, Ge XM, Weathers PJ (2010) Bioethanol and biodiesel: alternative liquid fuels for future generations. Eng Life Sci 10(1):8–18
2. Mussatto SI, Dragone G, Guimaraes PM, Silva JP, Carneiro LM, Roberto IC, Vicente A, Domingues L, Teixeira JA (2010) Technological trends, global market, and challenges of bio-ethanol production. Biotechnol Adv 28(6):817–830

3. Naik SN, Goud VV, Rout PK, Dalai AK (2010) Production of first and second generation biofuels: a comprehensive review. Renew Sustain Energy Rev 14(2):578–597
4. Mohr A, Raman S (2013) Lessons from first generation biofuels and implications for the sustainability appraisal of second generation biofuels. Energy Policy 63(100):114–122
5. Sanderson K (2011) Lignocellulose: a chewy problem. Nature 474(7352):S12–S14
6. Sayre R (2010) Microalgae: the potential for carbon capture. BioScience 60(9):722–727
7. Lü J, Sheahan C, Fu P (2011) Metabolic engineering of algae for fourth generation biofuels production. Energy Environ Sci 4:2451–2466
8. Dismukes GC, Carrieri D, Bennette N, Ananyev GM, Posewitz MC (2008) Aquatic phototrophs: efficient alternatives to land-based crops for biofuels. Curr Opin Biotechnol 19(3):235–240
9. Lee RA, Lavoi J-M (2013) From first- to third-generation biofuels: challenges of producing a commodity from a biomass of increasing complexity. Anim Front 3(2):6–11
10. Maity JP, Bundschuh J, Chen C-Y, Bhattacharya P (2014) Microalgae for third generation biofuel production, mitigation of greenhouse gas emissions and wastewater treatment: present and future perspectives–a mini review. Energy 78:104–113
11. Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. Nature 455:1101–1104
12. Büdel B (2011) Cyanobacteria: habitats and species. In: Lüttge U, Beck E, Bartels D (eds) Plant desiccation tolerance, ecological studies, vol 215. Springer, Heidelberg, pp 11–21
13. Price GD, Pengelly JJ, Forster B, Du J, Whitney SM, von Caemmerer S, Badger MR, Howitt SM, Evans JR (2013) The cyanobacterial CCM as a source of genes for improving photosynthetic CO_2 fixation in crop species. J Exp Bot 64(3):753–768
14. McGrath JM, Long SP (2014) Can the cyanobacterial carbon-concentrating mechanism increase photosynthesis in crop species? A theoretical analysis. Plant Physiol 164 (4):2247–2261
15. Quintana N, Van der Kooy F, Van de Rhee MD, Voshol GP, Verpoorte R (2011) Renewable energy from cyanobacteria: energy production optimization by metabolic pathway engineering. Appl Microbiol Biotechnol 91(3):471–490
16. Gao ZX, Zhao H, Li ZM, Tan XM, Lu XF (2012) Photosynthetic production of ethanol from carbon dioxide in genetically engineered cyanobacteria. Energy Environ Sci 5:9857–9865
17. Nozzi NE, Oliver JWK, Atsumi S (2013) Cyanobacteria as a platform for biofuel production. Front Bioeng Biotechnol 1:7
18. Taton A, Unglaub F, Wright NE, Zeng WY, Paz-Yepes J, Brahamsha B, Palenik B, Peterson TC, Haerizadeh F, Golden SS, Golden JW (2014) Broad-host-range vector system for synthetic biology and biotechnology in cyanobacteria. Nucleic Acids Res 42(17), e136
19. Hannon M, Gimpel J, Tran M, Rasala B, Mayfield S (2010) Biofuels from algae: challenges and potential. Biofuels 1(5):763–784
20. Singh RK, Tiwari SP, Rai AK, Mohapatra TM (2011) Cyanobacteria: an emerging source for drug discovery. J Antibiot 64(6):401–412
21. Kim B, Kim WJ, Kim DI, Lee SY (2015) Applications of genome-scale metabolic network model in metabolic engineering. J Ind Microbiol Biotechnol 42(3):339–348
22. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10(4):291–305
23. Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 15:107–120
24. Atsumi S, Higashide W, Liao JC (2009) Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. Nat Biotechnol 27(12):1177–1180
25. Olivera JWK, Machadoa IMP, Yonedaa H, Atsumia S (2013) Cyanobacterial conversion of carbon dioxide to 2,3-butanediol. Proc Natl Acad Sci U S A 110:1249–1254
26. Lana EI, Liao JC (2012) ATP drives direct photosynthetic production of 1-butanol in cyanobacteria. Proc Natl Acad Sci U S A 109:6018–6023

27. Shen CR, Liao JC (2012) Photosynthetic production of 2-methyl-1-butanol from CO_2 in cyanobacterium Synechococcus elongatus PCC7942 and characterization of the native acetohydroxyacid synthase. Energy Environ Sci 5:9574–9583
28. Zhou J, Zhang H, Zhang Y, Li Y, Ma Y (2012) Designing and creating a modularized synthetic pathway in cyanobacterium Synechocystis enables production of acetone from carbon dioxide. Metab Eng 14(4):394–400
29. Ungerer J, Tao L, Davis M, Ghirardi M, Maness P-C, Yu J (2012) Sustained photosynthetic conversion of CO_2 to ethylene in recombinant cyanobacterium Synechocystis 6803. Energy Environ Sci 5:8998–9006
30. Liu X, Sheng J, Curtiss R III (2011) Fatty acid production in genetically modified cyanobacteria. Proc Natl Acad Sci U S A 108(17):6899–6904
31. Ruffing AM, Jones HD (2012) Physiological effects of free fatty acid production in genetically engineered Synechococcus elongatus PCC 7942. Biotechnol Bioeng 109 (9):2190–2199
32. Griffiths MJ, Harrison STL (2009) Lipid productivity as a key characteristic for choosing algal species for biodiesel production. J Appl Phycol 21(5):493–507
33. Hirokawa Y, Suzuki I, Hanai T (2015) Optimization of isopropanol production by engineered cyanobacteria with a synthetic metabolic pathway. J Biosci Bioeng 119(5):585–590
34. Möllers KB, Cannella D, Jørgensen H, Frigaard N-U (2014) Cyanobacterial biomass as carbohydrate and nutrient feedstock for bioethanol production by yeast fermentation. Biotechnol Biofuels 7(64)
35. Deng MD, Coleman JR (1999) Ethanol synthesis by genetic engineering in cyanobacteria. Appl Environ Microbiol 65(2):523–528
36. Markou G, Angelidaki I, Nerantzis E, Georgakakis D (2013) Bioethanol production by carbohydrate-enriched biomass of arthrospira (Spirulina) platensis. Energies 6:3937–3950
37. Fang M, Jin L, Zhang C, Tan Y, Jiang P, Ge N, Heping L, Xing X (2013) Rapid mutation of Spirulina platensis by a new mutagenesis system of atmospheric and room temperature plasmas (ARTP) and generation of a mutant library with diverse phenotypes. PLoS One 8 (10), e77046
38. Baebprasert W, Jantaro S, Khetkorn W, Lindblad P, Incharoensakdi A (2011) Increased H2 production in the cyanobacterium Synechocystis sp. strain PCC6803 by redirecting the electron. Metab Eng 13(5):610–616
39. McNeely K, Xu Y, Bennette N, Bryant DA, Dismukes GC (2010) Redirecting reductant flux into hydrogen production via metabolic engineering of fermentative carbon metabolism in a cyanobacterium. Appl Environ Microbiol 76(15):5032–5038
40. Ducat DC, Way JC, Silveremail PA (2011) Engineering cyanobacteria to generate high-value products. Trends Biotechnol 29(2):95–103
41. Min H, Sherman LA (2010) Hydrogen production by the unicellular, diazotrophic cyanobacterium Cyanothece sp. strain ATCC 51142 under conditions of continuous light. Appl Environ Microbiol 76(13):4293–4301
42. Ananyev G, Carrieri D, Dismukes GC (2008) Optimization of metabolic capacity and flux through environmental cues to maximize hydrogen production by the cyanobacterium "Arthrospira (Spirulina) maxima". Appl Environ Microbiol 74(19):6102–6113
43. Aoyama KUI, Miyake J, Asada Y (1997) Fermentative metabolism to produce hydrogen gas and organic compounds in a cyanobacterium, Spirulina platensis. J Ferment Bioeng 83:17–20
44. Jeffries TW, Timourien H, Ward RL (1978) Hydrogen production by Anabaena cylindrica: effect of varying ammonium and ferric ions, pH and light. Appl Environ Microbiol 35:704–710
45. Varel VH, Chen TH, Hashimoto AG (1988) Thermophilic and mesophilic methane production from anaerobic degradation of the cyanobacterium Spirulina maxima. Resour Conserv Recycl 1(1):19–26
46. Samson R, LeDuyt A (1986) Detailed study of anaerobic digestion of Spirulina maxima algal biomass. Biotechnol Bioeng 28(7):1014–1023

47. Mussgnug JH, Klassen V, Schlüter A, Kruse O (2010) Microalgae as substrates for fermentative biogas production in a combined biorefinery concept. J Biotechnol 150:51–56
48. Zou Y, Pisciotta J, Billmyre RB, Baskakov IV (2009) Photosynthetic microbial fuel cells with positive light response. Biotechnol Bioeng 104(5):939–946
49. Tsujimura S, Wadano A, Kano K, Iked T (2001) Photosynthetic bioelectrochemical cell utilizing cyanobacteria and water-generating oxidase. Enzyme Microb Technol 29(4-5):225–231
50. Tanaka K, Tamamushi R, Ogawa T (1985) Bioelectrochemical fuel-cells operated by the cyanobacterium, Anabaena variabilis. J Chem Technol Biotechnol 35(3):191–197
51. Zhao J, Li X-F, Ren Y-P, Wang X-H, Jian C (2012) Electricity generation from Taihu Lake cyanobacteria by sediment microbial fuel cells. J Chem Technol Biotechnol 87 (11):1567–1573
52. Cheng D, He Q (2014) Assessment of environmental stresses for enhanced microalgal biofuel production–an overview. Front Energy Res 2:1–8
53. Markou G, Angelidaki I, Georgakakis D (2012) Microalgal carbohydrates: an overview of the factors influencing carbohydrates production, and of main bioconversion technologies for production of biofuels. Appl Microbiol Biotechnol 96:631–645
54. Aikawa S, Joseph A, Yamad R, Izumi Y, Yamagishi T, Matsuda F, Kawai H, Chang J-S, Hasunuma T, Kondo A (2013) Direct conversion of Spirulina to ethanol without pretreatment or enzymatic hydrolysis processes. Energy Environ Sci 6:1844–1849
55. Markou G, Nerantzis E (2013) Microalgae for high-value compounds and biofuels production: a review with focus on cultivation under stress conditions. Biotechnol Adv 31:1532–1542
56. Rosgaard L, Porcellinis AJ, Jacobsen JH, Frigaard N-U, Sakuragi Y (2012) Bioengineering of carbon fixation, biofuels, and biochemicals in cyanobacteria and plants. J Biotechnol 162 (1):134–147
57. Matsunaga T, Takeyama H (1995) Genetic engineering in marine cyanobacteria. J Appl Phycol 7(1):77–84
58. Koksharova O, Wolk C (2002) Genetic tools for cyanobacteria. Appl Microbiol Biotechnol 58(2):123–137
59. Vioque A (2007) Transformation of cyanobacteria. Adv Exp Med Biol 616:12–22
60. Xu Y, Alvey RM, Byrne PO, Graham JE, Shen G, Bryant DA (2011) Expression of genes in cyanobacteria: adaptation of endogenous plasmids as platforms for high-level gene expression in Synechococcus sp. PCC 7002. Methods Mol Biol 684:273–293
61. Kawamura M, Sakakibara M, Watanabe T, Kita K, Hiraoka N, Obayashi A, Takagi M, Yano K (1986) A new restriction endonuclease from Spirulina platensis. Nucleic Acids Res 14 (5):1985–1989
62. Singh DP, Singh N (1997) Isolation and characterization of a metronidazole tolerant mutant of the cyanobacterium Spirulina platensis exhibiting multiple stress tolerance. World J Microbiol Biotechnol 13(2):179–183
63. Lee JW, Na D, Park JM, Lee J, Choi S, Lee SY (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nat Chem Biol 8(6):536–546
64. Nogales J, Gudmundsson S, Thiele I (2013) Toward systems metabolic engineering in cyanobacteria: opportunities and bottlenecks. Bioengineered 4(3):158–163
65. Wang B, Wang J, Zhang W, Meldrum DR (2012) Application of synthetic biology in cyanobacteria and algae. Front Microbiol 3:344
66. Berla BM, Saha R, Immethun CM, Maranas CD, Moon TS, Pakrasi HB (2013) Synthetic biology of cyanobacteria: unique challenges and opportunities. Front Microbiol 4:246
67. Ramey CJ, Barón-Sola Á, Aucoin HR, Boyle NR (2015) Genome engineering in cyanobacteria: where we are and where we need to go. ACS Synth Biol 4(11):1186–1196
68. Calteau A, Fewer DP, Latifi A, Coursin T, Laurent T, Jokela J, Kerfeld CA, Sivonen K, Piel J, Gugger M (2014) Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. BMC Genomics 15:977

69. Stanley DN, Raines CA, Kerfeld CA (2013) Comparative analysis of 126 cyanobacterial genomes reveals evidence of functional diversity among homologs of the redox-regulated CP12 protein. Plant Physiol 161(2):824–835
70. Yoshikawa K, Hirasawa T, Ogawa K, Hidaka Y, Nakajima T, Furusawa C, Shimizu H (2013) Integrated transcriptomic and metabolomic analysis of the central metabolism of Synechocystis sp. PCC 6803 under different trophic conditions. Biotechnol J 8(5):571–580
71. Kopf M, Klähn S, Pade N, Weingärtner C, Hagemann M, Voß B, Hess WR (2014) Comparative genome analysis of the closely related Synechocystis strains PCC 6714 and PCC 6803. DNA Res 21(3):255–266
72. Dienst D, Georg J, Abts T, Jakorew L, Kuchmina E, Börner T, Wilde A, Dühring U, Enke H, Hess WR (2014) Transcriptomic response to prolonged ethanol production in the cyanobacterium Synechocystis sp. PCC6803. Biotechnol Biofuels 7:21
73. Dunlop MJ (2011) Engineering microbes for tolerance to next-generation biofuels. Biotechnol Biofuels 43:2
74. Zingaro KA, Papoutsakis ET (2012) Toward a semisynthetic stress response system to engineer microbial solvent tolerance. MBio 3(5):e00308–e00312
75. Qiao J, Wang J, Chen L, Tian X, Huang S, Ren X (2012) Quantitative iTRAQ LC-MS/MS proteomics reveals metabolic responses to biofuel ethanol in cyanobacterial Synechocystis sp. PCC 6803. J Proteome Res 11:5286–5300
76. Tian X, Chen L, Wang J, Qiao J, Zhang W (2013) Quantitative proteomics reveals dynamic responses of Synechocystis sp. PCC 6803 to next-generation biofuel butanol. J Proteomics 78:326–345
77. Liu J, Chen L, Wang J, Qiao J, Zhang W (2012) Proteomic analysis reveals resistance mechanism against biofuel hexane in Synechocystis sp. PCC 6803. Biotechnol Biofuels 5:68
78. Qiao J, Huang S, Te R, Wang J, Chen L, Zhang W (2013) Integrated proteomic and transcriptomic analysis reveals novel genes and regulatory mech- anisms involved in salt stress responses in Synechocystis sp. PCC 6803. Appl Microbiol Biotechnol 97:8253–8264
79. Huang S, Chen L, Te R, Qiao J, Wang J, Zhang W (2013) Complementary iTRAQ proteomics and RNA-seq transcriptomics reveal multiple levels of regulation in response to nitrogen starvation in Synechocystis sp. PCC 6803. Mol Biosyst 9:2565–2574
80. Pei G, Chen L, Wang J, Qiao J, Zhang W (2014) Protein network signatures associated with exogenous biofuels treatments in cyanobacterium Synechocystis sp. PCC 6803. Front Bioeng Biotechnol 2:48
81. Singh AK, Elvitigala T, Cameron JC, Ghosh BK, Bhattacharyya-Pakrasi M, Pakrasi HB (2010) Integrative analysis of large scale expression profiles reveals core transcriptional response and coordination between multiple cellular processes in a cyanobacterium. BMC Syst Biol 4:105
82. Sánchez BJ, Nielsen J (2015) Genome scale models of yeast: towards standardized evaluation and consistent omic integration. Integr Biol 7:846–858
83. Yen JY, Nazem-Bokaee H, Freedman BG, Athamneh AI, Senger RS (2013) Deriving metabolic engineering strategies from genome-scale modeling with flux ratio constraints. Biotechnol J 8(5):581–594
84. McCloskey D, Palsson BO, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol 9:661
85. Yang C, Hua Q, Shimizu K (2002) Metabolic flux analysis in Synechocystis using isotope distribution from 13C-labeled glucose. Metab Eng 4(3):202–216
86. Shastri AA, Morgan JA (2005) Flux balance analysis of photoautotrophic metabolism. Biotechnol Prog 21(6):1617–1626
87. Hong S-J, Lee C-G (2007) Evaluation of central metabolism based on a genomic database of Synechocystis PCC6803. Biotechnol Bioprocess Eng 12(2):165–173
88. Navarro E, Montagud A, Fernández de Córdoba P, Urchueguía JF (2009) Metabolic flux analysis of the hydrogen production potential in Synechocystis sp. PCC6803. Int J Hydrogen Energy 34(21):8828–8838

89. Fu P (2009) Genome-scale modeling of Synechocystis sp. PCC 6803 and prediction of pathway insertion. J Chem Technol Biotechnol 84(473483)
90. Knoop H, Zilliges Y, Lockau W, Steuer R (2010) The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. Plant Physiol 154(1):410–422
91. Montagud A, Navarro E, Fernandez de Cordoba P, Urchueguia JF, Patil KR (2010) Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. BMC Syst Biol 4:156
92. Montagud A, Zelezniak A, Navarro E, de Cordoba PF, Urchueguia JF, Patil KR (2011) Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium Synechocystis sp. PCC6803. Biotechnol J 6(3):330–342
93. Yoshikawa K, Kojima Y, Nakajima T, Furusawa C, Hirasawa T, Shimizu H (2011) Reconstruction and verification of a genome-scale metabolic model for Synechocystis sp. PCC6803. Appl Microbiol Biotechnol 92(2):347–358
94. Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. Proc Natl Acad Sci U S A 109(7):2678–2683
95. Saha R, Verseput AT, Berla BM, Mueller TJ, Pakrasi HB, Maranas CD (2012) Reconstruction and comparison of the metabolic potential of cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803. PLoS One 7(10), e48285
96. Knoop H, Grundel M, Zilliges Y, Lehmann R, Hoffmann S, Lockau W, Steuer R (2013) Flux balance analysis of cyanobacterial metabolism: the metabolic network of Synechocystis sp. PCC 6803. PLoS Comput Biol 9(6), e1003081
97. Knoop H, Steuer R (2015) A computational analysis of stoichiometric constraints and trade-offs in cyanobacterial biofuel production. Front Bioeng Biotechnol 3:47
98. Cogne G, Gros JB, Dussap CG (2003) Identification of a metabolic network structure representative of Arthrospira (spirulina) platensis metabolism. Biotechnol Bioeng 84 (6):667–676
99. Meechai A, Pongakarakun S, Deshnium P, Cheevadhanarak S, Bhumiratana S (2004) Metabolic flux distribution for γ-linolenic acid synthetic pathways in Spirulina platensis. Biotechnol Bioprocess Eng 9:506–513
100. Klanchui A, Khannapho C, Phodee A, Cheevadhanarak S, Meechai A (2012) iAK692: a genome-scale metabolic model of Spirulina platensis C1. BMC Syst Biol 6:71
101. Hamilton JJ, Reed JL (2012) Identification of functional differences in metabolic networks using comparative genomics and constraint-based models. PLoS One 7(4)
102. Vu TT, Hill EA, Kucek LA, Konopka AE, Beliaev AS, Reed JL (2013) Computational evaluation of Synechococcus sp. PCC 7002 metabolism for chemical production. Biotechnol J 8(5):619–630
103. Song HS, McClure RS, Bernstein HC, Overall CC, Hill EA, Beliaev AS (2015) Integrated in silico analyses of regulatory and metabolic networks of Synechococcus sp. PCC 7002 reveal relationships between gene centrality and essentiality. Life (Basel) 27(5):1127–1140
104. Vu TT, Stolyar SM, Pinchuk GE, Hill EA, Kucek LA, Brown RN, Lipton MS, Osterman A, Fredrickson JK, Konopka AE, Beliaev AS, Reed JL (2012) Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium Cyanothece sp. ATCC 51142. PLoS Comput Biol 8(4):e1002460
105. Mueller TJ, Berla BM, Pakrasi HB, Maranas CD (2013) Rapid construction of metabolic models for a family of Cyanobacteria using a multiple source annotation workflow. BMC Syst Biol 7(142)
106. Triana J, Montagud A, Siurana M, Fuente D, Urchueguía A, Gamermann D, Torres J, Tena J, de Córdoba PF, Urchueguía JF (2014) Generation and evaluation of a genome-scale metabolic network model of Synechococcus elongatus PCC7942. Metabolites 4(3):680–698
107. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5(1):93–121

108. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 42(Database issue):D459–D471
109. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32(Database issue):D277–D280
110. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39(Database issue):D691–D697
111. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43 (Database issue):D204–D212
112. Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, Sato S, Tabata S, Kaneko T, Nakamura Y (2010) CyanoBase: the cyanobacteria genome database update 2010. Nucleic Acids Res 38(Database issue):D379–D381
113. Sasaki NV, Sato N (2010) CyanoClust: comparative genome resources of cyanobacteria and plastids. Database 2010
114. Peter AP, Lakshmanan K, Mohandass S, Varadharaj S, Thilagar S, Abdul Kareem KA, Dharmar P, Gopalakrishnan S, Lakshmanan U (2015) Cyanobacterial KnowledgeBase (CKB), a compendium of cyanobacterial genomes and proteomes. PLoS One 10(8), e0136262
115. Senachak J, Cheevadhanarak S, Hongsthong A (2015) SpirPro: a Spirulina proteome database and web-based tools for the analysis of protein-protein interactions at the metabolic level in Spirulina (Arthrospira) platensis C1. BMC Bioinformatics 16:233
116. Arun PVPS, Bakku RK, Subhashini M, Singh P, Prabhu NP, Suzuki I, Prakash JSS (2012) CyanoPhyChe: a database for physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins. PLoS One 7(11), e49425
117. Yang Y, Feng J, Li T, Ge F, Zhao J (2015) CyanOmics: an integrated database of omics for the model cyanobacterium Synechococcus sp. PCC 7002. Database 2015:1–9
118. Hernandez-Prieto MA, Futschik ME (2012) CyanoEXpress: a web database for exploration and visualisation of the integrated transcriptome of cyanobacterium Synechocystis sp. PCC6803. Bioinformation 8(13):634–638
119. Kelly L, Huang KH, Ding H, Chisholm SW (2012) ProPortal: a resource for integrated systems biology of Prochlorococcus and its phage. Nucleic Acids Res 40(Database issue): D632–D640
120. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
121. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) The Pfam protein families database. Nucleic Acids Res 42:D222–D230
122. Lu Z (2010) PubMed and beyond: a survey of web tools for searching biomedical literature. Database 2011;2011:baq036
123. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11(1):40–79
124. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35(Web Server issue):W182–W185
125. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 42(Database issue):D206–D214

126. Swainston N, Smallbone K, Mendes P, Kell D, Paton N (2011) The SuBliMinaL toolbox: automating steps in the reconstruction of metabolic networks. J Integr Bioinform 8(2):186
127. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS Comput Biol 9(3)
128. Liao Y-C, Tsai M-H, Chen F-C, Hsiung CA (2012) GEMSiRV: a software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. Bioinformatics 28(13):1752–1758
129. Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC, Rocha M (2010) OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol 45(45)
130. Feng X, Xu Y, Chen Y, Tang YJ (2012) MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. BMC Syst Biol 6:94
131. Narang P, Khan S, Hemrom AJ, Lynn AM, Consortium OSDD (2014) MetaNET-a web-accessible interactive platform for biological metabolic network analysis. MC Syst Biol 8(130)
132. Kelley JJ, Lane A, Li X, Mutthoju B, Maor S, Egen D, Lun DS (2014) MOST: a software environment for constraint-based metabolic modeling and strain design. Bioinformatics 31:610–611
133. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BO (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6 (9):1290–1307
134. Baroukh C, Muñoz-Tamayo R, Steyer J-P, Bernard O (2014) DRUM: a new framework for metabolic modeling under non-balanced growth. Application to the carbon metabolism of unicellular microalgae. PLoS One 9 (8):e104499
135. Gomez JA, Höffner K, Barton PI (2014) DFBAlab: a fast and reliable MATLAB code for dynamic flux balance analysis. BMC Bioinformatics 2014(15):409
136. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143
137. Baart GJ, Martens DE (2012) Genome-scale metabolic models: reconstruction and analysis. Methods Mol Biol 799:107–126
138. Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. Proc Natl Acad Sci U S A 109:2678–2683
139. Montagud A, Zelezniak A, Navarro E, de Córdoba PE, Urchueguía JF, Patil KR (2011) Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium Synechocystis sp. PCC6803. Biotechnol J 6(3):330–342
140. Senger RS (2010) Biofuel production improvement with genome-scale models: the role of cell composition. Biotechnol J 5(7):671–685
141. Jeffrey DO, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28 (3):245–248
142. Mahadevan R, Schilling C (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276
143. Kramer DM, Evans JR (2011) The importance of energy balance in improving photosynthetic productivity. Plant Physiol 155(1):70–78
144. Chang RL, Ghamsari L, Manichaikul A, Hom EF, Balaji S, Fu W, Shen Y, Hao T, Palsson BO, Salehi-Ashtiani K, Papin JA (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. Mol Syst Biol 7:518
145. Baroukh C, Muñoz-Tamayo R, Steyer J-P, Bernard O (2015) A state of the art of metabolic networks of unicellular microalgae and cyanobacteria for biofuel production. Metab Eng 30:49–60

146. Subashchandrabose SR, Ramakrishnan B, Megharaj M, Venkateswarlu K, Naidu R (2013) Mixotrophic cyanobacteria and microalgae as distinctive biological agents for organic pollutant degradation. Environ Int 51:59–72
147. Moore LR (2013) More mixotrophy in the marine microbial mix. Proc Natl Acad Sci U S A 110(21):8323–8324
148. Baroukh C, Munoz-Tamayo R, Bernard O, Steyer JP (2015) Mathematical modeling of unicellular microalgae and cyanobacteria metabolism for biofuel production. Curr Opin Biotechnol 33:198–205
149. Mahadevan R, Edwards JS, Doyle FJ 3rd (2002) Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophys J 83(3):1331–1340
150. Erdrich P, Knoop H, Steuer R, Klamt S (2014) Cyanobacterial biofuels: new insights and strain design strategies revealed by computational modeling. Microb Cell Fact 13(1):128
151. Hädicke O, Klamt S (2010) CASOP: a computational approach for strain optimization aiming at high productivity. J Biotechnol 147(2):88–101
152. Hädicke O, Klamt S (2011) Computing complex metabolic intervention strategies using constrained minimal cut sets. Metab Eng 13(2):204–213

Adv Biochem Eng Biotechnol (2017) 160: 103–120
DOI: 10.1007/10_2016_45

Published online: 3 December 2016

Genome-Scale Modeling of Thermophilic Microorganisms

Sanjeev Dahal, Suresh Poudel, and R. Adam Thompson

Abstract Thermophilic microorganisms are of increasing interest for many industries as their enzymes and metabolisms are highly efficient at elevated temperatures. However, their metabolic processes are often largely different from their mesophilic counterparts. These differences can lead to metabolic engineering strategies that are doomed to fail. Genome-scale metabolic modeling is an effective and highly utilized way to investigate cellular phenotypes and to test metabolic engineering strategies. In this review we chronicle a number of thermophilic organisms that have recently been studied with genome-scale models. The microorganisms spread across archaea and bacteria domains, and their study gives insights that can be applied in a broader context than just the species they describe. We end with a perspective on the future development and applications of genome-scale models of thermophilic organisms.

Keywords Draft reconstruction, Flux balance analysis, Flux variability analysis, Gap-filling, Genome-scale models, Stoichiometric matrix

S. Dahal (✉) and S. Poudel
UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN, USA

Oak Ridge National Laboratory, Oak Ridge, TN, USA

BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: sdahal@vols.utk.edu

R.A. Thompson
BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN, USA

Contents

1 Introduction to Thermophilic Microorganisms 104
2 Uses of Thermophilic Microorganisms in Industry 105
3 Genome-Scale Modeling of Metabolism 105
4 Genome-Scale Modeling of Thermophilic Microorganisms 106
4.1 *Clostridium thermocellum* 106
4.2 *Thermotoga maritima* 107
4.3 *Thermus thermophilus* 109
4.4 *Sulfolobus solfataricus* 110
4.5 *Thermobifida fusca* 111
4.6 *Moorella thermoacetica* 112
4.7 *Streptococcus thermophilus* 114
4.8 *Thermoanaerobacterium saccharolyticum* 115
4.9 Models Deposited in BioModels Database 115
5 Conclusion 116
References 116

1 Introduction to Thermophilic Microorganisms

According to Brock, thermophiles are organisms that can grow and reproduce at high temperatures [1]. Generally, 50–60°C is regarded as the minimal temperature for bacteria and archaea to be considered thermophiles, because this is the known upper limit for eukaryotes. Even within thermophiles there is a distinction between hypothermophiles and hyperthermophiles based on their optimum temperature. Hypothermophiles prefer temperatures of up to 80°C whereas hyperthermophiles can have a temperature preference of up to 100°C.

Thermophilic microorganisms can be found in various habitats such as geothermal hot springs in places such as Yellowstone National Park where *Thermus aquaticus* was discovered [2]. Another major known habitat is the area around deep-sea vents from which *Methanococcus jannaschii* was found [3]. Nutritionally, thermophiles which span the metabolic range from phototrophy to chemotrophy, from autotrophy to heterotrophy, and from aerobic to anaerobic capabilities have been described in the literature [4].

Generally, studies have shown that, at optimal temperatures, thermophiles show lower growth yield compared to their mesophilic counterparts at their respective optimal temperatures. The lower yield is attributed to a higher energy for maintenance requirements such as turnover of proteins and nucleotides, cell mobility, and ionic maintenance. Because of their high temperature growth conditions, thermophiles require more energy for maintaining these conditions [5]. There are some exceptions to the rule, such as *Thermothrix thiopara* [6]. The observed reduced growth efficiency and higher maintenance requirements have made these organisms interesting in research as these organisms tend to produce various catabolic products in larger quantities than other organisms [7].

2 Uses of Thermophilic Microorganisms in Industry

Thermophilic organisms have been utilized in several industrial areas such as the fuel industry, waste management, and mining. Thermophiles and their enzymes have been widely regarded as the most efficient way to generate biofuels from lignocellulose (contains cellulose, hemicellulose, and lignin). The use of thermostable organisms and enzymes provides several advantages such as faster conversion of substrates, decreased risk of contaminations, and more compound recovery. Several thermophile-produced enzymes have been proposed for degradation of cellulosic biomass. Some of the enzymes are cellulases (which degrade cellulose) and xylanases (which degrade hemicellulose). Furthermore, thermophilic organisms have been suggested to be the microbial cell factory for consolidated bioprocessing (CBP) in which degradation of lignocellulose and fermentation of sugars are accomplished in one step. Examples of these organisms are *Clostridium thermocellum*, *Caldicellulosiruptor saccharolyticus*, and *Caldicellulosiruptor bescii*.

In addition to biofuels, thermophilic organisms also find their use in the area of waste management [8–11]. Studies have shown that the use of both mesophilic and thermophilic digesters could help recover energy from biowastes such as livestock manure and food waste. Furthermore, the use of thermophiles for the recovery of metals from industrial and municipal wastes has also been proposed [12, 13]. Bioleaching is the process through which microorganisms are used to extract metals from ores and waste products. This process has been used for the extraction of metals such as zinc, copper, gold, and molybdenum using organisms such as *Metallosphaera sedula* [14, 15]*, Sulfolobus* [14, 15]*, Sulfobacillus* [16], or *Ferroplasma* [17].

3 Genome-Scale Modeling of Metabolism

Genome-scale modeling is a powerful tool that has been used for many applications, such as the prediction of cellular phenotypes, elucidation of biological principles, rational strain design for metabolic engineering, simulation of co-cultures, and the interpretation of OMICs and other high-throughput datasets [18]. The most common method for analyzing a genome-scale metabolic network is called Flux Balance Analysis (FBA). In general, a metabolic network can be represented by a stoichiometric matrix $\boldsymbol{S} \in \mathbb{R}^{m \times n}$, consisting of m metabolites and n reactions, such that the entry $s_{i,j}$ is the stoichiometric coefficient of metabolite i in reaction j. A valid flux distribution vector $\boldsymbol{v} \in \mathbb{R}^{n \times 1}$ satisfies a steady-state condition

$$\boldsymbol{S} \cdot \boldsymbol{v} = 0$$

and is thus constrained by mass balance. The flux distribution vector is also constrained by thermodynamics such that

$$v_j \geq 0$$

for all irreversible reactions j. FBA relies on the stoichiometric and thermodynamic constraints to optimize a cellular objective, such as maximizing cell growth, maximizing product synthesis, or minimizing ATP hydrolysis [19].

Using this framework, optimization problems can be coupled in a multitude of ways to probe cellular metabolism, and multiple software packages have been developed to facilitate the construction and analysis of genome-scale models [20, 21]. In addition, many curated models of thermophiles and non-thermophiles have been deposited in the BiGG database [22]. As the metabolism of thermophiles are typically less well-understood than that of their mesophilic counterparts, and the challenges of living at higher temperatures favors alternative metabolic pathways, genome-scale models are effective tools to study thermophiles. The following section outlines several curated genome-scale models and how they have been used to study thermophilic metabolism, in particular increasing the understanding of deviations from model organisms and generating hypotheses for further study.

4 Genome-Scale Modeling of Thermophilic Microorganisms

4.1 Clostridium thermocellum

C. thermocellum is a gram-positive bacterium of great interest for consolidated bioprocessing of lignocellulose to biofuels because it exhibits one of the fastest growth rates on crystalline cellulose which it directly converts to the biofuels such as ethanol, hydrogen, and isobutanol.

The first genome-scale model of *C. thermocellum* was created for strain ATCC 27405 by Roberts et al. [23]. The model, called *i*SR432, consists of 577 reactions, 525 metabolites, and 432 genes. The model consists of the cellulosome data contained in its proteomic information. The cellulosome is a large extracellular protein complex, which is optimized for hydrolyzing cellulose into glucose oligomers of length 2–6. The draft reconstruction was based on the genome annotations from databases such as IMG, UniProt, and KEGG [24]. Additional transport reactions were added based on a reciprocal BLAST hit between *C. thermocellum* genome and the Transport Classification Database (TCDB) [25]. In the draft reconstruction, the investigators discovered that there were missing gaps, especially because of species-specific metabolism such as cellulosome production, cellulose and chitin degradation, biosynthesis of teichoic acid and peptidoglycan, steroid metabolism, and transport reactions. Therefore a manual gap-filling was carried out on the reconstruction to fill additional gaps using literature and experimental data. Furthermore, several other gaps were resolved using reciprocal blast hit between all the genes containing the missing Enzyme Commission (EC) number and the *C. thermocellum* ATCC 27405 genome. The process was iteratively performed until a positive flux for biomass synthesis was observed. The model was tested against data for growth on minimal media containing either cellobiose or fructose in continuous or batch cultures and was able to reproduce some phenotypes, although most fermentation product flux profiles were inaccurate. However, the addition of RNA-Seq data allowed for better predictions [26].

Following the construction of the model *i*SR432, much has been learned about the metabolism of *C. thermocellum*, particularly dealing with its atypical central carbon metabolism and redox processes [27, 28]. With these updates in mind, Thompson et al. have constructed and curated a new genome-scale model of *C. thermocellum*, this time of the genetically tractable strain DSM 1313 [29–31]. This new model of *C. thermocellum* DSM 1313 also incorporates a more dynamic cellulosome component, which allowed the researchers to predict more accurately the growth on soluble versus insoluble substrates. This is a key distinction because different substrates lead to different fermentation profiles and energetic requirements for growth. Using this updated model, the authors delved into the changes in metabolism between various substrates to propose a regulatory mechanism that explains the difference. The authors also used a strain design algorithm for optimal production of ethanol, hydrogen, and isobutanol, paving the way for future metabolic engineering [31]

4.2 Thermotoga maritima

T. maritima is a hyperthermophilic anaerobic bacterium believed to be one of the most ancient of eubacteria [32]. Its metabolism is classified as chemoorganotrophic, catabolizing sugars to produce CO_2, acetate, lactate, and hydrogen [33]. For a free-living organism, it has one of the smallest genomes [34]. Zhang et al. created the first metabolic reconstruction of *T. maritima* [35]. This model integrated structural information to examine the evolution of protein folds in the metabolism, and consists of 478 genes, 503 metabolites, and 645 reactions. The model was able to reproduce experimental results for growth and secretion profiles on different substrates. The protein information was gathered through literature data first followed by homology-based annotation databases. Finally, FBA and gap-filling were iteratively carried out until the model was able to replicate experimental growth results.

One important conclusion from the integration of structural data was the strengthening of the ‘patchwork’ hypothesis, which states that gene duplication events result in proteins that evolve to function in a similar manner to each other but in different pathways [36]. Furthermore, this study discovered that specific folds dominate the proteins involved in central metabolism, which suggests divergent evolution of ancient proteins. The core essential proteins, however, have relatively diverse folds because they catalyze highly specific reactions, which require particular enzymes.

Nogales et al. expanded the model created by Zhang et al. to study hydrogen production in *T. maritima* [37]. As *T. maritima* is a hyperthermophile and produces large amounts of hydrogen from various complex sugar polymers, it is an ideal candidate for microbial hydrogen bioproduction. The model expansion consists of modifying reactions involving hydrogen production and ferredoxin based on recent findings in *T. maritima*. Similarly, reactions are added for secretion of certain metabolites and in the Entner–Doudoroff (ED) pathway to improve the model. The predictive capability of the new reconstruction was found to be better than that of the original as determined by its ability to replicate more experimental results in silico. According to the model and experimental data, it was confirmed that *T. maritima* grows faster on polysaccharides than on other carbon sources. Moreover, it was determined that this organism mainly uses Embden–Meyerhoff (EM) and ED rather than the oxidative branch of the pentose phosphate pathway (OPP) to catabolize sugars for growth-coupled production of hydrogen. Furthermore, it was demonstrated that acetate production results in improved growth and hydrogen yield. The authors also concluded that sulfur is one of the important electron sinks in this organism based on model prediction and literature information.

The model was further used for developing an understanding of the redox balancing in the organism. The analysis suggested that carbon metabolism leads to surplus NADH, which is consumed in ED pathway coupled with sulfur reduction and subsequently causes the stoichiometric ratio of ferredoxin to NADH to be less than two. This result suggests the role of ED pathway in internal redox balancing. The investigators also used in silico knockouts to determine the mutants that have improved hydrogen yield on various substrates and limited sulfur condition. The simulations suggested that double mutants of acetate thiokinase (ACKr, converts acetyl phosphate to acetate) and L-lactate dehydrogenase (LDH_L, converts pyruvate to lactate) have improved hydrogen production. Mutation in either triose phosphate isomerase (TPI) or fructose bisphosphate aldolase (FBPA) also led to an increase in hydrogen yield when grown on glucose but at the expense of growth rate. Furthermore, it was determined that the addition of an efficient NADP+ regenerating enzyme could drive glucose metabolism through the OPP pathway and lead to more hydrogen production. The knock-in in combination with certain double knockouts resulted in a very high hydrogen yield. An additional confirmation from in silico simulation was that the introduction of NADP+ regenerating enzyme enabled the organism to grow on glycerol, on which the wild type cannot sustain growth. This knock-in coupled with triple mutations in FBPA, 2-dehydro-3-deoxy-phosphogluconate aldolase (EDA_R) and TPI led to improved hydrogen production of up to 5 mol per mole of glycerol [37].

4.3 Thermus thermophilus

T. thermophilus is a gram-negative organism that grows aerobically and anaerobically with the help of exogenous electron acceptors such as nitrate. It can consume a wide variety of protein substrates and carbohydrates, which is facilitated by a suite of proteases, glucosidases, and lipases to allow ideal growth between the temperatures of 65 and 72°C. Lee et al. created the first genome-scale model of this organism, which incorporates several distinct features present in thermophiles and is specific to *T. thermophilus* [38]. The reconstruction, called *i*TT548, contains 548 genes, 796 reactions, and 635 metabolites. The draft metabolic reconstruction was carried out in the usual manner, but the gap-filling required organism-specific information or knowledge from related organisms. Specifically, the pathways for carotenoid synthesis and those relating to the growth on various substrates were incorporated. One of the most distinct features of this model was that the biomass composition was determined using experimental results and literature information exclusive to thermophilic organisms.

When *i*TT548 was compared with *i*AF1260, the then most recent genome-scale model of *Escherichia coli*, it was determined that the amino acid biosynthesis pathways for lysine and methionine were different in these two organisms. Similarly, the model also clearly demonstrated that the carotenoid and polyamine biosynthesis were some of the unique characteristics of this organism compared to other gram-negative bacteria. These two groups of metabolites enable this organism to withstand high temperatures [39–42]. Furthermore, the model also predicted that the organism utilized amino acids to synthesize branched chain fatty acid. When constraint-based flux analysis was carried out on minimal and rich glucose media, the simulation results were consistent with experimental data for growth rate. The simulation demonstrated that certain amino acids were consumed at a higher rate than others to synthesize fatty acids. This pattern was distinct compared to *E. coli* because *T. thermophilus* consumed nutrients to drive fluxes more toward fatty acid synthesis rather than toward energy production.

Finally, gene essentiality was determined in this organism. It was demonstrated that genes involved in carotenoid biosynthesis were the most essential genes in both rich and minimal media. Most of the amino acid metabolism genes were found to be essential in minimal media except for genes involved in biosynthesis of tryptophan, proline, and tyrosine. Moreover, the model also showed that genes involved in oxidative phosphorylation and the citric acid cycle were also essential because *T. thermophilus* is an obligate aerobe. Finally, the simulations showed that *T. thermophilus* has a higher proportion of essential genes compared to *E. coli*. It was concluded that the more rigid network of *Thermus thermophilus* helps it to survive in higher temperatures.

4.4 Sulfolobus solfataricus

S. solfataricus is a hyperthermoacidophilic organism within the phylum *Crenarchaeota* found in volcanic hot springs [43]. It is strictly aerobic and can grow either autotrophically or heterotrophically. It also has the ability to oxidize sulfur. Moreover, its metabolism is very diverse, containing a bicarbonate fixation pathway, so it can grow chemolithoautotrophically and has the ability to grow on phenol [44, 45].

Ulas et al. created the first ever genome-scale model of *S. solfataricus* [43]. It was created by compiling the information from annotations created by EnzymeDetector software [46], various databases such as KEGG [47], MetaCyc [48], BRENDA [49], *Sulfolobus*-specific literature, and experimental data. Following this initial step, manual gap-filling was carried out and resulted in a genome-scale model *i*TU515 which contains 831 reactions involving 705 metabolites. The model was able to depict accurately the broad range of metabolism of *S. solfataricus*.

Following the reconstruction, the predictive capabilities of the model were determined. The organism was grown in silico on various carbon sources and FBA was carried out to compare with experimental results. *S. solfataricus* has a very low growth rate and utilizes only 25% of carbon toward biomass production, which the model was able to predict accurately. *S. sofataricus* has a modified ED pathway such that the glucose flux can go toward either semi-phosphorylative or non-phosphorylative branch. The model demonstrated that the carbon flux of glucose was divided in a ratio of 1:4 between the semi-phosphorylative/reverse ribulose monophosphate pathway and the non-phosphorylative/TCA cycle. In another analysis using this model, the effect of exopolysaccharide (EPS) on the growth of the organism was investigated. *S. solfataricus* produces EPS using the imported carbon flux. The model clearly demonstrated that when EPS-producing reactions were added, the growth rate decreased when grown on glucose media. Therefore, according to the model, the production of EPS causes lower biomass flux.

When flux variability analysis (FVA) was carried out on the model, it was determined that for an optimal flux toward biomass, 79 reactions showed variability. Most of the significant flux variation appeared because of the semi-phosphorylative and non-phosphorylative branches of the ED pathway. Similarly, FVA analysis for suboptimal analysis (in which up to 95% of optimal growth is considered) caused the number of reactions to increase from 79 to 352 that showed flux variability.

The ability of the organism to grow chemolithotrophically using the hydroxypropionate-hydroxybutyrate cycle under aerobic conditions was demonstrated on a bicarbonate source. FBA showed that the ED pathway was inactive and sulfur metabolism and the hydroxypropionate-hydroxybutyrate cycle was active when grown on bicarbonate. Furthermore, the model predicted the growth rate to be higher than that on glucose. In a similar analysis, growth on phenol was determined.

The FBA demonstrated some significant differences between growth on glucose and on phenol such as active phenol uptake and degradation and inactive ED pathways. Biomass production on phenol was determined to be one of the lowest possibly because of the requirement of ATP for the production of pyruvate from phenol.

Additionally, the model was used for the analysis of growth on various other carbon sources. Using FBA and normalization of carbon uptake rate for each carbon source to 1 mmol of carbon atoms per gram dry weight per hour, the model predicted growth on 35 different substrates. Around 13 of the substrates produced more biomass than on glucose. It was determined that carbon sources entering the central pathway through the TCA cycle resulted in lower biomass production, except 2-oxoglutarate which showed a higher yield. Glycerol was determined to be the source of highest biomass production because metabolism of glycerol produced twice the amount of ATP per six carbon atoms than other substrates.

Finally, a gene essentiality analysis on glucose media was carried out on this model. Around 18% of genes were determined to be essential because the in silico deletion of these genes resulted in biomass production of less than 2% of the original. The genes involved in the central metabolism such as the reverse ribulose-monophosphate pathway and gluconeogenesis were determined to be essential. The model predicted only some genes in ED pathway and TCA cycle to be crucial.

4.5 Thermobifida fusca

T. fusca is an aerobic, gram-positive bacterium of the *Actinomycetes* phylum [50]. Because of its stability at high pH and temperature, and possession of an efficient cellulolytic system consisting of several endo- and exocellulases, *T. fusca* could be useful in consolidated bioprocessing of lignocellulose for biofuel production. Deng and Fong demonstrated that *T. fusca* could be manipulated and optimized to produce propanol from untreated biomass [51]. Vanee et al. created three different genome-scale models of *T. fusca* through (1) an automated approach using Model SEED, (2) a semi-automated approach using KEGG, and (3) a proteomics-based model using proteome data for cells grown on cellobiose [52]. The Model SEED (Tfu_v1)-based reconstruction contains 1,302 reactions, but cannot predict growth on cellobiose. Similarly, the semi-automated approach (Tfu_v2) produced a model with 1,002 reactions, but it also could not predict the growth on cellobiose accurately. The proteomics-based genome-scale model (*i*Tfu296) consists of 975 reactions with 296 genes. The simulation with *i*Tfu296 predicted a growth rate similar to that experimentally derived. Between Tfu_v2 and *i*Tfu296, vast differences were observed in functions, and the study suggested that *i*Tfu296 was much closer to the in vivo phenotype. During growth on cellobiose, the *i*Tfu296 predicted 110 active reactions among which the majority of reactions were involved in carbohydrate and amino acid metabolism (Table 1).

The investigators were interested in terpenoid biosynthesis in *T. fusca*, and hence added 16 reactions to compute the feasibility of flux through the terpenoid backbone pathway. The objective function used for FBA was biomass production. Except for one reaction, all reactions in the mevalonate pathway were found to be active, and provided investigators with a hypothesis to test the presence of this pathway experimentally. The experimental results, in contrast, demonstrated that the non-mevalonate pathway is present in *T. fusca*. This further emphasizes the importance of genome-scale models because, with the help of the model, the researchers were able to investigate quickly the terpenoid biosynthesis pathway in *T. fusca*.

4.6 Moorella thermoacetica

M. thermoacetica is a strict anaerobe that can use both electron transport phosphorylation and substrate level phosphorylation to produce energy. It has the ability to convert substrates such as carbon dioxide, glucose, or fructose into acetate and produce ATP [53]. It primarily utilizes the Wood–Ljungdahl (WLD) pathway to produce acetate from CO_2 and hydrogen, and as such it has for decades been widely studied as a model organism in acetogenesis. A genome-scale reconstruction of *M. thermoacetica* was created by Islam et al. and was called *i*AI558. The model contains 558 genes and 705 reactions [53]. The highest number of active reactions was determined to be involved in the cofactor metabolism subsystem.

*i*AI558 was used for simulation of growth on various substrates such as H_2, CO_2, CO, and methanol for autotrophic growth and glucose, fructose, and xylose for heterotrophic growth. Additionally, ATP production and yield were also computed. The growth simulation was compared to experimental data. The model accurately predicted growth rates on H_2-CO_2 (syngas) and CO. The growth rate and yield for CO was determined to be the highest among autotrophic substrates. Growth on heterotrophic substrates produced higher yield, growth rate, and ATP production than that on autotrophic substrates.

Simulation of growth on syngas and glucose were also compared. The study demonstrated that for syngas, *M. thermoacetica* mainly used WLD and gluconeogenesis and conserved energy through electron transport phosphorylation (ETP) or anaerobic respiration. During growth on glucose, glycolysis was the most dominant process, but the WLD pathway was also highly active. The energy conservation/ production was carried out by substrate level phosphorylation in glycolytic reactions. Because substrate level phosphorylation is more efficient in ATP generation [53], heterotrophic growth produces higher ATP yield, which was corroborated by the model simulation.

The model was further used for study of ATP generation during autotrophic growth. Reactions were added to the model based on hypotheses proposed by

Table 1 **Salient features of some of the important genome scale models described in the text**

Features	iSR432[a]	*T. saccharolyticum*	iTZ479[a]	iTfu296[b]	iAI558	iTU515
Total reactions	631	537	645	975	705	718
Exchange	54	22[c]	83	30[c]	60	58
Gene-associated reactions	463	461	518	NA	620	606
Genes	432	315	479	296	558	515
Unique metabolites	525	502[c]	503	734[c]	630[c]	705
Notable complex carbon substrates	Cellobiose	NA	Starch, xylan, cellulose	Cellobiose	NA	Cellulose, starch, xylan
Dominant pathways	Amino acid metabolism	Nucleotide metabolism	Amino acid metabolism Nucleotide metabolism	Amino acid metabolism	Cofactor metabolism	Nucleotide metabolism

[a] Only first model used

[b] Proteomics model used

[c] Determined by parsing the model

previous studies. The first mechanism proposed by Mock et al. was the production of ATP through formate-hydrogen lyase (FHL) and methylene-tetrahydrofolate reductase (MTHFR), which act together to generate a proton gradient for electron transport phosphorylation [54]. The bifurcating ferredoxin:NAD hydrogenase (HYDFDNr} and electron bifurcating ferredoxin:NADP oxidoreductase (FRNDPRr) are also considered important for ATP generation during autotrophic growth. To investigate this mechanism, the investigators changed exchange fluxes for CO_2 and H_2 and monitored ATP flux. The study found that there was no change in ATP flux when such changes were made. The second mechanism proposed by Schuchmann and Muller assumes that ferredoxin hydrogenase (FRHD) is the enzyme required for energy conservation [55]. For this mechanism, HYDFDNr reaction with different stoichiometries than that used in the first mechanism is proposed. Using this hypothesis, the simulations were carried out. Similar results for ATP production were observed when the exchange fluxes for CO_2 and H_2 were varied. However, when the stoichiometry of reactions catalyzed by FRHD and HYDFDNr was changed that, the simulation results were comparable to experimental results for ATP production. A linear relationship between ATP flux and CO_2 and H_2 supply was observed during the simulation of the model containing the stoichiometric changes. Hence, the investigators have proposed that for energy conservation on autotrophic substrates, the proposed second mechanism requires modification.

4.7 Streptococcus thermophilus

S. thermophilus is an important organism in the dairy industry, especially in the production of yoghurt and cheese [56]. It is a borderline thermophile with an optimum growth temperature of 45°C [56]. The genome-scale reconstruction of *S. thermophilus* LMG18311 was constructed by Pastnik et al. to study its amino acid metabolism. The reconstruction was based on closely related organisms such as *L. planatarum* and *L. lactis*. Through a manual gap-filling procedure using literature and experimental evidence, the reconstruction was completed. The model consists of 429 genes and 522 reactions. The biomass was determined in the study itself, and hence makes this model more relevant.

The model could accurately predict that the organism cannot grow on histidine because of the lack of histidine biosynthesis genes. Similarly, the model analysis also determined that *ychE*, a gene involved in the synthesis of cysteine from methionine in *L. lactis*, is truncated in *S. thermophilus*, which could explain its apparent auxotrophy to either of these amino acids. Furthermore, the model could correctly suggest that homofermentative lactic acid production is the primary metabolism in *S. thermophilus*.

4.8 Thermoanaerobacterium saccharolyticum

T. saccharolyticum is a gram-positive anaerobe that is chemoorganotrophic in nature. It is known for high ethanol yields from hexoses and pentoses. Curie et al., to study the metabolic capabilities of *T. saccharolyticum*, created a genome-scale metabolic model of the organism by parsing information from the reconstruction of *C. thermocellum* ATCC 27405 [57]. After the construction of the initial model, gap-filling was carried out using literature information and FBA-Gap [57]. This gap-filling algorithm proposes a minimal set of reactions from a curated reactome database to be added to the model to support biomass synthesis. The refined model consists of 516 reactions, 315 genes, and 528 metabolites [57].

The investigators used experimental data to constrain hydrogenase reactions that contributed less toward hydrogen production. Specifically, the energy-conserving hydrogenase (ECH) was blocked, the bifurcating hydrogenase (BIFH2) and NADH hydrogenase (NADH2) were made irreversible, and the hydrogen export was constrained to reflect the experimental data. These constraints were shown to alter fluxes dramatically, and more accurately predicted the higher ethanol production observed experimentally.

The model was tested for gene knockouts to enhance the production of ethanol. The simulation predicted that the deletion of lactate dehydrogenase (LDH) and phosphotransacetylase (PTA) leads to optimal growth and high ethanol yield. The model also suggested that the deletion of reactions catalyzed by LDH and ferredoxin hydrogenase (HFS) leads to high ethanol yield at the expense of growth rate. The predictions were consistent with experimental results. Furthermore, the model indicated that the deletion of LDH, HFS, and glutamate dehydrogenase (GDH) leads to a marginal increase in ethanol production compared to that of LDH and HFS only. Overall, the model was successful in predicting the metabolic behavior of *T. saccharolyticum* when grown on cellobiose.

4.9 Models Deposited in BioModels Database

In addition to the curated models above, there have been a number of models of thermophiles automatically generated using genome annotations and deposited in the BioModels Database [58, 59]. These include *Pyrolobus fumarii* ([60], BMID:140676), *Pyrococcus furiosus* ([61], BMID:141276), *Archaeoglobus fulgidus* ATCC 49558 ([62], BMID:140871), *Methanococcus jannaschii* ([63], BMID:140493), *Aeropyrum pernix* ([64], BMID:142009), *Aquifex aeolicus* ([65], BMID:141549), *Hyperthermus butylicus* ([66], BMID:140823), *Desulfurococcus kamchatkensis* ([67], BMID:141539), *Desulfurococcus mucosus* ([68], BMID:141869), *Staphylothermus hellenicus* ([69], BMID:140958), and *Alicyclobacillus acidocaldarius* ([70], BMID:140735). As with all automatic reconstructions, these models need more curation, but they serve as a platform for further development and can only increase in applicability.

5 Conclusion

The range of metabolisms exemplified by thermophilic microorganisms is quite wide, considering the relatively few places on Earth where they thrive. Metabolic network modeling is an effective way to study the metabolism of thermophiles and to compare their metabolism to their mesophilic counterparts. We have chronicled several cases where thermophilic microorganisms have been studied with genome-scale models. However, there are still many challenges for expanding the scope of metabolic network models of thermophiles, such as estimating thermodynamic parameters at higher temperatures, measuring kinetic parameters of key metabolic enzymes, and fully understanding how cofactor usage changes at high temperature (e.g., the preference of ATP versus pyrophosphate as an energy carrier).

Despite the aforementioned issues, genome-scale models of thermophilic organisms are very useful tools for understanding and engineering the metabolisms of non-model strains to enhance their ability for use in the biofuel, waste management, and mining industries. As more data are acquired for thermophiles, in particular large OMICs datasets, the metabolic models continue to improve.

References

1. Brock TD (1985) Life at high temperatures. Science 230:132–138
2. Brock TD, Freeze H (1969) Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. J Bacteriol 98(1):289–297
3. Bult CJ, White O, Olsen GJ, Zhou L (1996) Complete genome sequence of the methanogenic archaeon. Methanococcus jannaschii. Science 273:1058
4. Brock TD (1967) Life at high temperatures. Science 158:1012–1019
5. Robb F, Antranikian G, Grogan D, Driessen A (2007) Thermophiles: biology and technology at high temperatures. CRC Press
6. Caldwell D, Brannan D, Kieft T (1983) Thermothrix thiopara: selection and adaptation of a filamentous sulfur-oxidizing bacterium colonizing hot spring tufa at pH 7.0 and 74 C. Ecol Bull 38:129–134
7. Zeikus J (1979) Thermophilic bacteria: ecology, physiology and technology. Enzyme Microb Technol 1:243–252
8. Shelef G, Kimchie S, Grynberg H (1980) High-rate thermophilic anaerobic digestion of agricultural wastes. In Biotechnol Bioeng Symp (United States). Environmental and Water Resources Engineering Dept., Technion, Haifa, Israel
9. Gajalakshmi S, Abbasi S (2008) Solid waste management by composting: state of the art. Crit Rev Environ Sci Technol 38:311–400
10. Cecchi F, Pavan P, Alvarez JM, Bassetti A, Cozzolino C (1991) Anaerobic digestion of municipal solid waste: thermophilic vs. mesophilic performance at high solids. Waste Manag Res 9:305–315

11. Micolucci F, Gottardo M, Cavinato C, Pavan P, Bolzonella D (2016) Mesophilic and thermophilic anaerobic digestion of the liquid fraction of pressed biowaste for high energy yields recovery. Waste Manag 48:227–235
12. Deveci H, Akcil A, Alp I (2004) Bioleaching of complex zinc sulphides using mesophilic and thermophilic bacteria: comparative importance of pH and iron. Hydrometallurgy 73:293–303
13. Krebs W, Brombacher C, Bosshard PP, Bachofen R, Brandl H (1997) Microbial recovery of metals from solids. FEMS Microbiol Rev 20:605–617
14. Barrett J (1990) Metal extraction by bacterial oxidation of minerals. Horwood
15. Rossi G (1990) Biohydrometallurgy. McGraw-Hill
16. Bobadilla-Fazzini RA, Cortés MP, Maass A, Parada P (2014) Sulfobacillus thermosulfidooxidans strain Cutipay enhances chalcopyrite bioleaching under moderate thermophilic conditions in the presence of chloride ion. AMB Express 4:1
17. Zhang L, Wu J, Wang Y, Wan L, Mao F, Zhang W, Chen X, Zhou H (2014) Influence of bioaugmentation with Ferroplasma thermophilum on chalcopyrite bioleaching and microbial community structure. Hydrometallurgy 146:15–23
18. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320
19. Varma A, Palsson B (1994) Metabolic flux balancing: basic concepts, scientific and practical use. Nat Biotechnol 12:994–998
20. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput Biol 9:e1002980
21. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6:1290–1307
22. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE (2016) BiGG models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res 44:D515–D522
23. Roberts S, Gowen C, Brooks JP, Fong S (2010) Genome-scale metabolic analysis of *Clostridium thermocellum* for bioethanol production. BMC Syst Biol 4:31
24. Roberts SB, Gowen CM, Brooks JP, Fong SS (2010) Genome-scale metabolic analysis of Clostridium thermocellum for bioethanol production. BMC Syst Biol 4:1
25. Milton H, Reddy VJ, Tamang D, Västermark A (2014) The transporter classification database. Nucleic Acids Res 42:251–258
26. Gowen CM, Fong SS (2010) Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of *Clostridium thermocellum*. Biotechnol J 5:759–767
27. Thompson RA, Layton DS, Guss AM, Olson DG, Lynd LR, Trinh CT (2015) Elucidating central metabolic redox obstacles hindering ethanol production in *Clostridium thermocellum*. Metab Eng 32:207–219
28. Zhou J, Olson DG, Argyros DA, Deng Y, van Gulik WM, van Dijken JP, Lynd LR (2013) Atypical glycolysis in Clostridium thermocellum. Appl Environ Microbiol 79:3000–3008
29. Feinberg L, Foden J, Barrett T, Davenport KW, Bruce D, Detter C, Tapia R, Han C, Lapidus A, Lucas S et al (2011) Complete genome sequence of the cellulolytic thermophile *Clostridium thermocellum* DSM1313. J Bacteriol 193:2906–2907
30. Tripathi SA, Olson DG, Argyros DA, Miller BB, Barrett TF, Murphy DM, McCool JD, Warner AK, Rajgarhia VB, Lynd LR et al (2010) Development of *pyrF*-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a pta mutant. Appl Environ Microbiol 76:6591–6599
31. Thompson RA, Dahal S, Garcia S, Nookaew I, Trinh CT (2016) Exploring complex cellular phenotypes and model-guided strain design with a novel genome-scale metabolic model of Clostridium thermocellum DSM 1313 implementing an adjustable cellulosome. Biotechnology Biofuels 9:194
32. Ozaki S, Fujimitsu K, Kurumizaka H, Katayama T (2006) The DnaA homolog of the hyperthermophilic eubacterium Thermotoga maritima forms an open complex with a minimal 149-bp origin region in an ATP-dependent manner. Genes Cells 11:425–438

33. Huber R, Langworthy TA, König H, Thomm M, Woese CR, Sleytr UB, Stetter KO (1986) Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. Arch Microbiol 144:324–333
34. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA et al (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329
35. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA (2009) Three-dimensional structural view of the central metabolic network of Thermotoga maritima. Science 325:1544–1549
36. Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425
37. Nogales J, Gudmundsson S, Thiele I (2012) An in silico re-design of the metabolism in Thermotoga maritima for increased biohydrogen production. Int J Hydrogen Energy 37:12205–12218
38. Lee N-R, Lakshmanan M, Aggarwal S, Song J-W, Karimi IA, Lee D-Y, Park J-B (2014) Genome-scale metabolic network reconstruction and in silico flux analysis of the thermophilic bacterium Thermus thermophilus HB27. Microb Cell Fact 13:1
39. Kaneda T (1991) Iso-and anteiso-fatty acids in bacteria: biosynthesis, function, and taxonomic significance. Microbiol Rev 55:288–302
40. Nordström KM, Laakso SV (1992) Effect of growth temperature on fatty acid composition of ten thermus strains. Appl Environ Microbiol 58:1656–1660
41. Pask-Hughes RA, Shaw N (1982) Glycolipids from some extreme thermophilic bacteria belonging to the genus Thermus. J Bacteriol 149:54–58
42. Oshima T (2007) Unique polyamines produced by an extreme thermophile, Thermus thermophilus. Amino Acids 33:367–372
43. Ulas T, Riemer SA, Zaparty M, Siebers B, Schomburg D (2012) Genome-scale reconstruction and analysis of the metabolic network in the hyperthermophilic archaeon Sulfolobus solfataricus. PLoS One 7:e43401
44. Teufel R, Kung JW, Kockelkorn D, Alber BE, Fuchs G (2009) 3-Hydroxypropionyl-coenzyme A dehydratase and acryloyl-coenzyme A reductase, enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in the Sulfolobales. J Bacteriol 191:4572–4581
45. Berg IA, Kockelkorn D, Ramos-Vera WH, Say RF, Zarzycki J, Hügler M, Alber BE, Fuchs G (2010) Autotrophic carbon fixation in archaea. Nat Rev Microbiol 8:447–460
46. Quester S, Schomburg D (2011) EnzymeDetector: an integrated enzyme function prediction tool and database. BMC Bioinformatics 12:376
47. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
48. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 42:D459–D471
49. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Söhngen C, Stelzer M, Thiele J, Schomburg D (2010) BRENDA, the enzyme information system in 2011. Nucleic Acids Res gkq1089
50. Wilson DB (2004) Studies of Thermobifida fusca plant cell wall degrading enzymes. Chem Rec 4:72–82
51. Deng Y, Fong SS (2011) Metabolic engineering of Thermobifida fusca for direct aerobic bioconversion of untreated lignocellulosic biomass to 1-propanol. Metab Eng 13:570–577
52. Vanee N, Brooks JP, Spicer V, Shamshurin D, Krokhin O, Wilkins JA, Deng Y, Fong SS (2014) Proteomics-based metabolic modeling and characterization of the cellulolytic bacterium Thermobifida fusca. BMC Syst Biol 8:1
53. Islam MA, Zengler K, Edwards EA, Mahadevan R, Stephanopoulos G (2015) Investigating Moorella thermoacetica metabolism with a genome-scale constraint-based metabolic model. Integr Biol 7:869–882

54. Mock J, Wang S, Huang H, Kahnt J, Thauer RK (2014) Evidence for a hexaheteromeric methylenetetrahydrofolate reductase in Moorella thermoacetica. J Bacteriol 196:3303–3314
55. Schuchmann K, Müller V (2014) Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. Nat Rev Microbiol 12:809–821
56. Hols P, Hancy F, Fontaine L, Grossiord B, Prozzi D, Leblond-Bourget N, Decaris B, Bolotin A, Delorme C, Ehrlich SD (2005) New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics. FEMS Microbiol Rev 29:435–463
57. Currie DH, Raman B, Gowen CM, Tschaplinski TJ, Land ML, Brown SD, Covalla SF, Klingeman DM, Yang ZK, Engle NL (2015) Genome-scale resources for Thermoanaerobacterium saccharolyticum. BMC Syst Biol 9:1
58. Chelliah V, Juty N, Ajmera I, Ali R, Dumousseau M, Glont M, Hucka M, Jalowicki G, Keating S, Knight-Schrijver V et al (2015) BioModels: ten-year anniversary. Nucleic Acids Res 43:D542–D548
59. Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B et al (2006) BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34:D689–D691
60. Anderson I, Göker M, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng J-F, Tapia R, Han C, Goodwin L et al (2011) Complete genome sequence of the hyperthermophilic chemolithoautotroph Pyrolobus fumarii type strain (1A(T)). Stand Genomic Sci 4:381–392
61. Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB, Dunn DM (2001) Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology. Methods Enzymol 330:134–157
62. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD et al (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature 390:364–370
63. Tsoka S, Simon D, Ouzounis CA (2004) Automated metabolic reconstruction for Methanococcus jannaschii. Archaea 1:223–229
64. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A et al (1999) Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, Aeropyrum pernix K1. DNA Res 6:83–101, 145–152
65. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M et al (1998) The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature 392:353–358
66. Brügger K, Chen L, Stark M, Zibat A, Redder P, Ruepp A, Awayez M, She Q, Garrett RA, Klenk H-P (2007) The genome of Hyperthermus butylicus: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 degrees C. Archaea (Vancouver, BC) 2:127–135
67. Ravin NV, Mardanov AV, Beletsky AV, Kublanov IV, Kolganova TV, Lebedinsky AV, Chernyh NA, Bonch-Osmolovskaya EA, Skryabin KG (2009) Complete genome sequence of the anaerobic, protein-degrading hyperthermophilic crenarchaeon Desulfurococcus kamchatkensis. J Bacteriol 191:2371–2379
68. Wirth R, Chertkov O, Held B, Lapidus A, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng JF, Tapia R et al (2011) Complete genome sequence of Desulfurococcus mucosus type strain (O7/1). Stand Genomic Sci 4:173–182
69. Anderson I, Wirth R, Lucas S, Copeland A, Lapidus A, Cheng JF, Goodwin L, Pitluck S, Davenport K, Detter JC et al (2011) Complete genome sequence of Staphylothermus hellenicus P8. Stand Genomic Sci 5:12–20
70. Mavromatis K, Sikorski J, Lapidus A, Glavina Del Rio T, Copeland A, Tice H, Cheng J-F, Lucas S, Chen F, Nolan M et al. (2010) Complete genome sequence of Alicyclobacillus acidocaldarius type strain (104-IA). Stand Genomic Sci 2:9–18

Adv Biochem Eng Biotechnol (2017) 160: 121–142
DOI: 10.1007/10_2016_38

Published online: 26 October 2016

Networking Omic Data to Envisage Systems Biological Regulation

Saowalak Kalapanulak, Treenut Saithong, and Chinae Thammarongtham

Abstract To understand how biological processes work, it is necessary to explore the systematic regulation governing the behaviour of the processes. Not only driving the normal behavior of organisms, the systematic regulation evidently underlies the temporal responses to surrounding environments (dynamics) and long-term phenotypic adaptation (evolution). The systematic regulation is, in effect, formulated from the regulatory components which collaboratively work together as a network. In the drive to decipher such a code of lives, a spectrum of technologies has continuously been developed in the post-genomic era. With current advances, high-throughput sequencing technologies are tremendously powerful for facilitating genomics and systems biology studies in the attempt to understand system regulation inside the cells. The ability to explore relevant regulatory components which infer transcriptional and signaling regulation, driving core cellular processes, is thus enhanced. This chapter reviews high-throughput sequencing technologies, including second and third generation sequencing technologies, which support the investigation of genomics and transcriptomics data. Utilization of this high-throughput data to form the virtual network of systems regulation is explained, particularly transcriptional regulatory networks. Analysis of the resulting

S. Kalapanulak and T. Saithong (✉)
Bioinformatics and Systems Biology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, Thailand

Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, Thailand
e-mail: saowalak.kal@kmutt.ac.th; treenut.sai@kmutt.ac.th

C. Thammarongtham (✉)
Biochemical Engineering and Pilot Plant Research and Development Laboratory, National Center for Genetic Engineering and Biotechnology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, Thailand
e-mail: chinae@biotec.or.th

regulatory networks could lead to an understanding of cellular systems regulation at the mechanistic and dynamics levels. The great contribution of the biological networking approach to envisage systems regulation is finally demonstrated by a broad range of examples.

Keywords High throughput sequencing, Network biology, Network reconstruction, Systems regulation

Contents

1 Introduction 122
2 High-Throughput Sequencing Technology to Discover Regulatory Elements 123
2.1 Second Generation Sequencing: Early Age of High-Throughput Sequencing Technology 123
2.2 Third Generation Sequencing: Empowering Detection of a Regulatory Molecule ... 125
3 Approaches to Link Regulatory Components to Reflect a Network of Systems Regulation 127
3.1 Biological Networks 127
3.2 Paths from Molecular Components to Biological Network 130
4 Network Analysis to Envisage the Systems Regulation: Inference and Applications 133
4.1 Inference of System Regulation by Means of Biological Networks 133
4.2 Applications of Biological Network Analysis to Understand the System Regulation: From Unicellular Organisms to Plants 134
5 Conclusions 136
References 136

1 Introduction

Studying the biology of organisms in the context of networks is a promising strategy to decipher the code of systems regulation in modern life science research. With the advances in omics technologies and a growing number of sequenced genomes in public databases, the systems biology approach has been applied to integrate all jigsaw information and demonstrate the global view of the regulation system. The omics data are exploited beyond their primary implication of the static expression and are utilized to infer rather dynamic regulation in the systems context (e.g., [1]). Here, we pursue the perspective of using network biology as a means to touch the systems regulation of a cell. The review highlights development of high-throughput sequencing technology and its applications to acquire a biological network. Lastly, we recapitulate our review using evidence from many success cases studied in a wide range of organisms from a simple single cell to a complex multi cell. These examples illustrate the synergized contribution of high-throughput sequencing technology and network biology approach to raising current biological research to the level of systems regulation inference.

2 High-Throughput Sequencing Technology to Discover Regulatory Elements

During the past few decades, technological advancement contributing to life sciences has been progressing rapidly. Among "systems technologies," high-throughput sequencing is one of the potent drivers of biological research. High-throughput sequencing technology has emerged over the last decade, after completion of the first human genome draft. Many genomics research scientists and genome research centers, including the National Human Genome Research Institute (NHGRI), have considered high-throughput sequencing technology as a core for driving genomics and systems biology research [2]. In addition to the vast sequence data (per run) obtained from this sequencing technology, the cost per base of sequencing is substantially lower compared to the traditional Sanger sequencing technology. A milestone of high-throughput sequencing was to achieve whole genome sequencing of an individual person at a cost of 1,000 US$ [3–5]. This produces a major type of omic data from sequencing data, which can be applied to various biological research. Interestingly, the emerging sequencing technology nowadays not only promotes genomic study but also makes an immense contribution to driving the transcriptomic, that is RNAseq [6, 7], and interatomic, for example ChIP-seq [8, 9], research. As a ground-breaking technology that broadly supports all dimensions of biological study, especially cellular regulation, the principles of high-throughput sequencing technologies have been described in several specific review articles [10–13] during recent years. At least three significant improvements in the technology are considered to shift the paradigm of high-throughput sequencing measurement and, as a consequence, empowering the accessibility to observe what really happens in a cell. Briefly, the GS instrument was launched on the market in 2005 by 454 Life Sciences. It was the first Next Generation Sequencing (NGS) system based on pyrosequencing. Later, other platforms of second generation sequencing technologies, namely AB SOLiD, Ion Torrent, and Illumina platforms, were introduced.

2.1 Second Generation Sequencing: Early Age of High-Throughput Sequencing Technology

In the pre-genomic era, a huge effort was required to obtain a gene sequence, where it was nearly impossible to attain whole genome sequences and to follow the abundance of expressed nucleotide sequences simultaneously. The introduction of the first high-throughput sequencing technology, 454 pyrosequencing, has changed the methods of biological study as it significantly pushes the limit of genetic decoding experimentation. The pyrosequencing-based GS instrument was introduced by 454 Life Sciences in 2005. At that time, the high-throughput of this technology made known the term "next generation sequencing." One year later the alternative platform of high-throughput sequencing, AB SOLiD (Sequencing by Oligo Ligation Detection) system, came onto the market. The key technology for the released AB SOLiD is ligation sequencing. Although earlier generations of SOLiD reads was 35 bp in length, but was also based on a two-base sequencing method, SOLiD could give high accuracy up to 99.9% [14]. Later, SOLiD produced longer reads (85 bp) with higher accuracy and a larger throughput of data.

A well-known sequencing system from Illumina is HiSeq 2000 which was brought to the market in 2010. Previously, the Genome Analyzer system was launched by Solexa in 2006. The company was then purchased by Illumina in 2007. This system platform is based on sequencing by synthesis technology. HiSeq 2000 gives 600 Gigabases (Gb) output per run when on high output run mode. It was claimed that HiSeq 2000 is the cheapest in terms of sequencing cost per Megabase compared to 454 and SOLiD [15]. A smaller scale system, Illumina MiSeq, which is based on similar technology, was released in 2011. The MiSeq system takes less running time, only 24 h, generating 4.5 Gb of sequence data using MiSeq reagent kit version 2. MiSeq can generate 250 bp reads with paired-end sequencing. The current platform of Illumina is HiSeq 2500 which can give up to 1 TB output based on HiSeq v4 chemistry. Although HiSeq was considered the industry standard for high-throughput sequencing technology according to its throughput, the read length obtained from this platform is approximately 250 bases or less. The required amplification of the template prior to sequencing can cause content-bias base error. Currently, Illumina offers a novel Illumina TruSeq synthetic long read strategy as an improved technology, which can give read lengths of 1.5–18.5 kbp with a very low error rate [16]. This technology was proposed to be a powerful technique for de novo whole genome assemblies.

The Ion Torrent platform was introduced by PostLight Sequencing Technology in 2010. This technology was later acquired and commercialized by Life Technologies Corp. Ion Torrent sequencing depends on monitoring hydrogen ions released as a by-product during nucleotide incorporation [17]. The sequencing reaction is performed within the micro-wells of the Ion Chip, a silicon semiconductor-sensing chip specifically designed to detect pH changes using hydrogen ion sensors at the base of the wells. This type of sensing eliminates the light, scanning, and cameras required for detecting sequencing process signals which reduce time for sequencing. In Ion Torrent sequencing reactions, native (unlabeled) nucleotides are used for polymerization. Therefore, the noise caused by fluorescence or blocking substances on the reactants is omitted. Sequences with homopolymer bases are a major concern for Ion Torrent sequencing, causing insertion or deletion errors [11]. In addition, errors of base substitution can arise. Nowadays, the throughput can average up to 1 Gb per run using the high well density Ion Chip. With 200-base reads, the Ion Proton platform can perform several sequencing applications, including transcriptome and multiplexing amplicons paired-end sequencing.

2.2 *Third Generation Sequencing: Empowering Detection of a Regulatory Molecule*

With the advent of second generation sequencing technology, much genome sequence data including the genomes of non-model organisms has been produced. However, short read-lengths and bias genome coverage may lead to fragmented genome assemblies. Therefore, third generation sequencing is being introduced with new technology for DNA sequencing. The key features of third generation sequencing include amplification-independent technology and real-time signal detection along with the sequencing reaction [18, 19]. Third generation sequencing technology has the potential to generate terabase-scale sequence data at little cost [20].

Pacific BioSciences introduced Single Molecule Real Time (SMRT) PacBio RS, a third generation sequencing platform. With this technology, during the enzymatic reaction of nucleotide incorporation into the complementary strand, the fluorescent dye linked with the incorporated nucleotide is cleaved off and detected for the signal immediately [21]. The explicit advantages over second generation sequencing technology include no amplification step, which reduces time and error of the polymerase chain reaction (PCR), a short sequencing run time within 1 day, and average longer (more than 1 kb) reads compared to the second generation sequencing. Although PacBio RS gives a lower throughput compared to those of the second generation sequencing, this platform receives much attention because of its longer read length and lower error. For the PacBio platform, the errors observed in sequencing reads are random errors, not context specific errors as found in the reads of other platforms [15, 22]. The error rates of single pass sequence reads are approximately 11%, according to company information. However, the consensus error rate is significantly lower as template DNA molecules are sequenced several times when the circular consensus mode is run. The single pass errors are reduced during the consensus building step, giving assembled contigs with high consensus accuracy (99.999%, Q50) regardless of the sequence context or GC content of the DNA templates [23, 24]. The non-sequence context bias is a key feature of the PacBio platform involved in producing high accuracy sequencing reads. This makes it possible to overcome sequencing of long tandem repeat regions [25]. Several studies have focused on the performance of the Pacific BioSciences sequencing platform. A portion of error-free sequencing reads without a single mismatch or indel, and was found to be 0% compared to reads of other sequencing platforms [15]. An evaluation performing chloroplast sequencing, contigs assembled from PacBio data concur mostly with Illumina contigs generated and can resolve unambiguous and misassemblies [26]. In a comparative study, the performance of this platform was considered to be outperforming the other platforms in terms of contig length obtained from de novo assembly [27]. With the recent advances, PacBio read lengths have been improved with the median and maximum of 10 kbp and 50 kbp, respectively [28]. Reasonably, PacBio sequencing platform is considered to be useful for both genome re-sequencing and de novo genome sequencing and assembly because the long sequencing reads are considered to be able to solve gap closing of the genome finishing process [29].

Oxford Nanopore MinION is another platform for third generation sequencing technology. This type of sequencing exploits nanopores of a particular protein, alpha hemolysin, facilitating the sequencing [30]. It is a thumb-drive sized device rather than a traditional sequencing instrument. Nanopore sequencing offers several advantages over other high-throughput sequencing platforms including its small size and low cost. Sequencing of single strand DNA is processed during depolymerization, not polymerization or synthesis, and therefore no PCR amplification. Several biochemical steps are not required, only exonuclease is used for depolymerization. In addition, fluorescent labeling is omitted as the detection is based on voltage disruption across the nanopore by different sized deoxyribinucleoside monophosphate released during DNA strand depolymerization. This results in a reduced time required for sample preparation. Another interesting advantage is that the Nanopore sequencing technique generates very long reads, potentially more than 5 kb at significantly high speed (1 bp/10 ns) [21]. Access to the MinION devices was available only for members of the testing program and little data are publicly available. Improvements for better quality accurate results have been made, including a hybrid method for read assembly [31].

These advanced sequencing technologies have facilitated the discovery or identification of several types of gene expression regulators and their functions, which are difficult to identify by traditional techniques. For example, the mechanism of a pathogenic bacterium regulator involved in stress response and chemotaxis has been studied [32]. Many regulatory genes related to activation of developmental processes have been identified in a non-model animal [33]. Wood plant long non-coding RNAs, an emerging type of regulator, involved in growth development and wood formation, have been identified from high-throughput sequencing data [34]. A number of regulatory elements in humans have been identified via high-throughput sequencing [35]. These are only a few examples of high-throughput sequencing-based identification contributing to the complex regulatory network of biological systems. As large numbers of regulatory elements remain undiscovered, further development of high-throughput sequencing technologies is key to unraveling many other regulatory components.

With the emergence of the high-throughput sequencing technology, life science research has progressed rapidly using a genomics and systems biology approach. Several techniques have been applied to various areas, namely medicine, clinical diagnosis, drug development, agricultural improvement, environmental investigation, and industrial biotechnology development of microbial products. Such techniques include whole genome re-sequencing, de novo genome assembly, transcriptome sequencing analysis, epigenomics, and metagenomics. To move the current understanding of cellular regulation forward, from component-based to systems-based, high-throughput technology was employed beyond the identification task. An example is the applications of RNAseq to explore the gene regulatory network through detection of gene expression patterns [36, 37].

3 Approaches to Link Regulatory Components to Reflect a Network of Systems Regulation

The first step to untangle the intracellular regulation underlying the behavioral phenotypes of an organism is the exploration and identification of the components involved (Fig. 1). The wide spectra of advanced and high-throughput technologies have been developed to provide a precise and accurate measurement when capturing exhaustive molecular components inside cellular space. As the progress of technology development is in a good position and direction, the greatest challenge is supporting the capability of analyzing large amounts of data. The interesting molecular components believed to play an important role in the studied conditions have been successfully identified many times via high-throughput data and advanced bioinformatics methods. However, how these molecular components interact, leading to the observed phenotypes such as high-disease resistant cultivars in plants [38–41] and high ethanol-producing strain in yeast [42, 43], is still a mystery. Understanding the interactions between intracellular components inside the cell is crucial for exploring biological regulation in each particular biological process across transcription, translation, and performing functions. One gene or one protein cannot perform the functions, but they work together as a network to demonstrate the phenotypes. Network biology has become an important field in systems biology research [44, 45]. Not only do biological networks contribute as an overview of the system under investigation but they are also an integration platform for combining biological data from many different studies into a single framework.

3.1 Biological Networks

Intracellular components, including genes, proteins, mRNAs, microRNAs, and metabolites, work together elaborately as a network. The relationships between these components are, therefore, usually represented in the form of a biological network where nodes of components are linked to their interactive partners. There are different types of biological networks which basically describe the nature of the regulatory activities occurring in each level of biological regulation. Table 1 presents examples of biological networks, their corresponding biological components in the networks, and the vital information for network reconstruction. Metabolic networks demonstrate the interactions between metabolites catalyzed by enzymes in the metabolism [46, 47]. Gene regulatory networks provide the matrix of gene-gene relationship which is normally inferred by their co-expression evidence in several conditions. Transcriptional regulatory networks (TRN), which emphasize more the gene transcription control, usually exhibit the functional association between transcription factors (TFs) and their target genes (TGs). For instance, TRN of yeast, a model organism of eukaryote, was constructed from the 12,873 known regulatory interactions between 157 TFs and their 4,410 target genes, shown

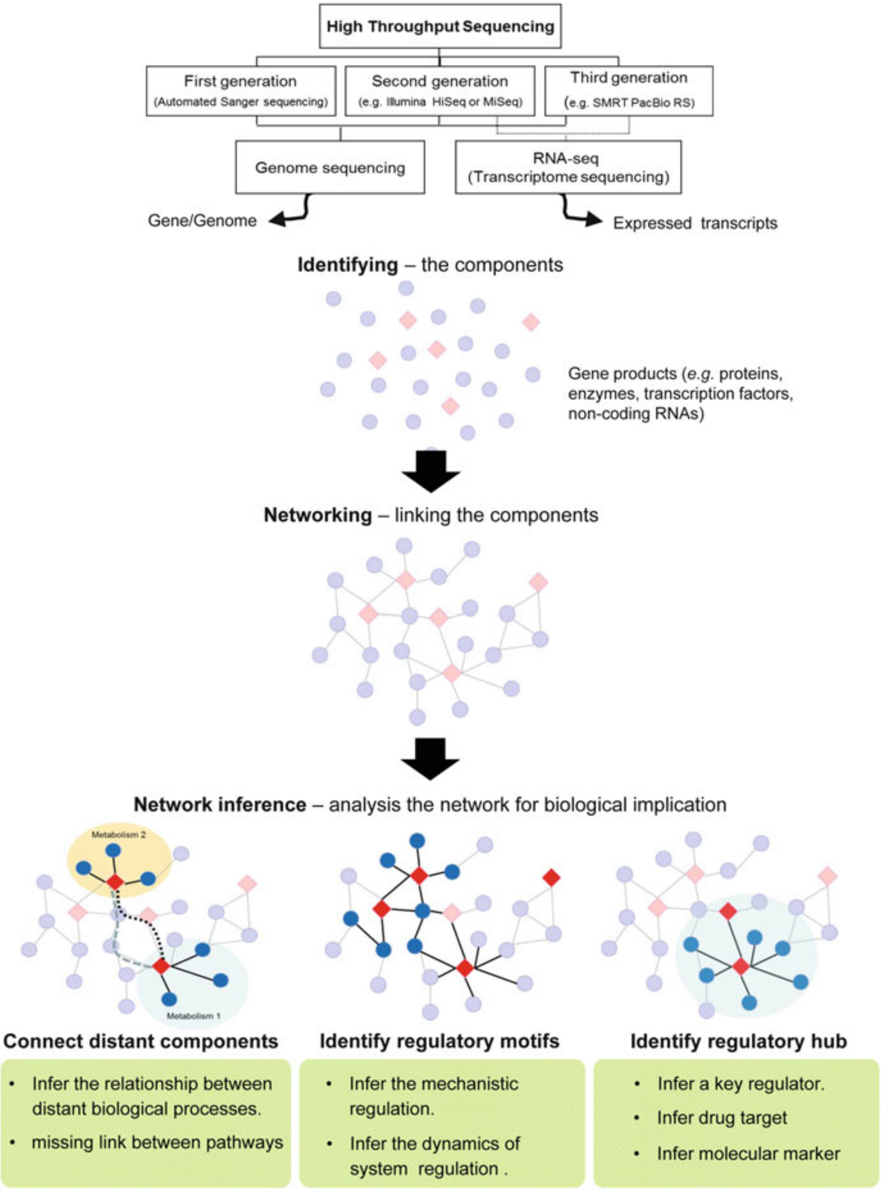

High Throughput Sequencing
First generation
(Automated Sanger sequencing)
Second generation
(e.g. Illumina HiSeq or MiSeq)
Third generation
(e.g. SMRT PacBio RS)
Genome sequencing
RNA-seq
(Transcriptome sequencing)
Gene/Genome
Expressed transcripts
Identifying – the components
Gene products (e.g. proteins, enzymes, transcription factors, non-coding RNAs)
Networking – linking the components
Network inference – analysis the network for biological implication
Metabolism 2
Metabolism 1
Connect distant components
Identify regulatory motifs
Identify regulatory hub
• Infer the relationship between distant biological processes.
• missing link between pathways
• Infer the mechanistic regulation.
• Infer the dynamics of system regulation .
• Infer a key regulator.
• Infer drug target
• Infer molecular marker

Fig. 1 Schematic of network-based regulatory study

as a graph in Fig. 2. The complex transcriptional regulatory interactions were demonstrated even in the very simple eukaryotic cell organism. To investigate inside the network, network decomposition into the small sub-networks, called network motifs, was examined to explore the function of each TF. The network

Table 1 Examples of biological networks

Types of biological network	Biological/molecular components in the network	Required information for biological network reconstruction
1. Metabolic network	• Metabolites • Enzymes	• Substrates and products of each enzyme • Direction of each enzymatic reaction • Compartmentalization of each reaction inside the cell
2. Gene regulatory network/ transcription regulatory network	• Transcription factor-coding genes • Promoter of target genes such as protein-coding genes, non-protein coding genes etc.	• Transcription factor (TF)–DNA interactions • Promoter region of target genes • Transcription factor binding sites (TFBS) • Condition specific for TF–DNA interactions
3. Cell signaling network	• Signal such as external metabolites, temperature, light, etc. • Signaling receptor proteins • Protein kinase • Protein phosphatases	• Ligand–receptor interactions • Receptor–intracellular component interactions • Protein–protein interactions • Protein–DNA interactions
4. Disease-gene network	• Distinct disorder or disease • Disorder-corresponding genes	• Genes causing disease for each disorder
5. Drug-target network	• Drugs • Target genes	• Drug–target interactions

motifs, basic units making up the network structure, are often occurring patterns of interactions among the regulatory proteins. The instances of regulatory network motifs often found in the transcriptional regulation of cells are auto-regulation, feed-forward loop, multi-component loop, single input, multiple input, and regulator chain (Fig. 2) [50]. For example, auto-regulation is where a TF-encoding gene translates to a TF protein and the TF protein itself functions as a regulator binding to the promoter region of the TF-coding gene. Moreover, the multiple input motif is described as one target gene that can be controlled by many transcription factors. Besides the two networks mentioned above, cell signaling network, kinase-substrate networks, protein–protein interaction networks, disease-gene networks, and drug-target networks have been reconstructed for several purposes [51–53] and visualized via many useful free software tools including Pajek [54], Cytoscape [55], Genes2Networks [56], and FANMOD [57].

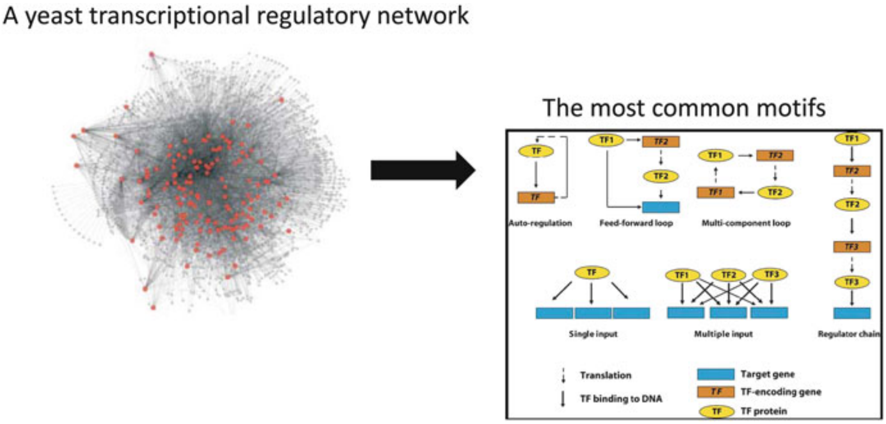


Fig. 2 Transcriptional regulatory network of yeast constructed from 12,873 known relationships between transcription factors and their target genes [48, 49] can be decomposed to be several common motifs [50]

3.2 Paths from Molecular Components to Biological Network

Systems biology approaches can allow biologists to move beyond a reductionistic approach. Biological networks derived from genome-wide information provide the global view of molecular components and their association under certain conditional contexts. They can expand the vision of researchers over the set of just a few genes within the scope of their experience. Furthermore, as an integrative approach, the systematic view could be achieved by linking the biological networks to conjecture the interactions of multiple levels of biological regulation through the relationship between communicating molecular components, such as metabolites, enzymes, genes, and transcription factors. A useful reconstructed framework has been provided for biologists to reveal the organization of life. Several different methods have been employed to exploit biological networks. Here, this review is more focused on the transcriptional regulatory networks (TRN) demonstrating the interactions between transcription factors and their target genes. There are three main computational approaches for TRN reconstruction as summarized in Table 2. First, the *template-based* method is derived based on the hypothesis that the relationship between genes and their transcriptional regulators could be inherited through evolutionary lineage. The orthologous transcription factors, in principle, regulate the expression of the corresponding orthologous target genes in the same manner as their templates. For this approach, a known regulatory network from a well-studied organism transfers the information about interactions to genes that have been determined to be orthologous in a target genome of interest [67]. However, the interested organisms should be closely related to the model organisms in terms of evolution to reduce the false positive in the inferred TRN. This approach

Table 2 Systems biology-based approaches for constructing transcriptional regulatory networks (TRNs)

Approaches and required information	Limitations	Advantages	Examples/ applications
1. Template-based method			
Required information • The well-characterized TRN of template organism • Whole genome sequences of template and interested organisms	• Closely-related organisms in terms of evolution between template and interested organism are required	• Simple method	• Evolution study of prokaryotic TRNs [58] • Proposing the transcription factors controlling Rubisco genes in cassava [59] • Bacterial TRNs construction [60]
2. Inferring network by predicting *cis*-regulatory elements			
Required information • Experimental-characterized TFBSs of related organisms • Upstream region of orthologous genes in multiple organisms for identifying conserved TFBS	• Scanning only known TFBSs on the promoter regions of interested organism	• Novel TFBSs can be identified	• Prediction of microbial regulatory elements controlling gene expression [61] • TRN construction of *Staphylococcus aureus* [62] • Identification of *cis*-regulatory in *Shewanella* genomes [63] • Identification of transcriptional regulatory region in *Drosophila* [64]
3. Reverse engineering using gene expression data			
Required information • Series-transcriptome data from microarray or RNA-sequencing platform	• Required more than two data points for gene co-expression profile analysis	• Condition-specific for TRNs	• TRN construction in human B cells [65] • TRN identification in bats [66]

has been applied to reconstruct the TRN of simple prokaryotic organism such as *Escherichia coli* and some multi-cellular organisms including plants [58–60]. Second, the networks are inferred by prediction of *cis-regulatory elements* in gene promoters. Based on the evolutionarily conserved transcription factor binding sites (TFBS), the promoter regions in the genome of interest are scanned using the known binding site profiles of characterized transcription factors. On the other hand, the novel transcription factor binding sites of the interested organism can be identified through the phylogenetic footprinting method [61]. The conserved sequences in the upstream region of orthologous genes from multiple organisms have been proposed as transcription factor binding sites of each gene in the investigated genome. It is based on the hypothesis that functional non-coding elements tend to evolve at a slower rate than non-functional surrounding elements because of the selective pressure, therefore demonstrating higher conservation during the time of evolution. Subsequently, the investigated genes predicted to have a binding site are hypothesized to be regulated by the corresponding transcription factor [62, 68, 69]. This approach has been widely applied for analysis of the *cis*-regulatory elements on the promoter regions in both prokaryotic and eukaryotic genomes, for instance *Shewanella* genomes [63] *and Drosophila* genome [64]. For these two approaches, comparative genomics has been employed to infer the unknown transcriptional regulation linkages between transcription factors and their target genes in the organisms of interest from well- characterized organisms such as *Arabidopsis*, a model of plants. Nevertheless, the gene regulation not only depends on evolutionary conservation but is also influenced by exposed conditions (i.e., biotic or abiotic stresses). Accordingly, the third method is developed based upon the transcriptional response under the studied conditions. It is named *reverse engineering* as it infers the TRN network using gene expression data. In this approach, patterns of gene expression from time-series experiments or from experiments conducted across several different conditions are employed for proposing the gene co-expression network. Normally, if a gene is upregulated following an increased expression of a transcription factor, or downregulated following the knockout of a transcription factor, a regulatory interaction between the two of them is inferred. On the other hand, for expression analysis over different experimental conditions, sets of genes with a similar expression profile across many conditions have been inferred to be co-regulated by the same set of transcription factors [65, 66, 70]. However, the inferences are more accurate if the number of expression data resolution increases because the distinguished direct regulatory interaction appears from the indirect or multi-step of regulation. The effect of the number of data points of gene expression data on the inferred TRN topology was investigated using two similar gene expression datasets of the *Arabidopsis* time-series microarray [71]. The TRN inference from high resolution datasets obtained significantly lower numbers of gene relationships than the other with low numbers of data points resulting in reduction of false positives in the inference network. Moreover, the TRN reconstructed from a high resolution dataset performed as a usual structure as in any biological network, namely with scale-free properties, in contrast to the TRN reconstructed from a low resolution dataset.

Although the aforementioned approaches were mainly developed from a single based principle, a great effort has been contributed to create a combinatorial strategy to increase the reliability of the predicted regulatory interactions [72, 73]. For example, a co-expression network constructed from microarray gene expression analysis has been combined with the scanning of all possible transcriptional factor binding sites on a promoter region of a particular gene member in the network [72]. Not only is the accuracy of predicted gene regulatory networks increased; directly regulated genes and undirected interactions between regulators and target genes can also be distinguished.

4 Network Analysis to Envisage the Systems Regulation: Inference and Applications

4.1 Inference of System Regulation by Means of Biological Networks

A biological network conveys more information than just the association of the relevant components (Fig. 1). It is the atlas of the cellular regulatory circuit, demonstrating the orchestration of molecular components to modulate the development and homeostasis of cells under an exposed environment. The characteristics of the biological network, including size, constituents, organization, and topology, relate to the nature of the cellular regulation [74, 75], whereas the dynamics of the network alteration infer the adaptive regulatory responses behind the observed phenotypes [76–78]. For decades analytical methods have, thus, been developed not only for investigating the relationship of the involved components but also for attempting to decipher the code of biological regulation inherent in the finding of networks (e.g., [75, 79]). The primary rationale is to learn the *systems regulation* derived from the collaborative actions of the regulatory components. The subsequent comprehensive study is to pinpoint the *key components* and to understand their crucial role in the context of systems regulation under the observed conditions. This comprehension enables us to envisage the landscape of the cellular regulatory systems which finally bring an understanding of the behavioral responses and overt phenotypes of the living organisms. Ultimately, the knowledge of the *cellular regulatory landscape* is conceptually transferred from a well-studied organism to gain more insight into the system of others on the basis of evolutionary conservation. Exploitations of biological networks to unravel the regulation inside the cells have been reported in extensive organisms and diverse aspects, for example (1) identification of a genetic mediator for prostate cancer [80] on which current research is moving toward the elucidation of the disease mechanism through network biology and modeling [81] and (2) investigation of the transcriptional regulation underlying storage root initiation of cassava [82]. Although biological networking is evidently an advantage for investigating the systems regulation of cells, its power is often constrained by the computational techniques and the quality and quantity of data [83, 84]. The precise inference of the biological networks always relies on the availability of data and methods of analysis.

4.1.1 Biological Networks Infer the Collaboration of Regulatory Components Underlying System Regulation

As a means of component-association networking, the regulation is primarily acquired from the predicted relationships between the molecular constituents. The biological network is a great demonstration of the complexity of cellular regulation. It shows that a single regulatory component could not accomplish the whole mechanistic process for cell response. Many studies have employed biological networks to describe the failure in demanding a cell behavior through one-gene attenuation [85, 86]. On the other hand, the biological networking approach complements the measured omics data to emphasize the highly elaborated regulatory mechanism driven by a large number of components. The associated pairs and their relationships comprising the network, which are deduced from the correlation of the measurable expression patterns, are presumed to indicate the mechanism of their cooperative function contributing to the regulatory system. Not only the closely related components but also the network are also obviously useful in discovering the connection between the distant components, enabling us to capture unexpected linkages between regulatory components which help improve the understanding of regulation thoroughly (Fig. 1, bottom-left).

4.1.2 Biological Networks Infer the Potential Key Regulator Modulating the System Regulation

Besides a map of the relationships, the biological network also brings identification of a key regulator (Fig. 1, bottom-middle) and dissection of characteristics of the regulatory system (Fig. 1, bottom-right). The key regulator is basically defined as a regulatory component with a high impact on the overall regulatory system. It can be acquired from the association network through various analytical approaches. Network topology analysis is one of the most used methods, which usually suggests the significance of a component based upon its number of associations (e.g., node degree and network modularity) [48, 49, 87, 88]. It is hypothesized that the important component should be tightly regulated so that it is expected to associate with many neighboring components. Despite a simple principle, these techniques successfully uncovered the predominant regulators in many studies and in broad organisms [87–90]. Finding the key regulators may reveal important components in the network, yet provide limited understanding about characteristics of the systems regulation. Analysis of the network module, such as feed-forward and feedback regulatory motifs, could access the properties of the regulatory system, which is believed to determine the dynamics and efficiency of the systems regulation [75, 91].

4.2 Applications of Biological Network Analysis to Understand the System Regulation: From Unicellular Organisms to Plants

These research practices have been successfully applied to acquire regulation of a range of biological systems and across taxa. Success has been found with prokaryotes, simple eukaryotes, humans, and plants. For example, transcriptional regulatory network (TRN) of *Escherichia coli*, one of the best-characterized prokaryotic organisms, has been constructed from both experiments and computational predictions and deposited in the useful resource database, namely RugulonDB [92]. Yan et al. analyzed the transcriptional regulatory network of *E. coli* and compared it with the computer operating systems in terms of topology and evolution of the regulatory control networks [93]. Based on the basic topology and hierarchical structure of the network, TRN of *E. coli* exhibited a characteristic pyramidal hierarchical layout. Only a few master transcriptional factors (master regulators) were on the top and most transcription factors were at the middle (middle managers), controlling a set of non-TF target genes as a workhorse. The master regulators or middle managers were identified and were useful to target genetic engineering with desired proposes. Not only have single-cell organisms been investigated; complex multi-cell organisms such as plants have also been explored to investigate their gene regulatory networks. *Arabidopsis thaliana* transcriptional regulatory networks under changing environmental conditions were inferred via the reverse-engineering approach [94]. Meta-analyses of several microarray data collections, including several growth conditions, developmental stages, biotic and abiotic stresses, and a variety of mutant genotypes, allowed us to identify regulatory and robust genetic structures. Moreover, TRN of starch metabolism in Arabidopsis were inferred using microarray data under a diurnal cycle and graphical Gaussian model [72]. Starch synthase 4 and its two predicted TFs were further validated with mutant lines. The knockout of two TFs led to the deformation of chloroplast and its contained starch granules. This systematic approach of microarray analysis promised successful discovery of the TRN of starch metabolism in *Arabidopsis* leaves. Besides the model plant, Arabidopsis, gene regulatory networks of important starchy crops, namely cassava, have been investigated [59, 78, 82]. The gene regulatory network of Arabidopsis comprising 11,354 interactions from 67 TFs and their gene targets was applied as the template inferring TRN of the starch metabolism in cassava [59]. *cis*-Regulatory element analysis on each promoter of target genes was verified through PlantPAN database, resulting in a basic-leucine zipper (bZIP) transcription factor family (cassava4.1_017720m.g) controlling two Rubisco genes in cassava (cassava4.1_017170m.g and cassava4.1_017243m.g). The predicted result may shed light into photosynthesis improvement in cassava. In addition to the TRN of metabolic regulation, analysis of the genome-wide microarray experiment revealed the transcription regulation underlying the storage root initiation of cassava, whereby development of fibrous root toward storage root were examined in 8-week-old plants [82]. Based on the reverse engineering approach and *cis*-regulatory element analysis, *KNOX1* gene and phytohormones were proposed as the key regulators to modulate the transition of cassava root toward storage stage. The hypothesis on transcriptional regulation of cassava storage root initiation was validated through hormone treatment experiments. Correspondingly, the exogenous treatment of phytohormones could induce the storage root initiation stage in the in vitro experiment. Furthermore, network inference approaches are capable of acquiring the transcriptional regulation in response to surrounding stresses. Transcriptional regulation underlying the adaptive development of cassava roots in different planting seasons was investigated through the time-series gene expression data, measured by semi-quantitative RT-PCR [78]. Gene expression in storage roots of cassava grown in two different seasons, beginning and end of rainy season (i.e., wet and dry seasons) was analyzed. The gene co-expression networks inferred transcriptional regulation governing the cassava root development under different exposed climates of the planting seasons. As a result, *AP2-EREBP* transcription factor (ERF1) was suggested to play an important role in regulating the cellular responses allowing cassava root development to adapt to wet and dry seasonal climates.

5 Conclusions

As the cellular processes are very complicated, regulation contributes to such complexity. Biological network analysis is considered an approach for exploring the mesh of regulatory elements association in cells and to reflect on how they work together. By virtue of advanced high-throughput sequencing technology and methods for constructing and analyzing available biological networks, the systems biology approach has been utilized for network investigation, leading to gene-to-phenotype mapping, in a wide range of organisms. These network analyses enhance our understanding of cellular regulation and could open the way for various applications.

References

1. Yu CP, Chen SC, Chang YM, Liu WY, Lin HH, Lin JJ, Chen HJ, Lu YJ, Wu YH, Lu MY, Lu CH, Shih AC, Ku MS, Shiu SH, Wu SH, Li WH (2015) Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. Proc Natl Acad Sci U S A 112(19)
2. Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422:835–847
3. Mardis ER (2006) Anticipating the $1,000 genome. Genome Biol 7:112
4. Service RF (2006) Gene sequencing. The race for the $1000 genome. Science 311:1544–1546
5. Wolinsky H (2007) The thousand-dollar genome. Genetic brinkmanship or personalized medicine? EMBO Rep 8:900–903
6. Martin LB, Fei Z, Giovannoni JJ, Rose JK (2013) Catalyzing plant science research with RNA-seq. Front Plant Sci 4:66
7. Wang L, Cao C, Ma Q, Zeng Q, Wang H, Cheng Z, Zhu G, Qi J, Ma H, Nian H, Wang Y (2014) RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications. BMC Plant Biol 14:169
8. Li C, Qiao Z, Qi W, Wang Q, Yuan Y, Yang X, Tang Y, Mei B, Lv Y, Zhao H, Xiao H, Song R (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. Plant Cell 27(3):532–545

9. O'Geen H, Henry IM, Bhakta MS, Meckler JF, Segal DJ (2015) A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Res 43 (6):3389–3404
10. Liu L, Li Y, Li S, Hu N, He Y, Pong R et al (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364
11. Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem 6:287–303
12. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Mol Ecol 21:1794–1805
13. Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics = Yi chuan xue bao 38(3):95–109
14. Huang YF, Chen SC, Chiang YS, Chen TH, Chiu KP (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6(Suppl 2):S10
15. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR et al (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341
16. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D et al (2014) Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One 9, e106689
17. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352
18. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T et al (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26:1146–1153
19. Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240
20. Schadt EE (2012) The changing privacy landscape in the era of big data. Mol Syst Biol 8:612
21. Timp W, Mirsaidov UM, Wang D, Comer J, Aksimentiev A, Timp G (2010) Nanopore sequencing: electrical measurements of the code of life. IEEE Trans Nanotechnol 9:281–294
22. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13:375
23. Koren S, Schatz MC, Walenz BP, Martin J, Howard J, Ganapathy G et al (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30:693–700
24. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D et al (2013) Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res 23:121–128
25. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10
26. Ferrarini M, Moretto M, Ward JA, Surbanovski N, Stevanovic V, Giongo L et al (2013) An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics 14:670
27. Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N et al (2014) Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. BMC Genomics 15:699
28. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33:623–630
29. Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110–120

30. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M (2012) Automated forward and reverse ratcheting of DNA in a nanopore at five angstrom precision. Nat Biotechnol 30:344–348
31. Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A et al (2015) Genome assembly using nanopore-guided long and error-free DNA reads. BMC Genomics 16:327
32. Wang H, Ayala JC, Benitez JA, Silva AJ (2015) RNA-seq analysis identifies new genes regulated by the histone-like nucleoid structuring protein (H-NS) affecting Vibrio cholerae virulence, stress response and chemotaxis. PLoS One 10, e0118295
33. Bassim S, Tanguy A, Genard B, Moraga D, Tremblay R (2014) Identification of Mytilus edulis genetic regulators during early development. Gene 551:65–78
34. Chen J, Quan M, Zhang D (2015) Genome-wide identification of novel long non-coding RNAs in Populus tomentosa tension wood, opposite wood and normal wood xylem by RNA-seq. Planta 241:125–143
35. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
36. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C et al (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100
37. Tang B, Hsu HK, Hsu PY, Bonneville R, Chen SS, Huang THM et al (2012) Hierarchical modularity in ER transcriptional network is associated with distinct functions and implicates clinical outcomes. Sci Rep 2:875
38. Buerstmayr M, Buerstmayr H (2015) Comparative mapping of quantitative trait loci for Fusarium head blight resistance and anther retention in the winter wheat population Capo × Arina. Theor Appl Genet 128:1519–1530
39. Duan G, Christian N, Schwachtje J, Walther D, Ebenhöh O (2013) The metabolic interplay between plants and phytopathogens. Metabolites 3:1–23
40. Owolade OF, Dixon AGO, Adeoti AYA (2006) Diallel analysis of cassava genotypes to anthracnose disease. World J Agric Sci 2:98–104
41. Vanderschuren H, Moreno I, Anjanappa RB, Zainuddin IM, Gruissem W (2012) Exploiting the combination of natural and genetically engineered resistance to cassava mosaic and cassava brown streak viruses impacting cassava production in Africa. PLoS One 7, e45277
42. Kim D, Song JY, Hahn JS (2015) Improvement of glucose uptake rate and production of target chemicals by overexpressing hexose transporters and transcriptional activator Gcr1 in Saccharomyces cerevisiae. Appl Environ Microbiol 81:8392–8401
43. Kurylenko OO, Ruchala J, Hryniv OB, Abbas CA, Dmytruk KV, Sibirny AA (2014) Metabolic engineering and classical selection of the methylotrophic thermotolerant yeast Hansenula polymorpha for improvement of high-temperature xylose alcoholic fermentation. Microb Cell Fact 13
44. Barabási A, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
45. Lucas M, Lapalaze L, Bennett MJ (2011) Plant systems biology: network matters. Plant Cell Environ 34:535–553
46. Chiang AW, Liu W, Charusanti P, Hwang M (2014) Understanding system dynamics of an adaptive enzyme network from globally profiled kinetic parameters. BMC Syst Biol 8:4
47. Dharmawardhana P, Ren L, Amarasinghe V, Monaco M, Thomason J, Ravenscroft D, McCouch S, Ware D, Jaiswal P (2013) A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice 6:15
48. Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J Mol Biol 360:213–227
49. Balaji S, Iyer LM, Aravind L, Babu M (2006) Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. J Mol Biol 360(1):204–212

50. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298:799–804
51. Ji C, Cao X, Yao C, Xue S, Xiu Z (2014) Protein–protein interaction network of the marine microalga Tetraselmis subcordiformis: prediction and application for starch metabolism analysis. J Ind Microbiol Biotechnol 41(8):1287–1296
52. Pirkl M, Hand E, Kube D, Spang R (2015) Analyzing synergistic and non-synergistic interactions in signalling pathways using Boolean nested effect models. Bioinformatics 32 (6):893–900
53. Wang J, Zhang S, Wang Y, Chen L, Zhang XS (2009) Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. PLoS Comput Biol 5:e1000521
54. Batagelj V, Mrvar A (1998) Pajek: a program for large network analysis. Connections 21:47–57
55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
56. Berger SI, Posner JM, Ma'ayan A (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. BMC Bioinf 8:372. doi:10.1186/1471-2105-1188-1372
57. Wernicke S, Rasche F (2006) FANMOD: a tool for fast network motif detection. Bioinformatics 22:1152–1153
58. Babu MM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358:614–633
59. Khampoosa B, Bumee S, Saithong T, Suksangpanomrung M, Kalapanulak S (2014) Construction of transcriptional regulatory network proposes bZIP transcription factor controlling Rubisco genes in cassava. Paper presented at the 26th annual meeting of the Thai Society for Biotechnology and international conference, Thailand
60. Lozada-Chavez I, Janga SC, Collado-Vides J (2006) Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res 34:3434–3445
61. Katara P, Grover A, Sharma V (2012) Phylogenetic footprinting: a boost for microbial regulatory genomics. Protoplasma 249:901–907
62. Alkema W, Lenhard B, Wasserman WW (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. Genome Res 14:1362–1373
63. Liu J, Xu X, Stormo GD (2008) The cis-regulatory map of Shewanella genomes. Nucleic Acids Res 36:5376–5390
64. Sosinsky A, Honig B, Mann RS, Califano A (2007) Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci U S A 104:6305–6310
65. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37:382–390
66. Rodenas-Cuadrado P, Chen XS, Wiegrebe L, Firzlaff U, Vernes SC (2015) A novel approach identifies the first transcriptome networks in bats: a new genetic model for vocal communication. BMC Genomics 16:836
67. Babu MM, Lang B, Aravind L (2009) Methods to reconstruct and compare transcriptional regulatory networks. Methods Mol Biol 541:163–180
68. Hu H, Li X (2010) Transcription factor binding site identification by phylogenetic footprinting frontiers in computational and systems biology, vol 15. Springer, London, pp 113–131
69. Wang T, Stormo GD (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci U S A 102:17400–17405
70. Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301:102–105

71. Wirojsirasak W, Saithong T, Sojikul P, Hirunsirisawat P, Kalapanulak S (2014) The effect of microarray data resolution on the inferred transcriptional regulatory network topology. Paper presented at the 2nd ASEAN plus three graduate research congress, Thailand
72. Ingkasuwan P, Netrphan S, Prasitwattanaseree S, Tanticharoen M, Bhumiratana S, Meechai A et al (2012) Inferring transcriptional gene regulation network of starch metabolism in Arabidopsis thaliana leaves using graphical Gaussian model. BMC Syst Biol 6:100. doi:10.1186/1752-0509-1186-1100
73. Xing B, van der Laan MJ (2005) A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. J Comput Biol 12:229–246
74. Xiaowei Z, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21:1010–1024
75. Zhiyuan L, Bianco S, Zhang Z, Tang C (2014) Generic properties of random gene regulatory networks. Quant Biol 1:253–260
76. Gitter A, Carmi M, Barkai N, Bar-Joseph Z (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. Genome Res 23:365–376
77. O'Neill PR, Giri L, Karunarathne WKA, Patel AK, Venkatesh KV, Gautam N (2014) The structure of dynamic GPCR signaling networks. Wiley interdisciplinary reviews. Syst Biol Med 6:115–123
78. Saithong T, Saerue S, Kalapanulak S, Sojikul P, Narangajavana J, Bhumiratana S (2015) Gene co-expression analysis inferring the crosstalk of ethylene and gibberellin in modulating the transcriptional acclimation of cassava root growth in different seasons. PLoS One 10, e0137602
79. Song F, Ollivier JF, Swain PS, Soyer OS (2015) BioJazz: in silico evolution of cellular networks with unbounded complexity using rule-based modeling. Nucleic Acids Res 43(19): e123. doi:10.1093/nar/gkv595
80. Ergün A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ (2007) A network biology approach to prostate cancer. Mol Syst Biol 3:82
81. Creixell P, Schoof EM, Erler JT, Linding R (2012) Navigating cancer network attractors for tumor-specific therapy. Nat Biotechol 30:842–848
82. Sojikul P, Saithong T, Kalapanulak S, Pisuttinusart N, Limsirichaikul S, Tanaka M, Utsumi Y, Sakurai T, Seki M, Narangajavana J (2015) Genome-wide analysis reveals phytohormone action during cassava storage root initiation. Plant Mol Biol 88:531–543
83. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8:717–729
84. Saithong T, Bumee S, Liamwirat C, Meechai A (2012) Analysis and practical guideline of constraint-based Boolean method in genetic network inference. PLoS One 7, e30232
85. Nakamichi N, Kita M, Ito S, Yamashino T, Mizuno T (2005) PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, together play essential roles close to the circadian clock of Arabidopsis thaliana. Plant Cell Physiol 46:686–698
86. Su SH, Suarez-Rodriguez MC, Krysan P (2007) Genetic interaction and phenotypic analysis of the Arabidopsis MAP kinase pathway mutations mekk1 and mpk4 suggests signaling pathway complexity. FEBS Lett 581:3171–3177
87. Chand Y, Alam MA (2012) Network biology approach for identifying key regulatory genes by expression based study of breast cancer. Bioinformation 8:1132–1138
88. Gargouri M, Park JJ, Holguin FO, Kim MJ, Wang H, Deshpande RR et al (2015) Identification of regulatory network hubs that control lipid metabolism in Chlamydomonas reinhardtii. J Exp Bot 66(15):4551–4566
89. Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M (2006) Target hub proteins serve as master regulators of development in yeast. Genes Dev 20:435–448
90. Zhang L, Yu S, Zuo K, Luo L, Tang K (2012) Identification of gene modules associated with drought response in rice by network-based analysis. PLoS One 7(5), e33748
91. Navlakha S, He X, Faloutsos C, Bar-Joseph Z (2014) Topological properties of robust biological and computational networks. J R Soc Interface 11(96):20140283

92. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, Garcia-Sotelo JS et al (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res 44(D1):D133–D143. doi:10.1093/nar/gkv1156

93. Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M (2010) Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. Proc Natl Acad Sci U S A 107:9186–9191

94. Carrera J, Rodrigo G, Jaramillo A, Elena SF (2009) Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions. Genome Biol 10 (9):R96. doi:10.1186/gb-2009-1110-1189-r1196

Adv Biochem Eng Biotechnol (2017) 160: 143–184
DOI: 10.1007/10_2016_46

Published online: 10 January 2017

Network Metamodeling: Effect of Correlation Metric Choice on Phylogenomic and Transcriptomic Network Topology

Deborah A. Weighill and Daniel Jacobson

Abstract We explore the use of a network meta-modeling approach to compare the effects of similarity metrics used to construct biological networks on the topology of the resulting networks. This work reviews various similarity metrics

This work was supported by the Plant-Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) in the Genomic Science Program, the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science, and the BERs BioEnergy Science Center (BESC) at the Oak Ridge National Laboratory (contract DE-PS02-06ER64304). Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the chapter for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy provides public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). The authors would also like to acknowledge the Centre for High Performance Computing and the Stellenbosch High Performance Computing Cluster for computing resources, and the South African National Research Foundation (www.nrf.ac.za) Technology and Human Resources Programme and Winetech. The financial assistance of the National Research Foundation (NRF) toward this research is hereby acknowledged. Opinions expressed and conclusions reached are those of the author and are not necessarily to be attributed to the NRF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

D.A. Weighill and D. Jacobson (✉)
Faculty of AgriSciences, Institute for Wine Biotechnology, Stellenbosch University, JH Neethling Building, Victoria Street, 7600 Stellenbosch, South Africa

The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, 444 Greve Hall, 821 Volunteer Blvd., Knoxville, TN 37996-3394, USA

Biosciences Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831, USA
e-mail: jacobsonda@ornl.gov

for the construction of networks and various topology measures for the characterization of resulting network topology, demonstrating the use of these metrics in the construction and comparison of phylogenomic and transcriptomic networks.

Keywords Network comparison, Network topology, Similarity metrics

Contents

1 Introduction 144
2 Similarity Metrics 145
2.1 Overview 145
2.2 Pearson Correlation Coefficient 145
2.3 Spearman Correlation Coefficient 146
2.4 Jaccard’s Index 146
2.5 Cosine Similarity 147
2.6 Sørensen Index 147
2.7 Czekanowski Index and Bray–Curtis Index 148
2.8 Canberra Distance 149
2.9 Jaccardized Czekanowski Index 150
2.10 Maximum Information Coefficient 151
3 Network Topology Measures 151
3.1 Node-Based Topology Measures 152
3.2 Global Network Topology Measures 154
4 Network Comparison and Network Overlap 156
4.1 Clustering Comparison 156
4.2 Network Profile Comparison 158
5 Similarity Metric Effect on Network Topology: Results and Discussion 158
5.1 Overview 158
5.2 Metric Comparison Though Network Topology Profiles and Clustering Comparison 158
5.3 Metric Comparison Through Network Merging and Cross-Network Topological Overlap 164
6 Conclusions 174
7 Similarity Metric Effect on Network Topology: Methods 177
7.1 Metric Comparison Though Network Topology Profiles and Clustering Comparison 177
7.2 Metric Comparison Through Network Merging and Cross-Network Topological Overlap 180
References 182

1 Introduction

Meta-modeling involves creating models of models to compare the outcomes of a model when different parameters are used. Network models involve modeling the similarity between pairs of objects of interest [1]. A parameter of such a network model could be the similarity metric chosen to quantify the similarity between nodes in order to weight the edges. Many similarity metrics exist, and were developed to quantify different aspects of similarity. Thus, using different similarity metrics to construct a network model should result in different results and thus affect the end biological interpretation.

This chapter focuses on network meta-modeling, exploring a selection of approaches for the comparison of networks. In particular, network models of particular datasets constructed using different similarity metrics are compared to investigate the effect the choice of similarity metric has on the resulting network topology.

2 Similarity Metrics

2.1 Overview

Networks are often constructed to represent the similarities and relationships between objects within biological systems. Objects are often represented as a vector of quantities. For example, when constructing gene co-expression networks, objects (genes) are represented by expression profiles. Networks are thus often constructed by performing an all-vs-all comparison of a set of objects of interest by calculating the similarity between all pairs of vectors representing the objects. To achieve this, similarity metrics are needed to provide a measure of similarity between two vectors. Various similarity metrics exist which all quantify different aspects of similarity.

2.2 Pearson Correlation Coefficient

Pearson's Correlation Coefficient was first introduced by Karl Pearson in 1895 [2] and is a widely used correlation metric. Pearson's correlation coefficient r between two variables X and Y can be expressed as

$$r = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i \left(X_i - \overline{X}\right)^2 \sum_I \left(Y_i - \overline{Y}\right)^2}} \tag{1}$$

where $\overline{X}$ and $\overline{Y}$ are the means of variables X and Y, respectively. Pearson's correlation coefficient takes on values between -1 and 1 and measures the linear association between two vectors [3]. Equation (1) can be expressed in an alternative form giving Pearson's correlation coefficient of vectors X and Y in terms of the covariance of the two vectors, scaled by their standard deviations (2):

$$r = \frac{\mathrm{Cov}\,(X,Y)}{S_X S_Y} \tag{2}$$

where $\mathrm{Cov}(X,Y)$ is the covariance of X and Y and S_X and S_Y are the standard deviations of X and Y, respectively [3].

2.3 Spearman Correlation Coefficient

The Spearman Correlation Coefficient [4] r_s for variables X and Y has a formula similar to the Pearson Correlation Coefficient except that, instead of using the actual values of the entries in the vectors, the ranks of the entries in the vectors are used. For vectors X and Y, let R_i denote the rank of value i in X, and let Q_i denote the rank of value i in Y. The Spearman Correlation Coefficient is then given by

$$r_s = \frac{\sum_i \left(R_i - \overline{R}\right)\left(Q_i - \overline{Q}\right)}{\sqrt{\sum_i \left(R_i - \overline{R}\right)^2 \sum_i \left(Q_i - \overline{Q}\right)^2}} \tag{3}$$

where $\overline{R}$ and $\overline{Q}$ are the means of rank variables R and Q, respectively [5]. The Spearman Correlation Coefficient measures the monotonicity of two vectors, that is, to what extent the values in the vector increase as the values in the other vector increase. Unlike the Pearson Correlation Coefficient, it does not measure the extent of a linear relationship between the two vectors [5].

2.4 Jaccard's Index

Jaccard's Index is a similarity index which was originally referred to as the "Coefficient of Community" [6]. It was developed to quantify the similarity between the plant species content of two areas. It is easily defined in terms of set intersects. Given two sets A and B, Jaccard's Index $J(A,B)$ is defined as [6]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

Jaccard's Index can also be defined in terms of vectors. Let the two sets be two binary vectors, X and Y. Jaccard's Index $J(X,Y)$ can then be defined in terms of inner products as [7]

$$J(X,Y) = \frac{\langle X,Y\rangle}{\langle X,X\rangle + \langle Y,Y\rangle - \langle X,Y\rangle} \tag{5}$$

Jaccard's Index takes on values between 0 and 1.

To apply Jaccard's Index to non-binary vectors, a vector X of integers can easily be converted to a binary vector X_B as follows:

$$X_{Bi} = \begin{cases} 1 & \text{if } X_i \geq 1 \\ 0 & \text{if } X_i = 1 \end{cases} \tag{6}$$

2.5 Cosine Similarity

The Cosine similarity of two vectors X and Y simply involves taking the cosine of the angle between the two vectors (7):

$$\text{Cosine Similarity} = \cos(\Theta_{XY}) \tag{7}$$

where Θ_{XY} is the angle between vectors X and Y. This equation can also be written in inner-product form, in which the cosine of the angle between two vectors is expressed in terms of the inner product of the vectors, divided by their norms [8] (8):

$$\cos(\Theta_{XY}) = \frac{\langle X,Y\rangle}{||X||\,||Y||} \tag{8}$$

Cosine similarity takes on values between 0 and 1 [8], assuming that both vectors contain only positive values. This is the case with most biological data.

2.6 Sørensen Index

The Sørensen Index [9] (also known as the Dice Coefficient [10]) is a similarity index developed for ecological purposes and (similar to Jaccard's Index) is also based on set intersections. For two sets A and B the Sørensen Index $S(A,B)$ is defined as

$$S(A,B) = \frac{|A \cap B|}{|A| + |B|} \tag{9}$$

where $\left|A\right|$ is the number of elements in A and $\left|B\right|$ is the number of elements in B. The Sørensen Index can also be formulated in terms of vector algebra. For two binary vectors X and Y the Sørensen Index $S(X,Y)$ is defined as

$$S(X,Y) = \frac{2\langle X,Y\rangle}{\sum_i x_i + \sum_i y_i} \tag{10}$$

$$= \frac{2\min(X,Y)}{\sum_i x_i + \sum_i y_i} \tag{11}$$

where x_i is the ith element of X and y_i is the ith element of Y.

2.7 Czekanowski Index and Bray–Curtis Index

The Czekanowski Index is a quantitative version of the Sørensen Index. For vectors X and Y the Czekanowski Index is defined as [11]

$$Cz = \frac{\sum_i 2\min(X_i,Y_i)}{\sum_i (X_i + Y_i)} \tag{12}$$

where X_i is the ith element of X and Y_i is the ith element of Y. The similarities between the forms of (11) and (12) are easy to see, indicating the relationship between the Czekanowski Index and the Sørensen Index.

The Bray–Curtis [12] Index is often confused with the Czekanowski Index [11]. Although the Bray–Curtis Index has the same form as the Czekanowski Index (12), the underlying normalization assumptions are different. The Bray–Curtis Index assumes that all vectors are normalized by the total sum of each vector, that is, the sum of all the entries in a vector is 1. Thus the Bray–Curtis Index $BC(X,Y)$ simplifies to [11, 12]

$$BC(X,Y) = \frac{\sum_i 2\min(X_i,Y_i)}{\sum_i (X_i + Y_i)} \tag{13}$$

$$= \frac{2\sum_i \min(X_i,Y_i)}{\sum_i X_i + \sum_i Y_i} \tag{14}$$

$$= \frac{2\sum_i \min(X_i,Y_i)}{1+1} \tag{15}$$

$$= \frac{2\sum_i \min(X_i, Y_i)}{2} \tag{16}$$

$$= \sum_i \min(X_i, Y_i) \tag{17}$$

2.8 Canberra Distance

The Canberra distance $Cb(X, Y)$ is a distance metric described as being the complement of Czekanowski's Index, and defined as [13]

$$Cb(X, Y) = \frac{\sum_i |X_i - Y_i|}{\sum_i (X_i + Y_i)} \tag{18}$$

Thus, the Canberra distance can be defined as

$$Cb(X, Y) = 1 - Cz(X, Y) \tag{19}$$

or, equivalently,

$$Cz(X, Y) = 1 - Cb(X, Y) \tag{20}$$

This can be derived as follows:

$$1 - Cb(X, Y) = 1 - \frac{\sum_i |X_i - Y_i|}{\sum_i (X_i + Y_i)} \tag{21}$$

$$= \frac{\sum_i (X_i + Y_i) - \sum_i |X_i - Y_i|}{\sum_i (X_i + Y_i)} \tag{22}$$

$$= \frac{\sum_i \big((X_i + Y_i) - |X_i - Y_i|\big)}{\sum_i (X_i + Y_i)} \tag{23}$$

It should be noted that (23) has the same denominator as the Czekanowski Index in (12). Thus, to show that $1 - Cb(X, Y) = Cz(X, Y)$, we need to show that the numerators of (23) and (12) are equal. To do this, consider the diagram in Fig. 1. Assume that for a given i, $X_i > Y_i$. Then

$$\sum_i \big((X_i + Y_i) - |X_i - Y_i|\big) = \sum_i 2Y_i. \tag{24}$$

Similarly, if for a given i, $Y_i > X_i$, then

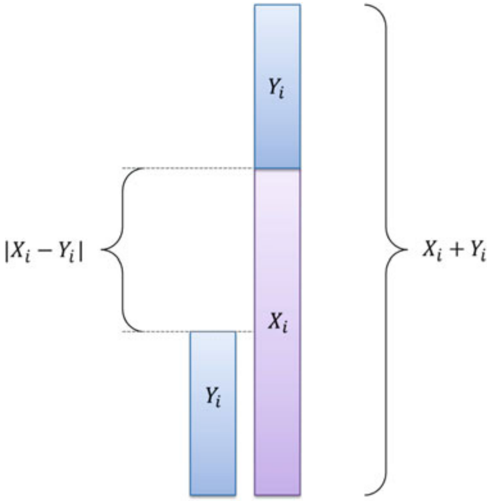


Fig. 1 Canberra distance vs Czekanowski similarity. A visual aid in the relatedness of the Czekanowski similarity index and the Canberra distance

$$\sum_i \left((X_i + Y_i) - \left|X_i - Y_i\right|\right) = \sum_i 2X_i. \tag{25}$$

Thus, combining the above two cases,

$$\sum_i \left((X_i + Y_i) - \left|X_i - Y_i\right|\right) = \sum_i 2\min(X_i, Y_i), \tag{26}$$

which is indeed the numerator of (12). Thus, the Czekanowski Index $Cz(X, Y)$ is the complement of the Canberra distance $Cb(X, Y)$ related as $1 - Cb(X, Y) = Cz(X, Y)$.

2.9 Jaccardized Czekanowski Index

The Jaccardized Czekanowski Index [14] is a new similarity metric which attempts to formulate a quantitative version of Jaccard's Index in the same sense that the Czekanowski Index is a quantitative version of the Sørensen Index. The Jaccardized Czekanowski Index is derived as follows [14]. First, the Jaccard Index J is related to the Sørensen Index S by the following equation:

$$S = \frac{2J}{J+1} \tag{27}$$

Rearranging (27) to make J the subject of the equation yields:

$$J = \frac{S}{2 - S} \tag{28}$$

Replacing the Sørensen Index S in (28) with the Czekanowski Index Cz thus yields a quantitative version of Jaccard's Index called the Jaccardized Czekanowski Index:

$$JCz = \frac{Cz}{2 - Cz} \tag{29}$$

The Jaccardized Czekanowski Index was then found not to be novel but actually the same as the Ružička Index developed in 1958 [15].

2.10 Maximum Information Coefficient

The Maximum Information Coefficient (MIC) between two vectors X and Y is a similarity metric which, unlike the Pearson Correlation Coefficient, can detect nonlinear correlations. The MIC is calculated as follows. Consider a set of ordered pairs (x_i, y_i) where x_i is the ith value in X and y_i is the ith value in Y. A partition is then created on the ordered pairs (x_i, y_i). This can be visualized as plotting a scatterplot of X vs Y, drawing a grid $m \times n$ on this scatter plot, and partitioning the points ((x_i, y_i) pairs) into blocks. Grids of different dimensions are drawn. Each grid results in a characteristic probability distribution of each variable, allowing the Mutual Information of the variables to be created. The Maximum Information Coefficient is the maximum Mutual Information Coefficient obtained across all grids of all dimensions considered [16, 17].

3 Network Topology Measures

Once networks have been constructed for a certain set of objects of interest within a system using a particular similarity metric and have been pruned to select for the most highly weighted edges, the networks exhibit certain topologies. Network topology can be described quantitatively through a number of network properties or network measures [18]. These measures quantify local properties of individual nodes within a network and topological properties of the entire network as a whole.

3.1 Node-Based Topology Measures

The following network measures are defined per node or per node pair for a given network, and include adjacency, connectivity, maximum adjacency ratio, topological overlap, TOM-connectivity, clustering coefficient, betweenness, and efficiency, as defined below.

3.1.1 Adjacency

For two nodes i and j, the adjacency a_{ij} is the entry ij in the adjacency matrix of the network. In an unweighted network, a_{ij} is 1 if nodes i and j are connected by an edge, and 0 otherwise. In a weighted network, a_{ij} is equal to the strength of the connection (i.e., the edge weight) between nodes i and j [19].

3.1.2 Connectivity

The connectivity k_i for a node i is defined as [18]

$$k_i = \sum_{j \neq i} a_{ij} \tag{30}$$

where a_{ij} is the adjacency of nodes i and j. It is an indication of how well-connected a node is to the network. For an unweighted network, the connectivity k_i of node i is the number of edges connected to node i, that is, the degree of the node. For a weighted network, the connectivity of node i it is the sum of the weights of the edges connected to node i.

3.1.3 Maximum Adjacency Ratio

The Maximum Adjacency Ratio (MAR$_i$) for a node i is an extension of the connectivity of a node and is defined as [18]

$$\mathrm{MAR}_i = \frac{\sum_{j \neq i} \left(a_{ij}\right)^2}{\sum_{j \neq i} a_{ij}} \tag{31}$$

MAR describes the extent to which a node has strong connections with its neighbors. Assuming that the network edges have weights between 0 and 1, the Maximum Adjacency Ratio obtains a maximum value of 1 when all the connections of a node have the maximum weight of 1 [18].

3.1.4 Topological Overlap

The Topological Overlap ω_{ij} between two nodes i and j quantifies how connected the two nodes are by taking into consideration the direct connection between the nodes and indirect connection via neighbors of the nodes [20], and is defined as [19]

$$\omega_{ij} = \frac{\left(\sum_u a_{iu}a_{uj}\right) + a_{ij}}{\min\left(k_i, k_j\right) + 1 - a_{ij}} \tag{32}$$

where a_{iu} and a_{uj} are the adjacencies and k_i and k_j are the connectivities of nodes i and j, respectively [19]. In an unweighted network, the term $\sum_u a_{iu}a_{uj}$ is equal to the number of neighbors shared between nodes i and j. Consider two nodes i and j with $k_i < k_j$. For an unweighted network, the topological overlap ω_{ij} is equal to 1 if every neighbor of i is also a neighbor of j and if a_{ij} is equal to 1. Put simply, this means that for the topological overlap between two nodes i and j to be 1, all neighbors of the node with smaller degree need to be neighbors of the node with larger degree, and the nodes i and j need to be directly connected. For the topological overlap to be zero, the nodes must not be connected and they must have no common neighbors [19].

3.1.5 TOM-Based Connectivity

The TOM-connectivity of a node is based on the topological overlap between nodes and is defined as [19]

$$k_i = \sum_{j \neq i} \omega_{ij} \tag{33}$$

where ω_{ij} is the topological overlap (32) between nodes i and j. A node thus has a high TOM-connectivity if it has a high topological overlap with its neighbors, that is, a node is connected to and shares a lot of neighbors with its neighbors [19].

3.1.6 Clustering Coefficient

The Clustering Coefficient [21] for a node is a measure indicating the local structure around the node, in particular how densely connected (cliquish) the node and its neighbors are [18]. For an unweighted network, the Clustering Coefficient C_i for a node i is defined as the number of edges present in the neighborhood around node i over the total possible number of edges in that neighborhood:

$$C_i = \frac{\sum_{l \neq i} \sum_{m \neq i, l} a_{il} a_{lm} a_{mi}}{k_i (k_i - 1)} \tag{34}$$

The Clustering Coefficient reaches its maximum value when each pair of a node's neighbors are connected to each other [18]. Zhang et al. [19] extended the Clustering Coefficient to apply to weighted networks:

$$C_i = \frac{\sum_{l \neq i} \sum_{m \neq i, l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il}\right)^2 \sum_{l \neq i} (a_{il})^2} \tag{35}$$

3.1.7 Betweenness

The Betweenness of a node i is the number of shortest paths between other pairs of nodes which run through node i [22]. This measure could indicate the importance of the node and how much it would affect the network should it be removed [22].

3.1.8 Efficiency

The Efficiency E_{ij} of a path between two nodes i and j is calculated as the inverse of the length of the shortest path between two nodes [23]:

$$E_{ij} = \frac{1}{d_{ij}} \tag{36}$$

where d_{ij} is the length of the shortest path between nodes i and j. The shorter the path between two nodes, the more efficient the path. If no path between nodes i and j exists in the graph, the distance d_{ij} between nodes i and j is defined to be $d_{ij} = \infty$ and thus the efficiency $E_{ij} = 0$ [23].

3.2 Global Network Topology Measures

The following network measures are global network measures calculated for a network as a whole and not on an individual node or node pair level, and include density, centralization, heterogeneity, path length, and degree correlation.

3.2.1 Network Density

The Density D of a network is a quantification of how densely connected the network is. For an unweighted network, Network Density is defined as the fraction of the number of edges in the network divided by the total number of possible edges given the number of nodes [24]:

$$D = \frac{s}{n(n-1)} \tag{37}$$

where s is the number of edges in the network and n is the number of nodes in the network. Network density can easily be extended for weighted networks and can be calculated as the mean of all the off-diagonal entries in the adjacency matrix [25]:

$$D = \frac{\sum_i k_i}{n(n-1)} \tag{38}$$

$$= \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} \tag{39}$$

where k_i is the connectivity of node i and a_{ij} is the entry ij in the adjacency matrix of the network.

3.2.2 Network Centralization

Network Centralization measures the extent to which there is a point in the network more central than all other points [26]. It has a maximum value of 1 when the network has a star topology (very centralized) and 0 if the connectivity of each node in the network is the same, for example, a square [18]. The Centralization C of a network is defined as [25]

$$C = \frac{n}{n-2}\left(\frac{k_{\max}}{n-1} - D_N\right) \tag{40}$$

where n is the number of nodes in the network, $k_{\max}$ is the maximum connectivity of the network, and D_N is the network density.

3.2.3 Network Heterogeneity

Network heterogeneity H quantifies how much the connectivity of the nodes in the network varies throughout the network in terms of the variance of the connectivities [24] and is defined as [25]

$$H = \frac{\sqrt{\mathrm{var}\,(k)}}{\mathrm{mean}\,(k)} \tag{41}$$

where var(k) is the variance in the connectivity of the network and mean(k) is the mean connectivity of the network. A very heterogeneous network has a large variation in the connectivities of the nodes whereas in a homogeneous network, connectivity is evenly distributed throughout the network.

3.2.4 Path Length

The Path Length of a network is the average length of all shortest paths between pairs of vertices [21].

3.2.5 Degree Correlation

The Degree Correlation quantifies how correlated the degrees of neighboring nodes are. Assortative networks arise if nodes of high degree are mostly connected to other nodes of high degree, whereas disassortative networks arise when nodes of high degree are mostly connected to nodes of low degree [22].

4 Network Comparison and Network Overlap

Once networks have been created, various methods exist to compare them to each other, based on how they cluster or on their topological characterization.

4.1 Clustering Comparison

A clustering C is a partition of a set of objects consisting of non-overlapping sets of objects [27]. These sets are called clusters which can be generated by clustering algorithms such as MCL [28, 29]. Sets of clusters can then be compared using clustering overlap metrics.

4.1.1 Jaccard Overlap

Several clustering overlap metrics are based on counting pairs of elements and how often pairs of elements fall in the same cluster or in different clusters [30]. An example of a pair counting metric is the Jaccard index for clustering overlaps. The Jaccard overlap between two clusterings C_i and C_j is calculated as [30, 31]

$$J\left(C_i, C_j\right) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{42}$$

where N_{11} is the number of pairs of elements (x, y) which are in the same cluster in C_i and C_j, N_{10} is the number of pairs of elements (x, y) which are in the same cluster in C_i but not C_j, and N_{01} is the number of pairs of elements (x, y) which are in the same cluster in C_j but not C_i.

4.1.2 Mutual Information

Other clustering overlap measures include those based on mutual information. These measures quantify the extent to which information about one clustering provides information about another clustering [31]. It is derived from the entropies of two clusterings as follows.

Let S denote the sample space of n objects. Let C_i and C_j denote clusterings of S. The Normalized Mutual Information between two clusterings C_i and C_j is defined as [30]

$$NMI\left(C_i, C_j\right) = \frac{I\left(C_i, C_j\right)}{\sqrt{H(C_i)H\left(C_j\right)}} \tag{43}$$

where $I(C_i, C_j)$ is the Mutual Information between clusterings C_i and C_j, $H(C_i)$ is the entropy of C_i, and $H(C_j)$ is the entropy of C_j. The Mutual Information between two clusterings $I(C_i, C_j)$ and the entropies $H(C_i)$ and $H(C_j)$ are defined as [31]:

$$I\left(C_i, C_j\right) = \sum_a \sum_b P(a, b) \log_2 \left(\frac{P(a, b)}{P(a)P(b)} \right) \tag{44}$$

$$H(C_i) = -\sum_a P(a) \log_2 (P(a)) \tag{45}$$

$$H\left(C_j\right) = -\sum_b P(b) \log_2 (P(b)) \tag{46}$$

where a is a cluster in clustering C_i and b is a cluster in clustering C_j. $P(a)$ is defined as $\frac{|a|}{n}$, $P(b)$ is defined as $\frac{|b|}{n}$, and $P(i, j)$ is defined as $\frac{|a \cap b|}{n}$.

Entropy ((45) and (46)) is a measure of the amount of uncertainty present in a clustering. This is best understood by the following thought experiment. Consider a clustering of n points and consider picking an arbitrary point from any cluster. Assuming each point has an equal chance of being picked, the probability of the point being in cluster k of size n_k is $\frac{n_k}{n}$ [31]. If there is only one cluster in the clustering, then $\frac{n_k}{n} = 1$, causing the entropy (uncertainty) to be zero ((45) and (46)). Thus, if there is only one cluster, there is no uncertainty/information present in the clustering. However, if the clustering contains more clusters with a more non-trivial probability distribution, the entropy (and information present in the clustering) increases. Mutual information is then derived from entropy, calculated as the information shared between two clusterings.

4.2 Network Profile Comparison

Another approach for comparing networks is implemented in a method called NetSimile [32]. This approach compares networks based on their topologies. For a set of networks to be compared, a selection of network topology measures are calculated for each network. These measures are compiled into a signature vector for each network. Network comparison then simply reduces to calculating the Canberra distance between the network's signature topology vectors [32].

5 Similarity Metric Effect on Network Topology: Results and Discussion

5.1 Overview

Two types of datasets were used for the exploration of network comparison approaches. The first dataset on which Clustering and Network topology profile comparisons were performed was a large grapevine microarray dataset, consisting of 472 microarray experiments, each containing 16,602 probesets. The co-expression networks generated from this dataset were very large, containing thousands of nodes and edges. The second type of dataset included the fully sequenced genomes of 71 fungi and 211 bacteria. The networks resulting from these two datasets were smaller and simpler, allowing visual inspection of the results of a new network comparison technique we developed, namely Cross-Network Topological Overlap.

5.2 Metric Comparison Though Network Topology Profiles and Clustering Comparison

Seven similarity metrics, namely the Pearson and Spearman Correlation Coefficients, Jaccard, Sørensen, Czekanowski, SPS Indices, and Euclidean Similarity (Table 1) were used as measures for gene co-expression across several grapevine

Table 1 Similarity metrics

Similarity metric	Formula
Pearson correlation [3]	$P(X,Y) = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}}$
Spearman correlation [5]	$S_p(X,Y) = \frac{\sum_i (R_i - \overline{R})(Q_i - \overline{Q})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (Q_i - \overline{Q})^2}}$
Sørensen Index [14]	$S_o(X,Y) = \frac{2\sum_i \min(X_{Bi}, Y_{Bi})}{\sum_i (X_{Bi} + Y_{Bi})}$
Jaccard Index [7, 14]	$J(X,Y) = \frac{\langle X,Y \rangle}{\langle X,X \rangle + \langle Y,Y \rangle - \langle X,Y \rangle}$
Czekanowski Index [11, 33]	$C_z(X,Y) = \frac{2\sum_i \min(X_i, Y_i)}{\sum_i (X_i + Y_i)}$
SPS Index	$\mathrm{SPS}(X,Y) = 1 - \frac{1}{n} \sum_i \frac{\lvert X_i^2 - Y_i^2 \rvert}{X_i^2 + Y_i^2}$
MIC [17]	Maximum mutual information
Euclidean Similarity	$E(X,Y) = 1 - \frac{D(X,Y)}{\max_{X,Y}(D(X,Y))}$

Definitions of similarity metrics. X and Y are vectors of length n. X_B and Y_B are the binary vectors associated with vectors X and Y, respectively, R and Q are the rank vectors associated with vectors X and Y, respectively, $D(X,Y)$ is the Euclidean distance between vectors X and Y, and $\langle X,Y \rangle$ is the inner product of vectors X and Y

microarray experiments. The SPS (Stringent Proportional Similarity) Index is a metric we created by modifying the Czekanowski Index (also known as the Proportional Similarity Index [33]) with the aim of creating a similarity metric which was still a quantitative overlap index similar to the Czekanowski Index, but more stringent, in that vectors have to be more similar in quantitative overlap to achieve the same score as with the Czekanowski Index.

The distributions of the co-expression values for each metric are shown in Fig. 2. It is evident that the different similarity metrics have very different distributions, although certain patterns do come forward. The Jaccard and Sørensen distributions are similar. This is to be expected as both of these metrics are based on set overlaps. For two sets A and B, the set overlap formulation of the Sørensen Index $S_o(A,B)$ and the Jaccard Index $J(A,B)$ are defined as [14]

$$S_o(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{47}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{48}$$

The Sørensen and Jaccard Indices are related to each other by the following equation [14]:

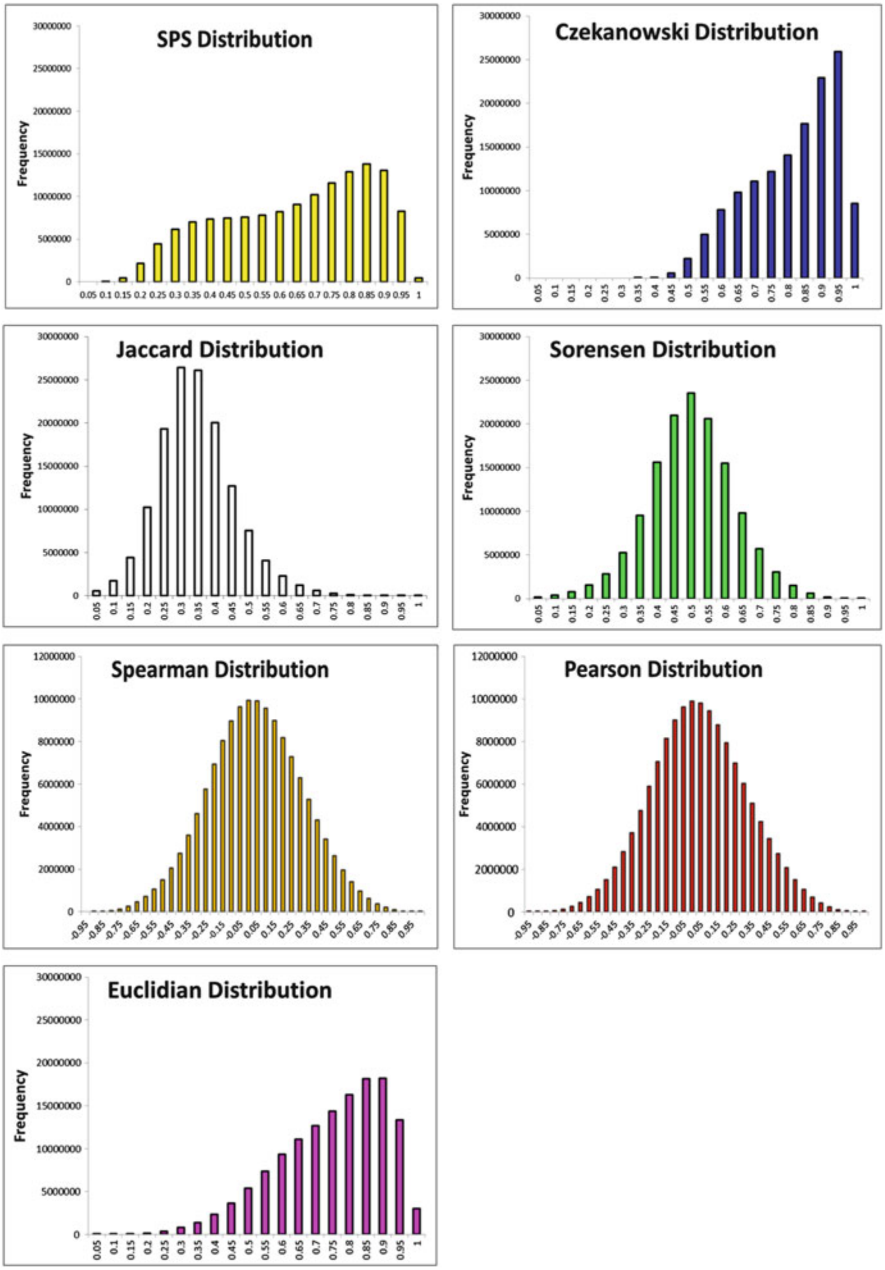


Fig. 2 Distributions. Frequency distributions of co-expression values for each of the similarity metrics when applied to the grapevine microarray expression dataset

$$S_o = \frac{2J}{J+1} \tag{49}$$

This relationship is reflected in the distributions in that the Jaccard distribution is skewed, having a longer right tail than the Sørensen distribution.

The Pearson and Spearman distributions are very similar. This seems logical as both are correlation coefficients with similar formulations (Table 1) and both measure to what extent the elements of two vectors follow the same pattern, the difference being that Pearson measures the linear relationship between two vectors and Spearman, being less stringent, measures the monotonic relationship between two vectors.

The SPS and Czekanowski distributions are similar in that they follow the same pattern of inflection points, although the SPS distribution is flatter, having less of a spike on the right side of the distribution, indicating that it is indeed more stringent than the Czekanowski Index.

5.2.1 Network Topology Profile Comparison

A network comparison method based on the principles of NetSimile [32] was developed, allowing the comparison of a set of networks in a pairwise manner. This method involved the calculation of several topology indices for each network. *Local indices* are calculated per node, and include clustering coefficients, connectivities, scaled connectivities, and maximum adjacency ratios. *Global indices* are calculated for a network as a whole, and include maximum connectivity, density, centralization, heterogeneity, and degree correlation (see Table 2). These topology indices form the variables in a topology profile for each network. Perl programs were written to calculate a series of local and global topology indices for a given set of networks and to construct topology profiles for these networks. Certain Perl programs made use of the Statistics::Basic Perl Module (Paul Miller, http://www.cpan.org/). The topology profiles form the rows of a matrix in which each row represents one of the input networks and each column represents a network topology index. Four different topology profile matrices were created with different sets of variables, namely:

1. Weighted global indices
2. Unweighted global indices
3. Weighted local indices
4. Unweighted local indices

These topology profiles can be further compared using multivariate methods such as Principal Components Analysis (PCA). To investigate further the relationships between and the effect of different similarity metrics on network topology, our network comparison method was used to compare grapevine co-expression networks generated using the seven different similarity metrics. Each co-expression network was pruned to maintain only the top 1% of edges. This pruning strategy

Table 2 Network topology indices

Topology index	Definition
Local indices	
Connectivity	$k_i = \sum_{j \neq i} a_{ij}$
Scaled connectivity	$k_i^{\text{scaled}} = \frac{\sum_{j \neq i} a_{ij}}{k_{\max}}$
Maximum adjacency ratio	$\text{MAR}_i = \frac{\sum_{j \neq i} \left(a_{ij}\right)^2}{\sum_{j \neq i} a_{ij}}$
Clustering coefficient	$CC_i = \frac{\sum_{l \neq i} \sum_{m \neq i, l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il}\right)^2 - \sum_{l \neq i} \left(a_{il}\right)^2}$
Global indices	
Maximum connectivity	$k_{\max} = \max\left(k_i\right)$
Density	$D_N = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)}$
Centralization	$C_N = \frac{n}{n-2}\left(\frac{k_{\max}}{n-1} - D_N\right)$
Heterogeneity	$H_N = \frac{\sqrt{\text{var}(k)}}{\text{mean}(k)}$
Degree correlation	Pearson (S, T)

Definitions of network indices [18, 22], where i and j are nodes, a_{ij} is the adjacency of nodes i and j, S is the vector of degrees of all source nodes, and T is the vector of degrees of all target nodes

was applied instead of a hard thresholding approach because the metrics have such varied distributions (Fig. 2).

Global and local topology indices were then calculated for each network. This resulted in the four topology profile matrices described above, each of which was analyzed with PCA. The score plot for the weighted local index matrix is shown in Fig. 3a. SPS-metric and Czekanowski Index cluster together, as do Pearson and Spearman Correlation Coefficients and Sørensen and Jaccard Indices. Intuitively, these groupings seem logical. Pearson and Spearman Correlation are both correlation coefficients and are calculated in a similar manner, except that Spearman uses ranks instead of actual variable values. Sorensen and Jaccard Indices are both set overlap measures and are calculated in a similar manner and thus would be expected to be similar. Lastly, the SPS Index was derived from the Czekanowski Index and thus it makes sense that they are similar. The score plot for the weighted global index matrix is shown in Fig. 3c. Similar groupings of metrics are seen in this score plot. It is interesting to note that the number of variables in the topology profile matrix in which the variables are local indices is vastly greater than that in which the variables are global indices. Because local indices are calculated for each node and there are thousands of nodes in each network, the number of variables in the local index topology profile matrix is very large. However, global indices are calculated only once per network, thus there are only five variables in the global index topology profile matrix. It is interesting that even though there are far fewer variables in the global index topology profile matrix than in the local index

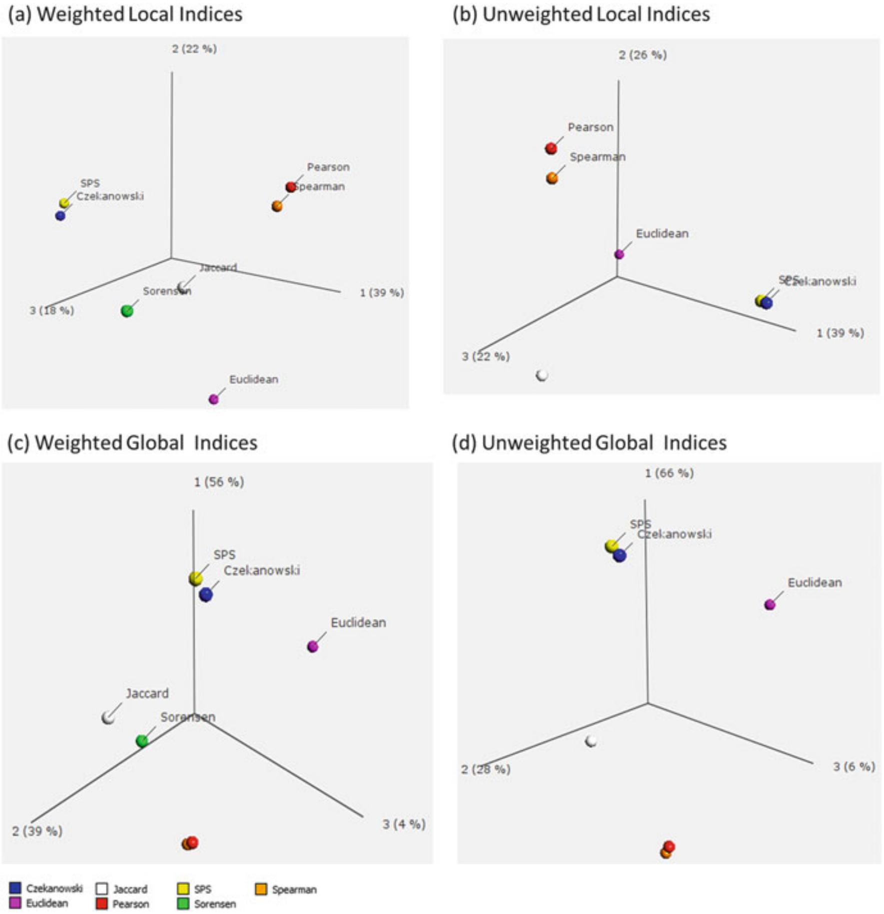


Fig. 3 Score plots: PCA of topology profiles. Score plots resulting from PCA of the topology profile matrices in which variables are (**a**) weighted local topology indices, (**b**) unweighted local topology indices, (**c**) weighted global topology indices, and (**d**) unweighted global topology indices. Scores of the Jaccard and Sørensen Index networks in (**b**) and (**d**) are identical, and thus their points in the score plots are superimposed and cannot both be visualized or labeled. Score plots were generated using Qlucore [34]

topology profile matrix, both give similar groupings in their respective PCA score plots.

In general, the score plots resulting from PCA of the matrices with unweighted indices as variables (Fig. 3b, d) have similar but tighter groupings than those resulting from PCA of matrices with weighted indices as variables (Fig. 3a, c). The Jaccard and Sørensen scores are in fact identical in both score plots resulting from using unweighted indices as variables (Fig. 3b, d).

5.2.2 Network Clustering Comparison

The seven pruned similarity networks were all clustered using MCL [28, 29] and the resulting clusterings were compared using three clustering comparison metrics, namely Average-Maximum Overlap, Jaccard Overlap, and Normalized Mutual Information (see Methods). This resulted in three all-vs-all networks in which each node represented a similarity metric and each edge represented similarity between those two similarity metrics, based on how similar the clusterings of the two respective co-expression networks were. These three clustering comparison networks are show in Fig. 4. All three clustering comparison approaches give similar results. From the thickness of the edges, it can be seen that the Pearson network clustering is most similar to the Spearman clustering, Jaccard is most similar to Sørensen, SPS is most similar to Czekanowski, and Euclidean is quite different from all other metrics. These are the same groupings which were seen in the score plots resulting from PCA of the network topology profiles and suggested by the distributions of the metrics.

5.3 Metric Comparison Through Network Merging and Cross-Network Topological Overlap

Phylogenomic networks were constructed to represent the evolutionary relationships and similarities between 71 fungal species and 211 bacterial species based on gene family content. For the fungal dataset, 8 similarity metrics were used to calculate the similarity between the gene family content of 71 fungal species. A similar procedure was used to calculate the similarity between the gene family content of 211 bacterial species, using 7 different similarity metrics.

This resulted in eight fungal MSTs and seven bacterial MSTs in which each node represented a species (either fungal or bacterial) and each edge represented similarity between the gene family content of the two species the edge connected, quantified using a particular similarity metric. In the fungal MSTs (Fig. 5), nodes were colored according to high order taxonomic groupings, whereas in the bacterial MSTs (Fig. 6), nodes were colored according to genus. From the MSTs it can be seen that, in general, all similarity metrics seem to group the species within their taxonomic groupings or genera. Thus, globally, the choice of similarity metric does not make much difference. However, locally, the choice of similarity metric results in different topologies. To visualize this better, all fungal MSTs and all bacterial MSTs were merged into two Union MSTs, one for fungi (Fig. 5i) and one for bacteria (Fig. 6h). These merged views give a good visualization on how much the similarity metrics agree on a global and a local scale. The presence of multiple edges between nodes indicates that multiple similarity metrics place these two nodes adjacent in their respective MSTs. From the Union networks in Figs. 5i and 6h, it can clearly be seen through the color distributions that these similarity

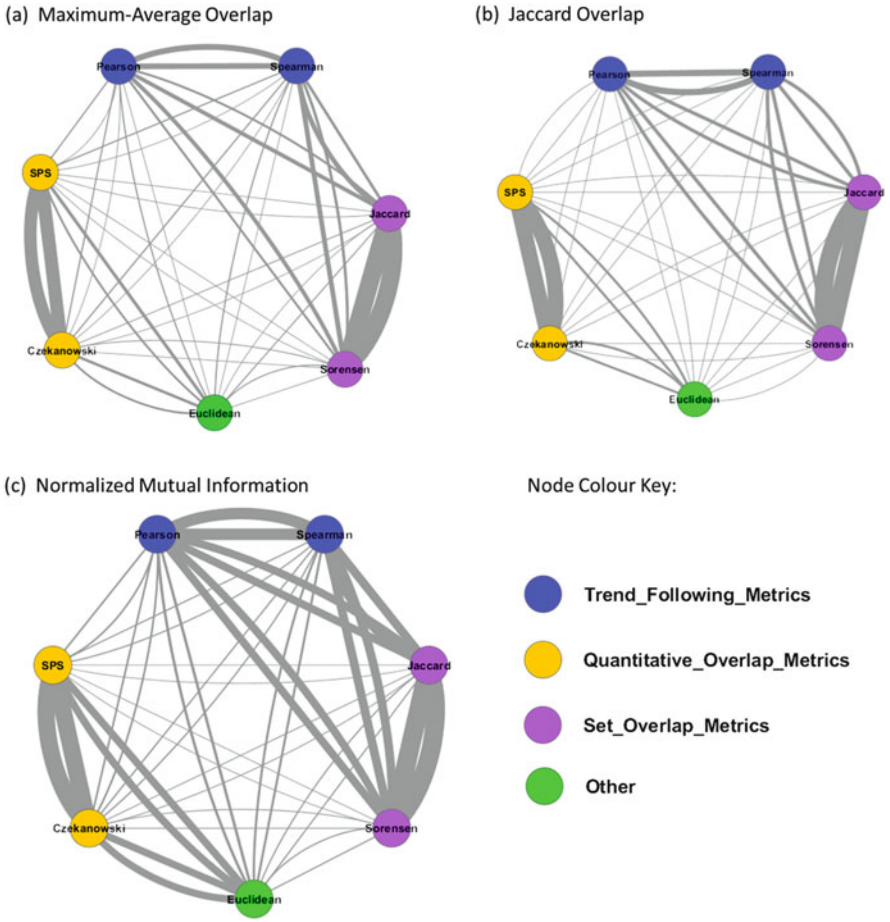


Fig. 4 Clustering similarity. Each node represents a network (in particular a gene-co-expression network) constructed using a particular similarity metric as the measure of gene co-expression. The similarity between these seven similarity metrics (nodes) is quantified by calculating the similarity between the MCL clusterings of these networks through the use of (**a**) Maximum Average Clustering Overlap, (**b**) Jaccard Clustering Overlap, and (**c**) Normalized Mutual Information between clusterings. Edge thickness corresponds to the weight of the edges based on the particular clustering similarity measure. Network visualizations were created using Cytoscape [35]

metrics generally agree on a global scale, grouping species within their taxonomic/ genera groupings. However, the similarity metrics differ on a local scale. This is illustrated by the connections between nodes which are present in only a few of the MSTs.

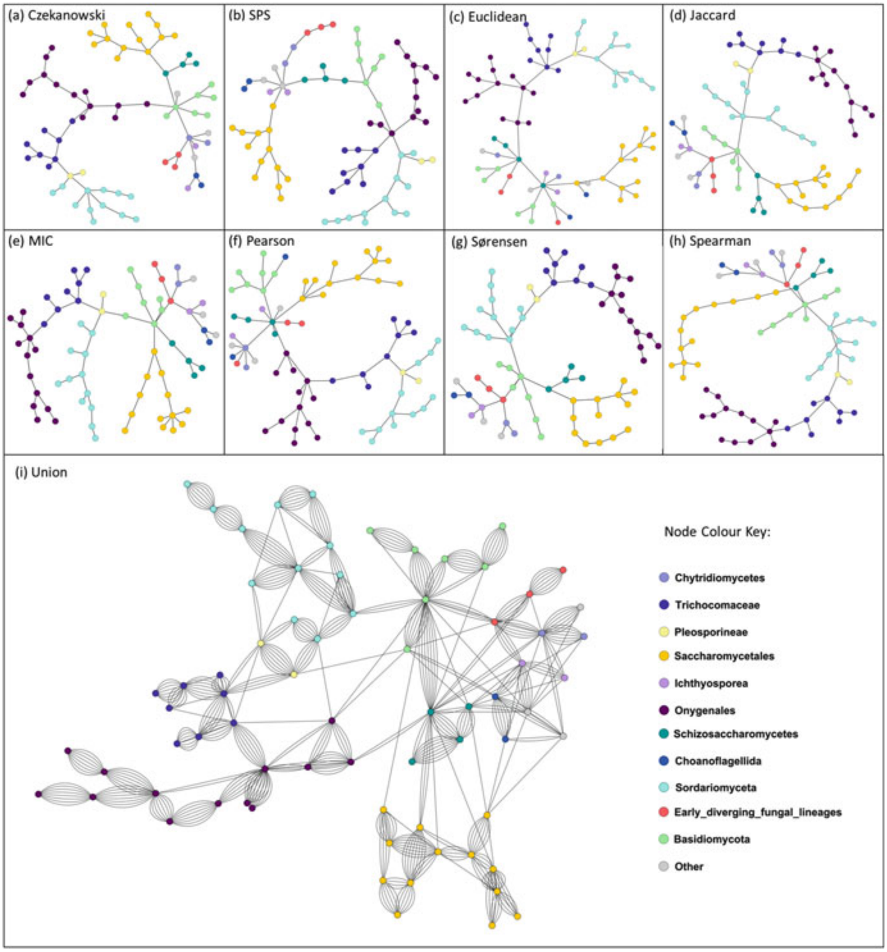


Fig. 5 Fungal Gene Family Content MSTs. Each MST shows the similarity between the gene family content of fungal species, each calculated using a different similarity metric. In each network, each node represents a fungal species and each edge represents similarity between the gene family content of two species calculated using a different similarity metric, namely (**a**) Czekanowski Index, (**b**) SPS Index, (**c**) Euclidean Similarity, (**d**) Jaccard Index, (**e**) Maximum Information Coefficient, (**f**) Pearson Correlation Coefficient, (**g**) Sorensen Index, and (**h**) Spearman Correlation Coefficient. (**i**) Union of the MSTs in (**a**)–(**h**). Species nodes are colored according to their taxonomic groupings. Network visualizations were created using Cytoscape [35]

5.3.1 Cross-Network Topological Overlap

Topological Overlap [20] is a network measure which quantifies the extent to which two nodes within a network are connected through direct connections between the two nodes and indirect connections through shared neighbors of the two nodes. We

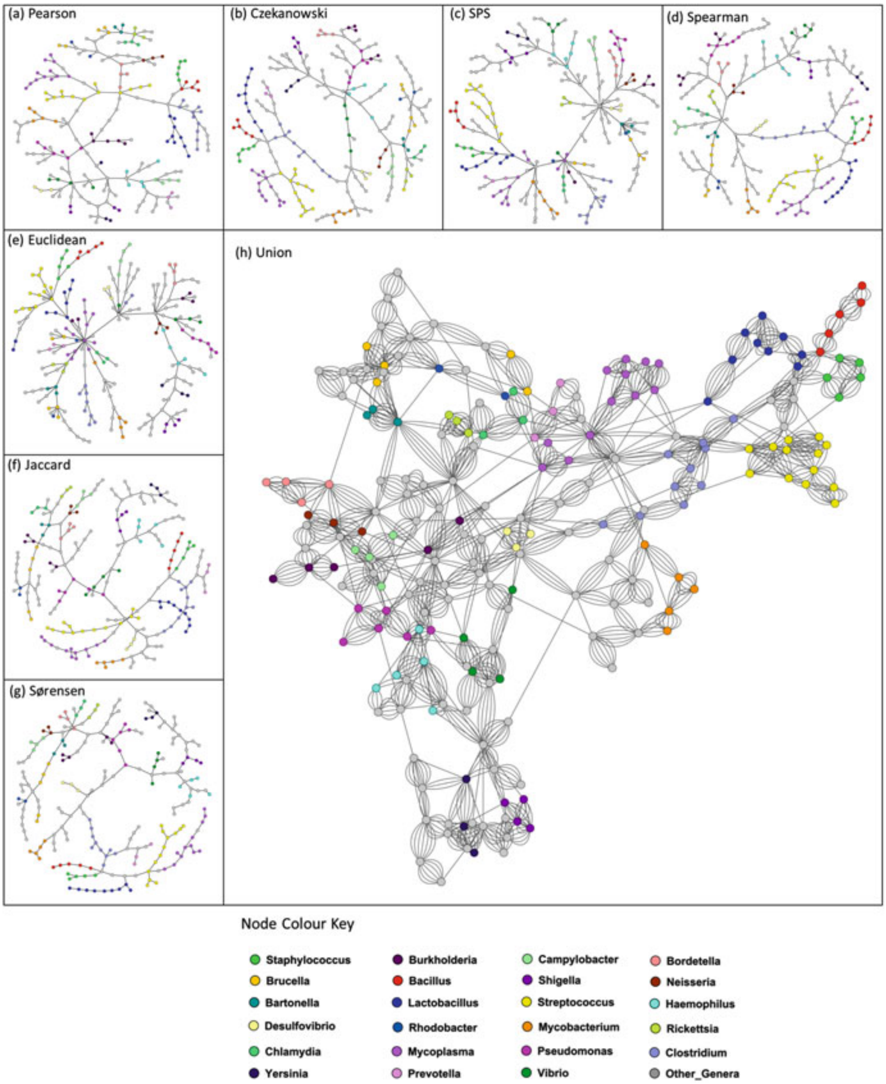


Fig. 6 Bacterial Gene Family Content MSTs. Each MST shows the similarity between the gene family content of bacterial species, each calculated using a different similarity metric. In each network, each node represents a bacterial species and each edge represents similarity between the gene family content of two species calculated using a different similarity metric, namely (**a**) Pearson Correlation Coefficient, (**b**) Czekanowski Index, (**c**) SPS Index, (**d**) Spearman Correlation Coefficient, (**e**) Euclidean Similarity, (**f**) Jaccard Index, and (**g**) Sorensen Index. (**h**) Union of the MSTs in (**a**)–(**g**). Species nodes are colored according to their genus. Network visualizations were created using Cytoscape [35]

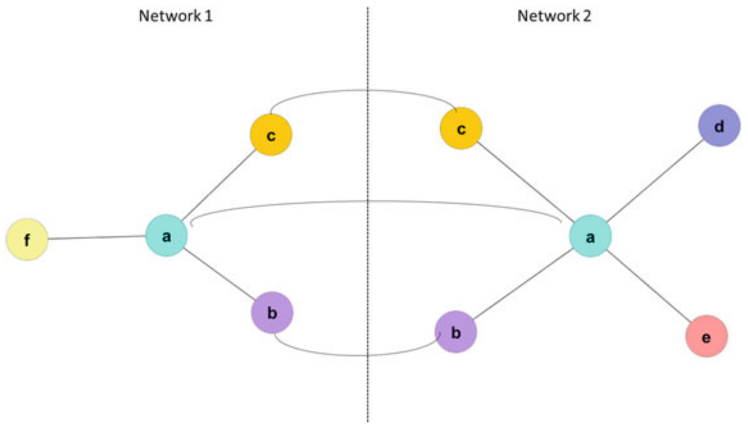


Fig. 7 Cross-Network Topological Overlap. Subnetworks of the neighborhood of node i in two hypothetical networks are shown. *Solid edges* represent edges within a network and *dashed edges* represent edges constructed to link each node with its corresponding node in the other network

extended this concept and introduce a formulation of Topological Overlap, called Cross-Network Topological Overlap, which calculates the topological overlap between nodes in different networks, quantifying the similarity between the neighborhoods of two nodes in different networks (Fig. 7). Selecting best-hits for a node in another network thus selects nodes which are topologically most similar to that node. This provides a node-by-node based approach for comparing networks.

Consider two networks, A and B. Let A_i denote the ith node in network A and B_j denote the jth node in network B. We define Cross-Network Topological Overlap (CNTO) in a directional manner. Let $CN_{TO}(A_i, B_j)$ be the CNTO of node A_i *onto* node B_j. Then, the two directional CNTOs are defined as

$$CN_{TO}\left(A_i, B_j\right) = \frac{n_{A_i,B_j} + d_{A_i,B_j}}{k_{A_i} + 1} \tag{50}$$

$$CN_{TO}\left(B_i, A_j\right) = \frac{n_{A_i,B_j} + d_{A_i,B_j}}{k_{B_j} + 1} \tag{51}$$

where n_{A_i,B_j} is the number nodes which are neighbors of both A_i and B_j, k_{A_i} is the connectivity of node A_i, k_{B_j} is the connectivity of node B_j, and d_{A_i,B_j} is defined by

$$d_{A_i,B_j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{52}$$

Thus, the directed CNTO is equal to 1 if the two nodes are in fact the same node, and if they share all their neighbors. The symmetrical CNTO of two nodes is then defined as the average of the two respective directional topological overlaps.

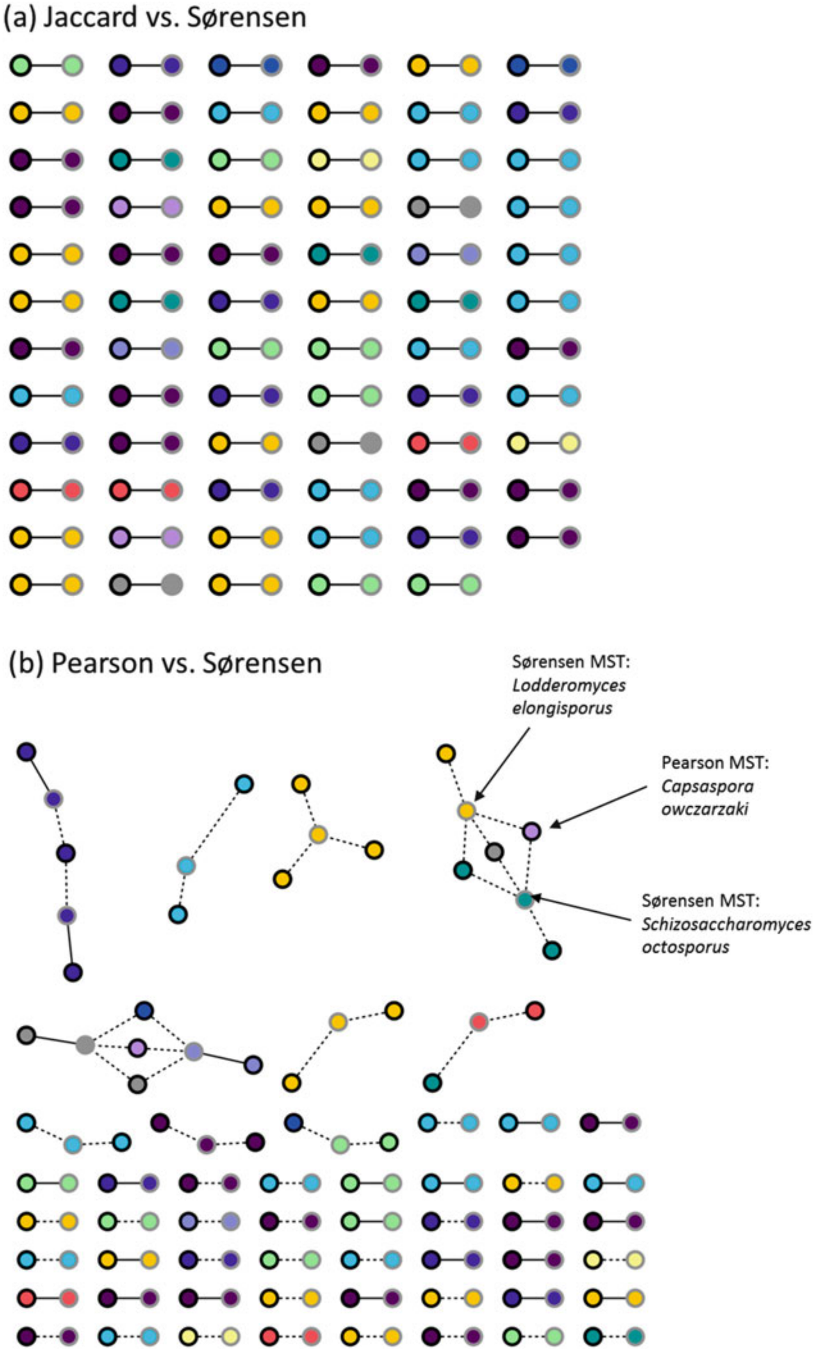

(a) Jaccard vs. Sørensen
(b) Pearson vs. Sørensen
Sørensen MST: *Lodderomyces elongisporus*
Pearson MST: *Capsaspora owczarzaki*
Sørensen MST: *Schizosaccharomyces octosporus*

Fig. 8 CNTO Networks: Comparison of Fungal MSTs. CNTO networks resulting from the comparison of fungal MSTs. Each node represents a fungal species from an MST corresponding to one similarity metric and is connected to the node(s) in an MST from another metric with which it has the highest CNTO. (**a**) CNTO network from the comparison of Jaccard and Sørensen fungal MSTs. *Black-bordered* nodes represent fungal species nodes from the Jaccard MST and *gray*

To investigate this new CNTO measure and how it can be used to compare networks, we applied it to compare the phylogenomic MSTs as these networks were simple and small enough for the output to be visually inspected.

CNTO was used to compare the Jaccard and Sørensen and the Pearson and Sørensen fungal MSTs. The Pearson and Sørensen networks were chosen for comparison because these metrics have very different definitions and were shown to result in very different network topologies when applied to the transcriptomic dataset. For a given pair of MSTs, *A* and *B*, the CNTO was calculated for all pairs of nodes *i* and *j* in which *i* is a node in *A* and *j* is a node in *B*. For each node *i* in an MST *A*, the node(s) in MST *B* with the highest topological overlap with node *i* were selected. Pairwise CNTO networks were then constructed. These networks contained two copies of each node, one from each of the two MSTs being compared, and each node is connected to the node(s) from the other MST with which they have the highest CNTO. The CNTO networks for the comparison of the Jaccard and Sørensen and the Pearson and Sørensen MSTs can be seen in Fig. 8a, b, respectively. These networks very clearly show the degree of similarity in the topologies of two networks. Figure 8a shows that the topologies of the Jaccard MST and the Sørensen MST are identical, as each node from the Jaccard MST (black bordered nodes) connects only to its corresponding node in the Sørensen MST (gray bordered nodes) with CNTO = 1. Figure 8b illustrates the similarities and differences in the topologies of the Pearson and Sørensen MSTs. Certain nodes are topologically similar between these two MSTs (illustrated by the pairs of nodes at the bottom of the network in Fig. 8b), although the disagreement of the two similarity metrics is shown largely in the top half of the network.

To illustrate how CNTO selects the most topologically similar node in another network, consider the three labeled nodes in Fig. 8b. The network shows that the nodes in the Sørensen MST most topologically similar to the species node *Capaspora owczarzaki* in the Pearson MST are *Lodderomyces elongisporus* and *Schizosaccharomyces octosporus*. The position of *Capaspora owczarzaki* in the Pearson MST is illustrated in Fig. 9a. The only information we have topologically about this node is that it is a neighbor of the node *Schizosaccharomyces japonicus*. Thus, logically, the most topologically similar nodes in the Sørensen MST should be neighbors of *Schizosaccharomyces japonicus*. Consider the Sørensen MST in Fig. 9b. Neighbors of *Schizosaccharomyces japonicus* are *Schizosaccharomyces octosporus*, *Lodderomyces elongisporus*, or *Cryptococcus neoformans*. CNTO chose *Schizosaccharomyces octosporus* and *Lodderomyces elongisporus* as more topologically similar than *Cryptococcus neoformans*, as their degrees are lower and

Fig. 8 (continued) bordered nodes represent fungal species nodes from the Sørensen MST. (**b**) CNTO network from the comparison of Pearson and Sørensen fungal MSTs. *Black-bordered* nodes represent fungal species nodes from the Pearson MST and *gray bordered* nodes represent fungal species nodes from the Sørensen MST. *Solid edges* represent CNTO = 1 (nodes are identical and share all their neighbors) whereas *dashed edges* represent CNTO < 1. Network visualizations were created using Cytoscape [35]

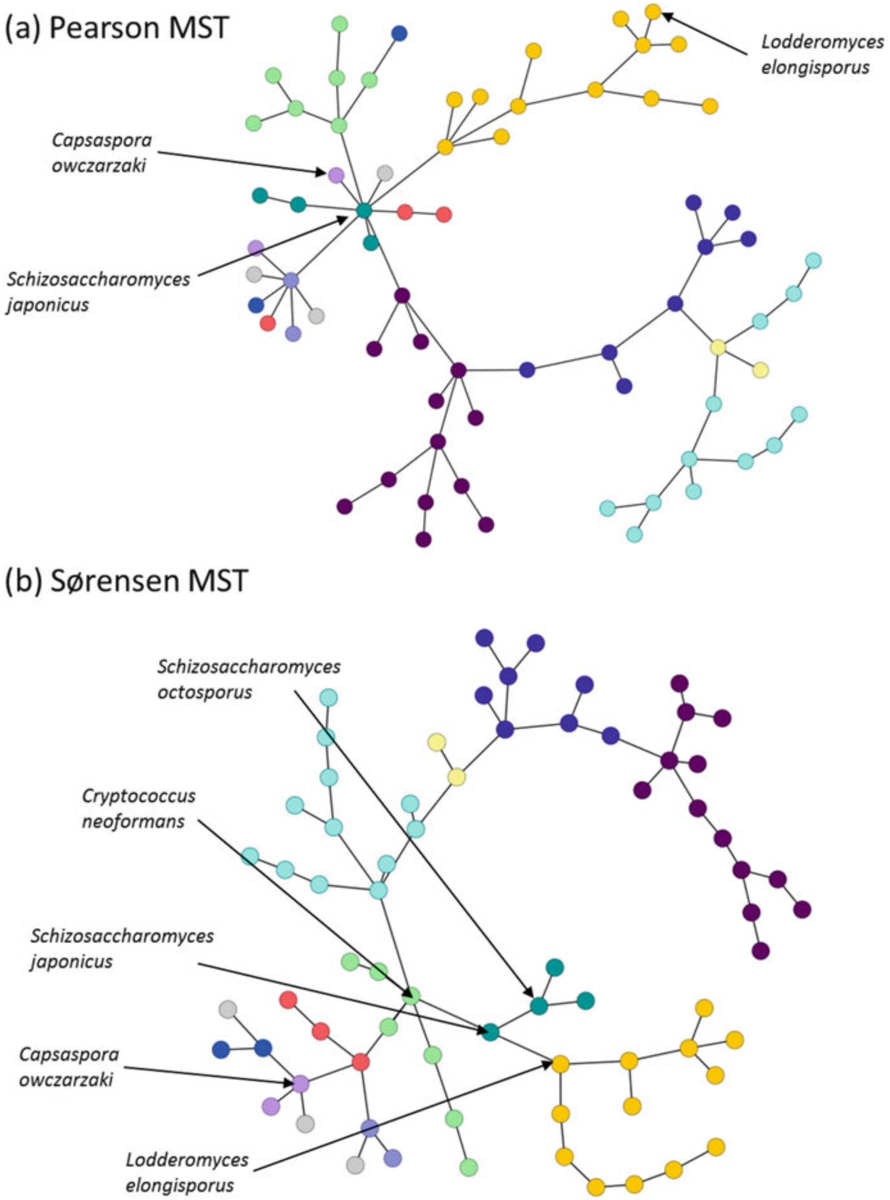

(a) Pearson MST
Lodderomyces elongisporus
Capsaspora owczarzaki
Schizosaccharomyces japonicus
(b) Sørensen MST
Schizosaccharomyces octosporus
Cryptococcus neoformans
Schizosaccharomyces japonicus
Capsaspora owczarzaki
Lodderomyces elongisporus

Fig. 9 Pearson and Sørensen Fungal MSTs. Fungal MSTs in which nodes represent fungal species and edges represent similarity between the gene family content of species quantified using (**a**) the Pearson Correlation Coefficient and (**b**) the Sørensen Index. Network visualizations were created using Cytoscape [35]

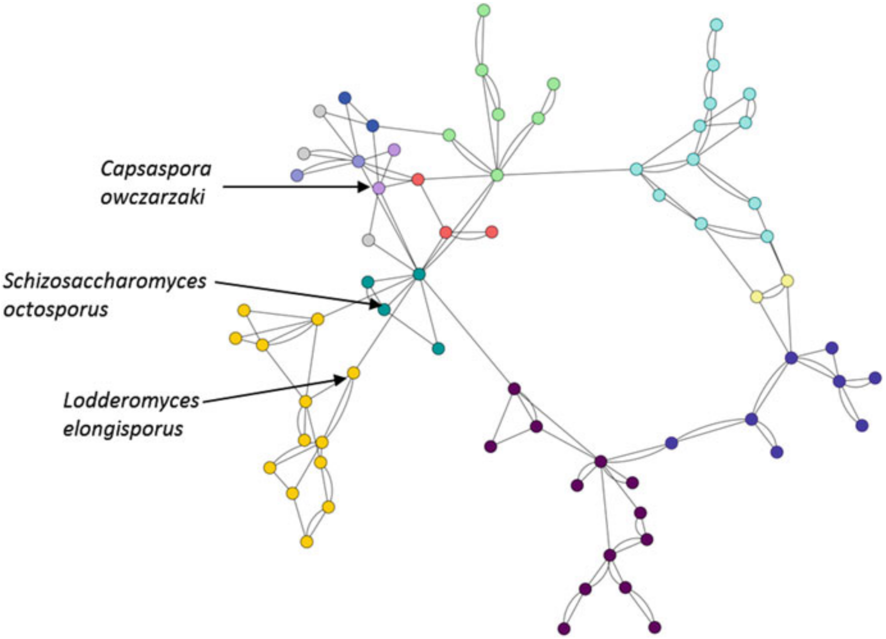


Fig. 10 Union of Pearson and Sørensen MSTs. Merged Sørensen and Pearson fungal MSTs from Fig. 9. Network visualizations were created using Cytoscape [35]

thus they have a higher fraction of shared neighbors with *Capaspora owczarzaki* than does *Cryptococcus neoformans*.

As illustrated, for a given node in a particular network, CNTO selects the node (s) in a corresponding network with the most similar topological surroundings in terms of fraction of shared neighbors. CNTO networks such as those in Fig. 8 reveal different information than would be gained from simply merging the two networks being compared. For example, consider the merged fungal Sørensen MST and Pearson MST shown in Fig. 10. This merged view gives an indication of shared edges, but does not easily show which nodes are most topologically similar in terms of shared neighbors as is shown by the CNTO networks.

MSTs, in general, have very simple topologies. They have no cycles and are very minimalistic in topology. They were chosen as example networks to develop and explore this method of network comparison because of their simplicity and ease of visualization. To explore the results of this method on networks with more complex topology, the Pearson and Sørensen all-vs-all bacterial networks were pruned to maintain the top 2.5% of edges. The resulting networks can be seen in Fig. 11. These networks have more complex topologies than the MSTs, having a much larger variance in node connectivities. CNTO was then calculated between all pairs of nodes in these two pruned networks, and a CNTO network constructed (Fig. 12). This network clearly indicates the differences in the local topologies of nodes in the two pruned bacterial networks being compared.

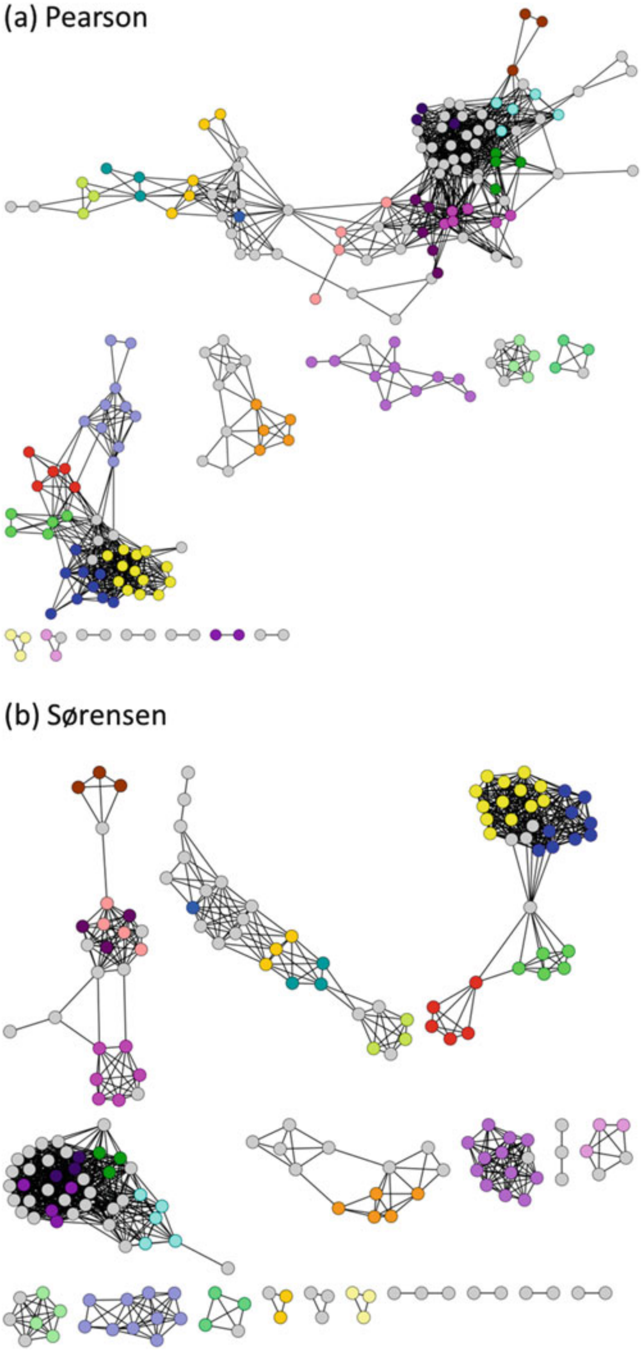


Fig. 11 Pearson and Sørensen Pruned Bacterial Networks. Pruned phylogenomic networks in which nodes represent bacterial species and edges represent similarity between the gene family content of bacterial species quantified using (**a**) the Pearson Correlation Coefficient and (**b**) the

Consider the labeled nodes in Fig. 12. The species nodes in the Sørensen bacterial network (Fig. 11b) most similar to *Lactobacillus acidophilus* in the Pearson bacterial network (Fig. 11a) are *Enterococcus faecium* and *Lactococcus lactis*. The common and uncommon neighbors of these nodes are illustrated in Fig. 13. On the left of each panel is the node in question, *Lactobacillus acidophilus* connected to its neighbors in the Pearson bacterial network in (Fig. 11a). On the right of each panel is a node from the Sørensen bacterial network (Fig. 11b) connected to its neighbors. The neighbors shared between *Lactobacillus acidophilus* from the Pearson network and the node from the Sørensen network in the right panel are enclosed in a rectangle. This figure illustrates why the CNTO measure selected *Enterococcus faecium* and *Lactococcus lactis* in the Sørensen network as more topologically similar to *Lactobacillus acidophilus* in the Pearson network rather than its equivalent node, *Lactobacillus acidophilus*, in the Sørensen network. As can be seen in Fig. 13, *Lactobacillus acidophilus* from the Pearson network shares proportionally many more neighbors with *Enterococcus faecium* and *Lactococcus lactis* in the Sørensen network than with *Lactobacillus acidophilus* from the Sørensen network.

6 Conclusions

In this study, different similarity metrics were applied to construct networks from three different datasets, and the effect of different similarity metrics on the resulting network topology was investigated through various network comparison approaches. Two new network comparison approaches and one existing approach were investigated, including PCA of network topology profiles, Cross Network Topological Overlap, and Clustering Comparison [30]. It is evident from all these investigations that the similarity metric chosen can have a large impact on the topology of the resulting network. These differences in network topology also carry through to the results of further analysis, such as clustering. The choice of similarity metric could thus greatly impact the resulting biological interpretation of the networks. A potential limitation of using network topology measures to compare networks is that certain topology measures, such as shortest path, are computationally time consuming to calculate and may become infeasible for very large networks. However, with the appropriate High Performance Computing resources, they can be applied to larger networks.

The fact that different similarity metrics result in different biological interpretations can be exploited as an advantage. As each similarity metric describes and quantifies a different aspect of similarity, the use of multiple similarity metrics

Fig. 11 (continued) Sørensen Index. Nodes are colored according to genus. These networks are pruned to maintain only the top 2.5% of edges. Network visualizations were created using Cytoscape [35]

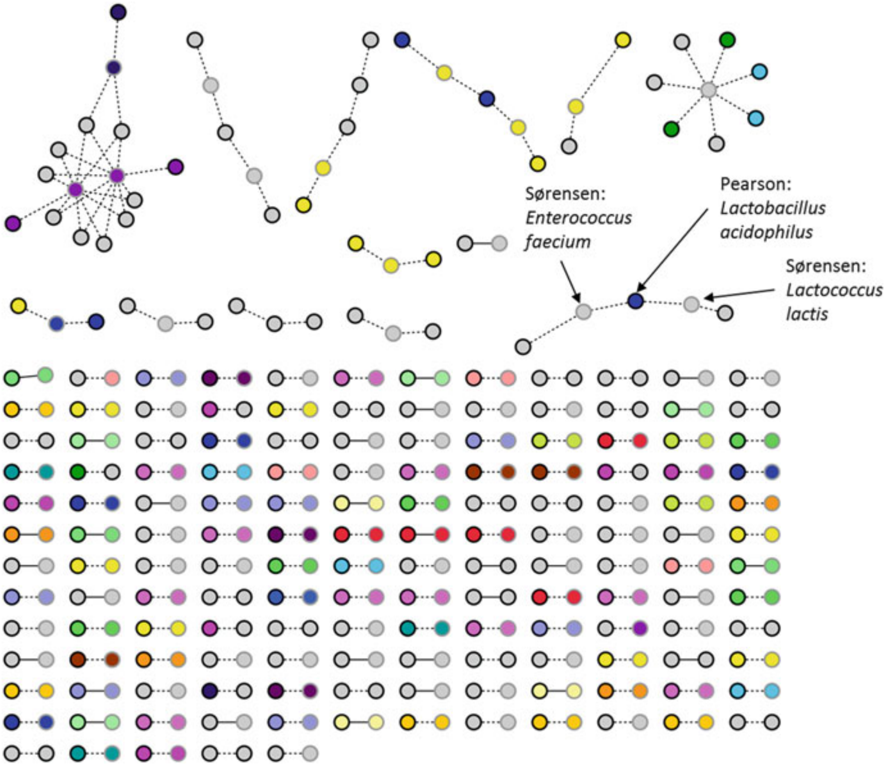


Fig. 12 CNTO Network: Pearson and Sørensen Bacterial Networks. Comparison of the pruned bacterial networks in Fig. 11 through CNTO. *Black bordered* nodes represent nodes from the pruned Pearson bacterial network (Fig. 11a) and *gray bordered* nodes represent nodes from the pruned Sørensen bacterial network (Fig. 11b). Each node is connected to the node(s) in the other network with which it has the highest CNTO. *Solid edges* represent CNTO $= 1$ (nodes are identical and share all their neighbors) and *dashed edges* represent CNTO < 1. Network visualizations were created using Cytoscape [35]

provides multiple perspectives on the data, each of which is valuable. An agglomerative approach in which many different similarity metrics are used to gain different perspectives and insights into a dataset is thus appealing.

Furthermore, with the Cross-Network Topological Overlap method presented here, it is relatively easy to identify the portions of the network affected by the choice of similarity metric. This approach provides different information than would be gained from merging two or more networks being compared. CNTO specifically highlights areas of the networks with conflicting topologies in a node-based manner, connecting nodes to their most topologically similar nodes in another network, whereas network merging is an edge-based approach, simply revealing shared edges between the two networks.

This ability of CNTO to highlight and zoom in on these areas of interest is a very useful attribute, especially when comparing large networks. In addition, CNTO

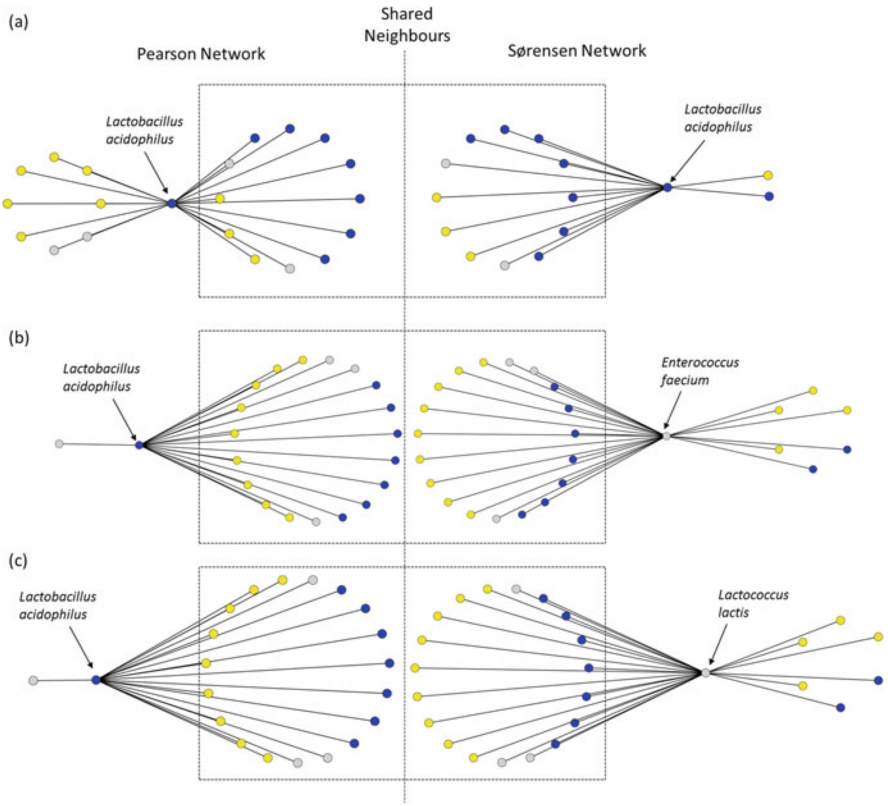


Fig. 13 Shared Neighbors. Neighbors of *Lactobacillus acidophilus* in the Pearson network, as shown in Fig. 11a, which are shared with nodes in the Sørensen network, as shown in Fig. 11b, are illustrated within this figure. *Lactobacillus acidophilus* and its neighbors in the Pearson network are shown on the *left side*, nodes in the Sørensen network and their neighbors are shown on the *right side*, and neighbors shared between the node on the left and the node on the right are enclosed in *rectangles*. Network visualizations were created using Cytoscape [35]

potentially has broader applications to network comparisons in a wide variety of real-world networks, including communication networks, transport networks, and social networks. This approach can be used to compare any kind of network with another of its kind, highlighting regions of the networks with conflicting topologies. The further application of CNTO in the comparison of various types of networks is suggested for future work.

7 Similarity Metric Effect on Network Topology: Methods

7.1 Metric Comparison Though Network Topology Profiles and Clustering Comparison

7.1.1 Co-Expression Similarity Network Construction

A total of 472 grapevine Affymetrix microarray experiments were downloaded from Gene Expression Omnibus and normalized using RMA [36]. In the resulting expression matrix E, the columns represented microarray experiments, rows represented probesets, and each entry Xi represented the $\log_2$(expression) value of probeset X in experiment i. Seven metrics were then used to calculate the similarity ("co-expression") between all pairs of probesets.

Let X and Y denote rows of the expression matrix E corresponding to the expression profiles of genes x and y, respectively. Let X_B and Y_B denote the binary vectors corresponding to X and Y, calculated as

$$X_{Bi} = \begin{cases} 1 & \text{if } X_i \geq \overline{X} \\ 0 & \text{if } X_i = 0 \end{cases}. \tag{53}$$

where X_{Bi} is the ith entry of X_B and $\overline{X}$ is the mean of X. Seven similarity metrics (defined in Table 1) were then calculated between all pairs of genes. The Pearson and Spearman Correlation Coefficients and Czekanowski, SPS, and Euclidean Distance Indices were calculated using the original vectors, and Sørensen and Jaccard Indices were calculated using the binary vectors defined in (53). The mcxarray program from MCL-Edge [28] available from http://micans.org/mcl/ was used to calculate the Pearson and Spearman correlation coefficients. Customized Perl scripts were written to calculate the other similarity metrics. This resulted in seven similarity networks (one for each similarity metric) in which each node represented a probeset and each edge represented similarity between the expression profiles of the probesets the edge was connecting, according to a particular similarity metric. These similarity networks were subsequently pruned to maintain only the top 1% of edges, including reciprocal edges but not including self-loops. Network visualizations were created using Cytoscape [35].

7.1.2 Metric Distribution Construction

A distribution was constructed for each similarity metric using a bin size of 0.05. All similarity metrics with a range of 0 to 1 (Sørensen, Jaccard, Czekanowski, and SPS Indices and Euclidean Similarity) thus had 20 bins. Pearson and Spearman Correlation Coefficients have a range of -1 to 1 and thus needed 40 bins.

7.1.3 Network Comparison Through Topology Indices

For each of the seven pruned co-expression networks, a series of network topology indices were calculated. Weighted versions of the topology indices use the actual similarity value as the weight of the edges, whereas unweighted topology indices do not acknowledge edge weights, only the presence or absence of edges in the pruned networks.

The weighted and unweighted versions of the following global (whole-network) indices were calculated for each of the co-expression networks:

1. Density
2. Centralization
3. Heterogeneity
4. Degree Correlation
5. Maximum Connectivity

The following weighted local (node-based) indices were calculated for each node in each network:

1. Clustering Coefficient
2. Scaled Connectivity
3. Connectivity
4. Maximum Adjacency Ratio

The same unweighted local indices were calculated for each network, with the exception of Maximum Adjacency Ratio, which is meaningless in the context of an unweighted network.

Topology profile matrices were then constructed in which each row represents one of the input networks and columns represent topology indices. Four topology profile matrices were constructed for which the variables were weighted local indices, unweighted local indices, weighted global indices, and unweighted global indices, respectively. PCA was performed on these matrices using Qlucore [34].

7.1.4 Network Comparison Through Clustering Comparison

The pruned similarity networks were clustered using MCL [28, 29] available from http://micans.org/mcl/. This produced a clustering (a set of clusters) for each similarity network. Perl scripts were then written to compare all pairs of clusterings using three measures, namely Average-Maximum Overlap, Jaccard Clustering Overlap, and Normalized Mutual Information.

Let C_i and C_j be two clusterings. Then the Average-Maximum Overlap between clusterings C_i and C_j was calculated as follows. For each pair of clusters (a, b) where $a \in C_i$ and $b \in C_j$, the Jaccard Index was calculated as

$$J(a,b) = \frac{|a \cap b|}{|a \cup b|} \tag{54}$$

This results in the matrix in which the rows represent clusters from clustering C_i, columns represent clusters from clustering C_j, and each entry (a,b) is the Jaccard overlap of a from clustering C_i and b from clustering C_j. The maximum value of each row is then taken, representing the "best hit" overlap for each cluster in clustering C_i. The average of these maxima is then taken, giving a score for how similar clustering C_i is to C_j. The matrix is then transposed and the process repeated, as this similarity score is not symmetric. A network was then created in which each node represented a co-expression network (constructed using a specific similarity metric) and each edge represented the Average-Maximum Overlap score between the clusterings of the two nodes (networks) the edge is connecting. The network was visualized in Cytoscape [35] and can be seen in Fig. 4a.

The Jaccard clustering overlap between clusterings C_i and C_j was calculated as [31]:

$$J\left(C_i, C_j\right) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{55}$$

where N_{11} is the number of pairs of elements (x,y) which are in the same cluster in C_i and C_j, N_{10} is the number of pairs of elements (x,y) which are in the same cluster in C_i but not C_j, and N_{01} is the number of pairs of elements (x,y) which are in the same cluster in C_j but not C_i. A network was created in which each node represented a co-expression network (constructed using a specific similarity metric) and each edge represented the Jaccard overlap between the clusterings of the two nodes (networks) which the edge is connecting. The network was visualized in Cytoscape [35] and can be seen in Fig. 4b.

The normalized Mutual Information clustering overlap between clusterings C_i and C_j was calculated as [30]

$$\mathrm{NMI}\left(C_i, C_j\right) = \frac{\sum_a \sum_b P(a,b) \log_2\left(\frac{P(a,b)}{P(a)P(b)}\right)}{\sqrt{\sum_a P(a) \log_2(P_a) \sum_b P(b) \log_2(P_b)}} \tag{56}$$

where a is a cluster in clustering C_i, b is a cluster in clustering C_j, $P(a)$ is defined as $\frac{|a|}{n}$, $P(b)$ is defined as $\frac{|b|}{n}$, and $P(a,b)$ is defined as $\frac{|a \cap b|}{n}$. The Normalized Mutual Information was calculated between all pairs of the seven clusterings (one for each co-expression network) and a network was constructed in which each node represented a co-expression network (calculated using a specific similarity metric) and each edge represented the normalized mutual information between the clusterings of the two nodes (networks) connected by that edge. The resulting network was visualized in Cytoscape and can be seen in Fig. 4c.

7.2 *Metric Comparison Through Network Merging and Cross-Network Topological Overlap*

Two datasets, one consisting of the fully sequenced genomes of 71 fungal species (downloaded from the Broad Institute [http://www.broadinstitute.org/] and the *Saccharomyces* Genome Database [http://www.yeastgenome.org/download-data] and the other consisting of the fully sequenced genomes of 211 bacterial species (downloaded from NCBI, [http://www.ncbi.nlm.nih.gov/]) were obtained and gene families were constructed.

7.2.1 Gene Family Construction

Gene families were constructed across 71 fungal species using a parallel version of OrthoMCL [37]. Gene families were constructed across the 211 bacterial species using TribeMCL [38]. TribeMCL constructs less stringent families does than OrthoMCL, although TribeMCL is faster and was thus chosen for the larger dataset of 211 bacterial genomes. In both cases, an inflation value of 2 was used during the MCL [28] clustering step. All families of size 2 or less were excluded from further analysis. From the resulting gene families, two matrices (named Species-Family Matrices or SF-matrices) of gene family content profiles were constructed, one containing fungal gene family profiles and the other containing bacterial gene family profiles. In both matrices, each column represented a species, each row represented a gene family, and each entry *ij* represented the number of genes in gene family *i* present in species *j*.

7.2.2 Phylogenomic Network Construction and Pruning

The similarity between the gene family content of all pairs of fungal species was calculated using eight different similarity metrics. Let X_i and Y_i represent the *i*th element in column *X* and column *Y* in the SF-matrix (i.e., the number of members of gene family *i* in species *X* and species *Y*, respectively. Let X_B be the binary vector associated with vector *X* and Y_B be the binary vector associated with vector *Y*, calculated as

$$X_{Bi} = \begin{cases} 1S & \text{if}_i X \geq 1 \\ 0 & \text{if} X_i = 0 \end{cases} \tag{57}$$

Eight similarity metrics (defined in Table 1) were then used to calculate the similarity between the gene family content of all pairs of fungal species *X* and *Y*. Pearson and Spearman Correlation Coefficients, MIC, Euclidean Similarity, and Czekanowski and SPS Indices were calculated using the original vectors, and the

Sørensen and Jaccard Indices were calculated using the binary vectors defined in (57).

The same procedure was performed to calculate the similarity between all pairs of bacterial species, although, in this case, only seven similarity metrics were used, as the MINE package which is used to calculate MIC failed to run on the bacterial dataset because of memory limitations.

The mcxarray program from MCL-Edge [28, 29] available from http://micans.org/mcl/ was used to calculate the Pearson and Spearman correlation coefficients. The MINE Java program [17] was used to calculate the Maximum Information Coefficient.

Applying each of these similarity metrics yielded eight all-vs-all similarity networks for the fungal dataset and seven all-vs-all similarity networks for the bacterial dataset in which each node represented a species and each edge represented the similarity between the two species which the edge connected based on the particular similarity metric. The all-vs-all networks were then pruned by calculating a Maximum Spanning Tree (MST) for each similarity network using the Perl program for MST construction used in [39]. This Perl program calculates MSTs by converting each edge weight w from a similarity value to distance value $w' = 1 - w$ and calculating a Minimum Spanning Tree on the resulting distance network using the Dijkstra algorithm from the Graph Perl Module (Jarkko Hietaniemi, http://www.cpan.org/). The resulting fungal MSTs were visualized using Cytoscape [35] and are shown in Fig. 5 and the bacterial MSTs are shown in Fig. 6. The fungal species nodes were colored by their taxonomic groupings determined using the NCBI Taxonomy Browser [40]. Bacterial species nodes were colored according to genus. The default color is gray, and thus the color gray does not indicate any specific genus or taxonomic grouping.

Two other pruned networks were created from the all-vs-all Sørensen and Pearson bacterial networks. Each of these networks were pruned by selecting the top 2.5% of edges (not including reciprocal edges or self-loops).

7.2.3 MST Merging

The two Union MSTs (Figs. 5i and 6h) were constructed by merging all fungal and bacterial MSTs, respectively, using the Cytoscape Advanced Network Merge Plugin.

7.2.4 Cross-Network Topological Overlap Networks

The Cross-Network Topological Overlap was calculated between all pairs of nodes for a selection of pairs of networks, namely:

1. Jaccard Fungal MST vs Sørensen Fungal MST
2. Pearson Fungal MST vs Sørensen Fungal MST

3. Pearson Bacterial pruned network vs Sørensen Bacterial pruned network (networks pruned to maintain only the top 2.5% of edges)

Pairs of nodes which shared no neighbors across two networks in question were excluded. For each node, the nodes in the other network with the highest topological overlap were selected, and the resulting CNTO networks were visualized in Cytoscape [35].

Author's Contributions and Acknowledgments D. Weighill and D. Jacobson conceived of and designed the methods, D. Weighill wrote the code and created the networks, D. Weighill and D. Jacobson discussed and interpreted the networks, D. Weighill drafted the manuscript, and D. Jacobson critically revised and edited the manuscript.

The research reported in this chapter was performed at Stellenbosch University, South Africa as part of a Master's thesis [41], and subsequent editing for publication in this book was performed at Oak Ridge National Laboratory and University of Tennessee, Knoxville.

Competing Interests The authors declare that they have no competing financial interests.

References

1. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
2. Pearson K (1895) Note on regression and inheritance in the case of two parents. Proc R Soc Lond 58(347–352):240–242
3. Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42(1):59–66
4. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101
5. Pinto da Costa J, Soares C (2005) A weighted rank measure of correlation. Aust N Z J Stat 47 (4):515–529
6. Jaccard P (1912) The distribution of the flora in the alpine zone. 1. New Phytol 11(2):37–50
7. Lipkus AH (1999) A proof of the triangle inequality for the Tanimoto distance. J Math Chem 26(1–3):263–265
8. Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, Rousseau R, Vanhoutte A (1989) Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. Inform Process Manage 25(3):315–318
9. Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter 5:1–34
10. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26 (3):297–302
11. Yoshioka PM (2008) Misidentification of the Bray-Curtis similarity index. Mar Ecol Prog Ser 368:309–310
12. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27(4):325–349
13. Lance G, Williams W (1966) Computer programs for hierarchical polythetic classification ("similarity analyses"). Comput J 9(1):60–64
14. Schubert A (2013) Measuring the similarity between the reference and citation distributions of journals. Scientometrics 96(1):305–313

15. Schubert A, Telcs A (2014) A note on the Jaccardized Czekanowski similarity index. Scientometrics 98(2):1397–1399
16. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets - supplementary material. http://www.sciencemag.org/content/334/6062/1518/suppl/DC1. Accessed Feb 2013
17. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detecting novel associations in large data sets. Science 334(6062):1518–1524
18. Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4(8):e1000117
19. Zhang B, Horvath S et al (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4(1):5144–6115
20. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555
21. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393 (6684):440–442
22. Reijneveld JC, Ponten SC, Berendse HW, Stam CJ (2007) The application of graph theoretical analysis to complex networks in the brain. Clin Neurophysiol 118(11):2317–2331
23. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87 (19):198701
24. Snijders TA (1981) The degree variance: an index of graph heterogeneity. Soc Networks 3 (3):163–174
25. Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Syst Biol 1:24
26. Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Netw 1 (3):215–239
27. Meilă M (2005) Comparing clusterings: an axiomatic view. In: Proceedings of the 22nd international conference on machine learning, ACM, pp 577–584
28. Van Dongen S (2000) Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht
29. Van Dongen S (2008) Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl 30(1):121–141
30. Wagner S, Wagner D (2007) Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik
31. Meilă M (2007) Comparing clusterings - an information based distance. J Multivar Anal 98 (5):873–895
32. Berlingerio M, Koutra D, Eliassi-Rad T, Faloutsos C. A scalable approach to size-independent network similarity. Available: http://arxiv.org/pdf/1209.2684.pdf
33. Bloom SA (1981) Similarity indices in community studies: potential pitfalls. Mar Ecol Prog Ser 5(2):125–128
34. Qlucore (2008) http://www.qlucore.com/. Accessed 14 Feb 2013
35. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
36. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264
37. Li L, Stoeckert C, Roos D (2003) Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189
38. Enright A, Van Dongen S, Ouzounis C (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30(7):1575–1578
39. Setati ME, Jacobson D, Andong UC, Bauer F (2012) The vineyard yeast microbiome, a mixed model microbial map. PLoS One 7(12):e52609
40. Federhen S (2012) The NCBI taxonomy database. Nucleic Acids Res 40(D1):D136–D143
41. Weighill DA (2014) Exploring the topology of complex phylogenomic and transcriptomic networks. Master's thesis, Stellenbosch University

Adv Biochem Eng Biotechnol (2017) 160: 185–196
DOI: 10.1007/10_2016_49

Published online: 26 October 2016

Molecular Phylogenetics: Concepts for a Newcomer

Pravech Ajawatanawong

Abstract Molecular phylogenetics is the study of evolutionary relationships among organisms using molecular sequence data. The aim of this review is to introduce the important terminology and general concepts of tree reconstruction to biologists who lack a strong background in the field of molecular evolution. Some modern phylogenetic programs are easy to use because of their user-friendly interfaces, but understanding the phylogenetic algorithms and substitution models, which are based on advanced statistics, is still important for the analysis and interpretation without a guide. Briefly, there are five general steps in carrying out a phylogenetic analysis: (1) sequence data preparation, (2) sequence alignment, (3) choosing a phylogenetic reconstruction method, (4) identification of the best tree, and (5) evaluating the tree. Concepts in this review enable biologists to grasp the basic ideas behind phylogenetic analysis and also help provide a sound basis for discussions with expert phylogeneticists.

Keywords Evolutionary trees, Molecular phylogenetics, Phylogenetic analysis, Phylogenetic markers, Phylogeny

Contents

1 Introduction 186
2 Phylogenetic Tree 186
3 Molecular Markers for Building a Tree 189
4 Sequence Alignment 190
5 Phylogenetic Reconstruction 191

P. Ajawatanawong (✉)
Department of Microbiology, Faculty of Science, Mahidol University, 272 Rama VI, Rachathewi, Bangkok 10400, Thailand
e-mail: pravech.aja@mahidol.ac.th

5.1 Distance-Based Approach 191
5.2 Character-Based Approach 192
6 Conclusion 194
References 194

1 Introduction

Biological research has changed rapidly in the last decade, particularly following the launch of next-generation sequencing (NGS) technology [1]. This is because NGS dramatically reduces sequencing prices, speeds up the process, and generates high-throughput DNA sequencing results. Moreover, the advancements in several "*-omic*" areas also drive biological research in the new era of bioinformatics, systems biology and networking biology. Because the generation of data is easy today, the bioresearch paradigm has shifted from the generation of sequence data to analysis efficacy and power.

Molecular phylogenetics is a disciplinary study of evolutionary relationships amongst organisms using molecular sequences. The analysis methods used in molecular phylogenetics were originally developed to reveal evolutionary pathways, yet today molecular phylogenetics is used in several fields, such as systematic biology and biodiversity [2], molecular epidemiology [3–5], identification of gene functions [6], and microbe identification in microbiome studies [7–9]. For these reasons, molecular phylogenetics is a fundamental field in science of which most biologists require background knowledge.

This review aims to introduce network biologists who are new to the field of molecular phylogenetics to the basic concepts and ideas behind phylogenetic analysis. It begins with the frequently used terminology, characteristics of sequencing markers, and general methods for tree reconstruction and tree evaluation. It then discusses some popular computer programs and critical points that need to be considered in the analysis.

2 Phylogenetic Tree

A phylogenetic tree or phylogeny is a tree-like diagram used to visualize evolutionary relationships among a set of operational taxonomic units (OTUs). The OTU generally represents a species, but can also represent individual organisms in a population, a gene or protein sequence or a taxon at any taxonomic rank (e.g., family, order, class, phylum). The tree is composed of nodes and branches (Fig. 1). Nodes at the tips of the tree are called '*external nodes*.' These are used to represent the OTUs. Another type of node, called '*internal nodes*,' represents a recent common ancestor (RCA). Between these are lines, called '*branches*,' used to connect newer and older nodes and show the evolutionary relationships among

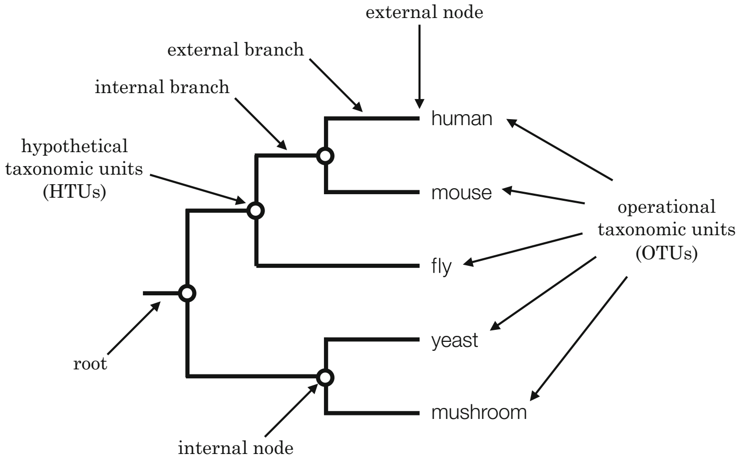


Fig. 1 Composition of a phylogenetic tree. Terminology frequently used in phylogenetic trees is labeled on the tree

the taxa. A branch linking two internal nodes is an '*internal branch*,' which shows an ancient relationship. Conversely, the branch joining an internal node with an external node to show a modern relationship is called an '*external branch*.'

The deepest branch of the tree represents the '*root*' or the '*most recent common ancestor*' (MRCA) of all taxa in the tree. Generally, phylogenetic software can only reconstruct an '*unrooted tree*' or a tree showing who is closely related to whom. To give the tree more meaning in an evolutionary context, the '*rooted tree*' is reconstructed by identifying the origin of all taxa. The best way to root a phylogenetic tree is by adding an '*outgroup*' in the dataset. Theoretically, the root of the tree is located between the outgroup and the remaining taxa. So the best outgroup is an organism or group of organisms recently diverged from the remainder of the organisms in the tree. If an outgroup is unknown or if an ideal outgroup is unavailable (e.g., if there are no data or closely related specimen available), the middle point of the longest branch on the tree can be used as the root of the tree.

The branching pattern dividing the two new nodes is called a '*bifurcation*' or a '*dichotomy*.' This fits with the concept of speciation, in which organisms split from one ancestor into two new species. The tree that contains only bifurcating nodes is a '*fully resolved tree*.' If a deeper node branches into more than two new nodes (three or more), this branching pattern is said to be '*multifurcating*' or a '*polytomy*.'

To read a tree properly, one needs to understand that all branches on the tree can be rotated around a node while retaining the same meaning in the context of evolutionary relatedness (Fig. 2a). Sometimes unrooted phylogenies are drawn in a star-like shape, also called a '*star tree*.' In this case, all branches can be rotated too (Fig. 2b) and the angles of all nodes are meaningless. A phylogenetic tree clusters taxa based on their evolutionary relationships. The closely related taxa are grouped together and share an RCA, whereas more distantly related taxa share a deeper (earlier) common ancestor. All taxa that are descended from the same ancestor

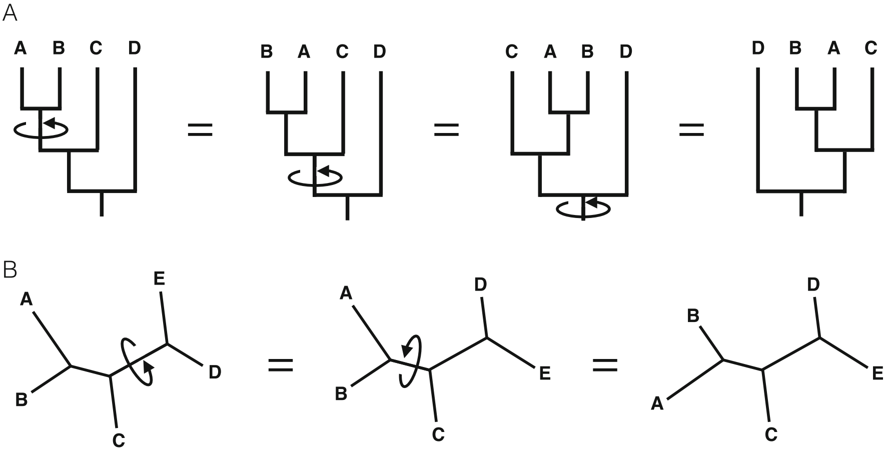


Fig. 2 A phylogenetic tree is similar to a mobile. Rotating the branches on the tree does not change the topology (branching pattern) and meaning of the evolutionary relationship in both rooted (**a**) and unrooted (**b**) phylogenies

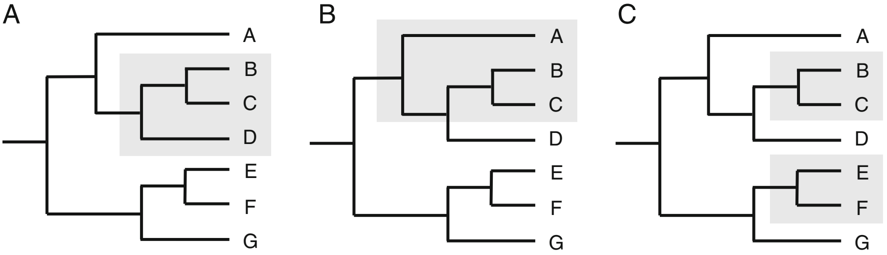


Fig. 3 Examples of a monophyletic group (**a**), a paraphyletic group (**b**), and a polyphyletic group (**c**) are shown in *gray*

make up a '*monophyletic group*' or '*clade*' (Fig. 3a). However, a group of organisms that shares the same ancestor, but does not include all members descending from that ancestor, is called a '*paraphyletic group*' or '*glade*' (Fig. 3b). Another type of group in phylogenetics is a '*polyphyletic group*' (Fig. 3c). This term refers to a group of taxa that are homoplasy. This means that they are not derived from the same ancestor and the term is usually uses for describing convergent evolution.

Molecular phylogenetic analysis must begin with a set of homologous sequences. The homology in molecular sequences is based on the sequences being derived from the same ancestor. With this in mind, molecular homology can be classified into three different types based on genetic mechanisms that separate the daughter sequences. The first type is '*orthologous genes*.' This means that the sequence was once present in the genome of an ancestor, and was transferred to the new species by speciation. This kind of gene is potentially informative for molecular phylogeny. Conversely, some genes are duplicates of other genes in the same genome and are called ‘*paralogous genes*.’ They can cause confusion in the tree reconstruction. Finally, the type of homologous genes that must be avoided in molecular phylogenetic reconstructions are ‘*xenologous genes*.’ These arise from horizontal gene transfer from one species to another. This type of gene can be problematic for a gene tree reconstruction and so usually are best avoided.

3 Molecular Markers for Building a Tree

Over the last few decades, DNA sequences have been accepted and widely used as molecular characters for phylogenetic tree reconstructions, surpassing the use of morphological characters [10]. This is because the sequence states of DNA, which can be only adenine, thymine, cytosine, or guanine, are clearer than morphological states. Molecular sequences also provide a large number of characters for phylogenetic analysis. For example, a phenotype regulated by single gene or a group of genes can be recognized as one character, but almost all positions in a gene’s DNA sequence are useful characters for phylogenetic analysis. In addition, sequence-based phylogeny allows scientists to compare organisms across higher taxonomic ranks, such as class, phylum, or even kingdom, despite a lack of comparable morphology (see [11] for further discussion).

Ribosomal RNA (rRNA) sequences in the small subunit (SSU) of the ribosome (16S rDNA sequences for prokaryotes and 18S rDNA sequences for eukaryotes) are the most widely used molecular region for phylogenetic analyses [12–15]. There are several reasons why the SSU is a very powerful marker [16, 17]. First, it is an ancient molecule which emerged during the very early stages of life and it codes for a function necessary for the survival of all cellular organisms. It is therefore present in all organisms. This allows different organisms with no morphology in common to be compared. Second, this molecule is vertically transferred with a low rate of mutation. This means the SSU is very conserved in its sequence, structure, and function. Third, the SSU sequence has multiple variable regions (V1–V9) which are all flanked with conserved blocks. This is convenient for finding oligonucleotide primers to amplify a piece of the SSU DNA for testing the diversity of sequences. In addition to the SSU, there are many other sequence markers which are potentially useful and have been used for phylogenetic analyses. Generally, a potentially useful marker sequences should be single copy and located in either the genome of the nucleus or organelles [18] such as the mitochondrial or plastid genomes (see further details in [19]). They can be either coding or non-coding sequences.

The tree built from a gene is called a ‘*gene tree*.’ Normally, a gene tree can illustrate the evolutionary history of that gene, which is not necessarily the same as the story of the species’ evolution. As such, it is probable that the topology (branching pattern) of the gene tree might not be identical to the ‘*species tree*.’

Phylogenetic tree reconstruction based on multiple genes is an alternative way to improve the resolution of a gene tree and avoid the biases that come with a tree generated from a single gene. Phylogenetic signals from different genes can be combined by concatenating all the aligned sequences. This approach aims to integrate the signal from each gene to make it more intense.

Rokas and Holland [20] proposed the term '*rare genomics changes*' (RGCs), which refers to regions in the genomes of organisms in a particular clade that have rare mutational changes, which can be used as novel markers in molecular phylogeny and evolution. Some examples of RGCs include indels (insertions/deletions), sequence signatures, and amino acid composition changes, which show the potential of RGCs as evolutionarily informative markers [21–24]. There have been some attempts to use RGCs as data for phylogenetic tree reconstruction, but it is very difficult to measure the rate of evolution in these markers and there is also no accepted weighting method for them.

4 Sequence Alignment

DNA and protein sequences are the most frequently used data types in molecular phylogenetic analysis. To study deep phylogeny, one needs ancient, universal, orthologous sequences to form the dataset. However, these sequences might be very diverse and may not align properly. To circumvent this problem, protein sequences are a better choice. This is because mutations appear to have fewer effects on protein sequences. On the other hand, the study of recent evolution or phylogenetic analysis of OTUs within the same species needs DNA sequences, which are less conserved in their sequences than are proteins. Moreover, the analysis of non-coding sequences can be carried out on DNA sequences only.

Molecular phylogenetic analysis relies heavily on the accuracy of the sequence alignment. The programs used for the alignment of sequences are developed from several algorithmic approaches. One of the most popular algorithms is '*progressive sequence alignment*,' which has been implemented in several software packages, such as MUSCLE [25, 26], MAFFT [27, 28], and Clustal Omega [29]. The general concept on which progressive sequence alignment is based is the construction of a '*guide tree*,' which is not meant to be accurate. The guide tree is used to identify sequences with the highest similarity to align first. That is because they are the easiest sequences to align. Then the algorithm keeps adding less similar sequences to the previous alignment. If a gap is needed, it is inserted into the previous sequence alignment and added to all sequences. Once all sequences are aligned, a better tree, which is built from more sophisticated methods, is created and used as a guideline for improving the final alignment.

Most alignment algorithms were developed to perform a good alignment of conserved regions, but none are powerful enough to handle indel (insertion/deletion) regions properly. Moreover, most tree reconstruction methods are developed based on substitution models. Therefore, all indel regions should be removed from the alignment to avoid errors in the analysis. There are some programs that can identify conserved regions and help the user eliminate indel regions from an analysis, such as SeqFIRE [30] and GBLOCKS [31, 32].

5 Phylogenetic Reconstruction

Methods for phylogenetic reconstruction can be classified into two main approaches: distance-based methods and character-based methods. The concept behind the former is the transformation of all sequence information into a distance matrix, which is then analyzed using an algorithm for clustering the taxa. Building a tree with this method is fast but all sequence information is lost in the process. The latter method is time-consuming because all the sequence information is used for the evaluation of the best phylogenetic tree. The calculation of phylogenetic trees using this method can be carried out using several approaches, such as maximum parsimony (MP), maximum likelihood (ML), or Bayesian analyses.

5.1 Distance-Based Approach

The key concept behind distance matrix methods is the conversion of a pairwise sequence alignment into distant values. Because a multiple sequence alignment (MSA) must contain three or more sequences, distance values from all possible pairwise sequences generate a distance matrix. Once a matrix is developed, the alignment is no longer used for the phylogenetic reconstruction. At this point, the matrix is used as the input for the tree building. Different tree building approaches used include the unweighted pair group method with arithmetic mean (UPGMA), weighted pair group method with arithmetic mean (WPGMA), neighbor-joining (NJ), least square (LS), and minimum evolution (ME) methods.

To infer sequence evolution, substitution models are used to calculate a distance value. The simplest method, which can infer the distance from both nucleotide and protein sequences, is *p*-distance. This is based on the level of sequence similarity for each pair in the alignment. Jukes–Cantor's one-parameter (JC69) model assumes that all changes in nucleotides occur at the same rate [33], whereas Kimura's two parameters (K80) model treats the occurrence of transitions and transversions as different rates [34]. The JC69 and K80 models both assume nucleotide substitution moves toward an equilibrium, which means the frequency of each nucleotide is close to 0.25. In the case of disequilibrium, one needs to employ another substitution model which fits the observed mutations. Some other models include F81 [35], HKY85 [36], TN93 [37], and more (see details in [38, 39]). Using an appropriate model for phylogenetic tree reconstruction is important to avoid errors in the clustering step. There are a number of software packages used for testing the applicability of the relevant model against the MSA, such as ModelTest [40] and jModelTest [41].

It is more complicated to infer protein substitutions. This is because changes in protein sequences result from substitutions in the DNA. However, there have been some attempts to observe amino acid substitutions in protein sequences by using a protein substitution matrix. There are two main matrix approaches generally used in sequence analysis software, including those used in phylogenetic analysis. One of these is called the percentage accepted mutation (PAM) matrix [42] and the other is the blocks substitution matrix or BLOSUM [43]. The PAM models with a higher number (e.g., PAM250) and the lower number BLOSUM matrices (e.g., BLOSUM30) are suitable for more diverse amino acid sequences, whereas the PAM models with a lower number (e.g., PAM60) and the higher number BLOSUM matrices (e.g., BLOSUM90) are suitable for the highly conserved amino acid sequences.

The major advantage of distance matrix methods is their rapid calculation speed. This is possible because the method dramatically reduces the amount of data from a long sequence alignment into a single distance matrix. Moreover, this method may give reliable results if homoplasy is rare and randomly distributed throughout the tree. However, reduction of the data leads to a loss of sequence information and can sometime generate negative branch lengths, which lack biological meaning. Instead, distance-based approaches (e.g., the NJ method) are recommended for large datasets ($>$1,000 sequences) with high sequence similarity.

5.2 Character-Based Approach

There are several methods that have been developed from character-based approaches, such as maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference methods. These approaches aim to reconstruct a phylogeny directly from the sequence data, without any transformation. They make extremely slow calculations but the final tree is said to be very accurate. Briefly, the algorithm used in these begins with scoring all possible phylogenies that can be generated from the n taxa. Then the optimal tree is assumed to be the tree with the best score. However, it is nearly impossible to score all of the individual trees when the number of taxa is larger than 20 (as this means the number of possible trees is larger than 2.21×10^{18}) by using a greedy method that searches all possible trees. Some computational search algorithms allow the user to score and select from all possible trees simultaneously. They also reduce the number of possible trees by skipping the theoretically impossible topologies from the possible trees, resulting in an increased search speed. Two popular search algorithms, which are implemented in most current phylogenetic software, are the '*branch-and-bound*' and '*heuristic*' methods. The process of the former method starts with the generation of a core tree: a three-taxa phylogeny. Then a random new taxon from the dataset is added into the core tree, and the only the new trees with an improved score have the fifth taxon added to them. This process is continued until the algorithm reaches the last taxon. The heuristic method is generally similar to the branch-and-bound method, but instead of adding new taxa into the tree with an improved score over the previous tree, the heuristic method uses only the tree with the best score in each round of taxon addition.

The maximum parsimony (MP) method—the oldest phylogenetic method—is a substitution model-free method for phylogenetic tree reconstruction. It is mostly used for building trees from morphology-based data, where it is difficult to measure the rate of evolutionary change. When this method is applied to molecular sequences, each column in the MSA is treated as a individual character. Even though each molecular sequence contains numerous characters, not every position is useful in the MP analysis (e.g., invariable sites). Characters (columns in the MSA) having at least two states (more than two types of nucleotide or amino acid) are called '*parsimony informative sites*,' and only these are included in the MP analysis. The MP method searches for the '*the most parsimonious tree*' or '*the maximum parsimony tree*,' which requires the minimum number of steps to build. Phylogenetic tree reconstruction using this method can give a reliable result if homoplasy occurs in the sequence data either randomly or infrequently. Moreover, this method can be easily applied to any novel type of data, such as indel positions. However, most sequences do not simply evolve at a low rate, and as a result sequences can be difficult to align, which makes MP less efficient, particularly when alignment patterns are complicated. MP is a time-consuming method, and it is not recommended when multiple-gene sequences are concatenated or with sequences with high levels of variation [44].

The second popular method for phylogenetic tree reconstruction is maximum likelihood (ML). ML is a statistical method used to estimate the parameters of a model given the data, and was first applied in phylogenetic analyses of DNA and protein sequences by Felsenstein [35]. In phylogenetic analysis, the ML method estimates the branch lengths and topology of the tree based on the substitution model and the sequence alignment. The numerical output of the ML analysis is the probability that a tree topology and model fit to the sequences. The calculation is repeated for all possible tree topologies that can be generated from *n* taxa. The tree topology with the highest maximum likelihood value is then reported as the best tree or the '*maximum likelihood tree*.' The strong point of the ML method is that it is claimed to be very accurate. This is because the analysis relies heavily on the evolutionary model. Because of this, all substitution models that can be used in the distance matrix methods can also be used for tree selection. Unlike the MP method, the ML method uses all the information in the sequences to calculate the maximum likelihood value. However, this results in a slow calculation time. Likewise, another weak point of the ML method is that it is impractical for large data sets. This is because the calculation is robust and requires significant computational resources.

Bayesian statistics is the newest method, which was first used for phylogenetic tree reconstruction about two decades ago [45]. This method depends on Bayesian statistics, and aims to search for the tree that maximizes the chance of seeing the model given the data (see details in [46–48]). In brief, the Bayesian phylogenetic algorithm searches for the tree that has the highest posterior probability. To deal with the enormous number of possible trees, Bayesian phylogenetic inference uses a Markov chain Monte Carlo (MCMC) algorithm to search for the best tree. This technique is more sophisticated than that used in the ML method because every new tree that is explored can produce a lower score than the tree in the previous step. This allows the Bayesian inference algorithm to find the best tree efficiently. There are some popular programs that implement the Bayesian inference algorithm, such as MrBayes [49, 50], PhyloBayes [51, 52], and BEAST [53].

Once a tree is reconstructed it is necessary to visualize it. There are no set rules for presenting a tree, but using color and renaming taxa to something easy to understand are always beneficial to the reader. Generally it is best to try to avoid using sequence codes or accession numbers to label OTUs. Likewise, it is critical to write a summary of the method used to build the tree to present in the figure legend. This helps the user to understand the tree more easily [54].

6 Conclusion

Phylogenetic analysis is one of the important techniques in the networking biologist's toolbox. It can be used to identify the evolutionary relationships among organisms, as well as gene or protein sequences. To analyze an evolutionary pathway, one needs to start with orthologous sequences and perform the analysis properly. However, single gene phylogenies generally have less evolutionary signal. As genomes are now being widely sequenced, the possibility of tree reconstruction based on entire or nearly complete genomes is emerging. This approach may replace traditional techniques in molecular evolution in the near future.

Acknowledgement I thank Prof. Prasit Palittapongarnpim and Dr. Nancy Irwins for comments on this review. This work was funded by Mahidol University.

References

1. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141
2. Senés-Guerrero C, Schüßler A (2016) A conserved arbuscular mycorrhizal fungal core-species community colonizes potato roots in the Andes. Fungal Divers 77:317–333
3. Baele G, Suchard MA, Rambaut A, Lemey P (2016) Emerging concepts of data integration in pathogen phylodynamics. Syst Biol. p ii: syw054
4. Bentley SD, Parkhill J (2015) Genomic perspectives on the evolution and spread of bacterial pathogens. Proc Biol Sci 282:20150488
5. Kenah E, Britton T, Halloran ME, Longini IM Jr (2016) Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. PLoS Comput Biol 12, e1004869

6. Chang AB, Lin R, Keith Studley W, Tran CV, Saier MH Jr (2004) Phylogeny as a guide to structure and function of membrane transport proteins. Mol Membr Biol 21:171–181
7. Carrillo-Araujo M, Taş N, Alcántara-Hernández RJ, Gaona O, Schondube JE, Medellín RA, Jansson JK, Falcón LI (2015) Phyllostomid bat microbiome composition is associated to host phylogeny and feeding strategies. Front Microbiol 6:447
8. Martiny JBH, Jones SE, Lennon JT, Martiny AC (2015) Microbiomes in light of traits: a phylogenetic perspective. Science 350:aac9323
9. Matsen FA (2015) Phylogenetics and the human microbiome. Syst Biol 64:e26–e41
10. Lumbsch HT, Leavitt SD (2011) Goodbye morphology? A paradigm shift in the delimitation of species in lichenized fungi. Fungal Divers 50:59–72
11. Hillis DM (1987) Molecular versus morphological approaches to systematics. Ann Rev Ecol Syst 18:23–42
12. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S (2007) Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73:278–288
13. Ettoumi B, Guesmi A, Brusetti L, Borin S, Najjari A, Boudabous A, Cherif A (2013) Microdiversity of deep-sea *Bacillales* isolated from Tyrrhenian sea sediments as revealed by ARISA, 16S rRNA gene sequencing and BOX-PCR fingerprinting. Microbes Environ 28: 361–369
14. Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173:697–703
15. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12: 635–645
16. Nadler SA (1995) Advantages and disadvantages of molecular phylogenetics: a case study of ascaridoid nematodes. J Nematol 27:423–432
17. Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 9, e93827
18. Zhang N, Zeng L, Shan H, Ma H (2012) Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol 195:923–937
19. Patwardhan A, Ray S, Roy A (2014) Molecular markers in phylogenetic studies—a review. J Phylogenetics Evol Biol 2:2
20. Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459
21. Ajawatanawong P, Baldauf SL (2013) Evolution of protein indels in plants, animals and fungi. BMC Evol Biol 13:140
22. Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci U S A 90:11558–11562
23. Chernikova D, Motamedi S, Csürös M, Koonin EV, Rogozin IB (2011) A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biol Direct 6:26
24. Janečka JE, Miller W, Pringle TH, Wiens F, Zitzmann A, Helgen KM, Springer MS, Murphy WJ (2007) Molecular and genomic data identify the closest living relative of primates. Science 318:792–794
25. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113
26. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
27. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

28. Katoh K, Standley DM (2016) A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics 32:1933–1942
29. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539
30. Ajawatanawong P, Atkinson GC, Watson-Haigh NS, Mackenzie B, Baldauf SL (2012) SeqFIRE: a web application for automated extraction of indel regions and conserved blocks from protein multiple sequence alignments. Nucleic Acids Res 40:W340–W347
31. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552
32. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577
33. Jukes TH, Cantor CR (1969) In: Munro HN (ed) Mammalian protein metabolism. Academic, New York, pp 121–123
34. Kimura MA (1980) Simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol 16:111–120
35. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
36. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174
37. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526
38. Liò P, Goldman N (1998) Models of molecular evolution and phylogeny. Genome Res 8: 1233–1244
39. Sullivan J, Joyce P (2005) Model selection in phylogenetics. Annu Rev Ecol Evol Syst 36: 445–466
40. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818
41. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256
42. Dayhoff MO, Schwartz R, Orcutt BC (1978) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, vol 5, supplement 3rd edn. Nat Biomed Res Found. pp 345–358
43. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919
44. Stewart CB (1993) The powers and pitfalls of parsimony. Nature 361:603–607
45. Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304–311
46. Alfaro ME, Holder MT (2006) The posterior and the prior in Bayesian phylogenetics. Annu Rev Ecol Evol Syst 37:19–42
47. Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275–284
48. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314
49. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 7:754–755
50. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542
51. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Phylogenetics 25:2286–2288
52. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109
53. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973
54. Baldauf SL (2003) Phylogeny for the faint of heart: a tutorial. Trends Genet 19:345–351

Index

A

AACENs. *See* Amino acid contact energy networks (AACENs)

Acetate thiokinase, 108

Acetone, 78

Adjacency, 152

AEF. *See* Alternate electron flow pathways (AEF)

Aeropyrum pernix, 115

Alicyclobacillus acidocaldarius, 115

Alternate electron flow pathways (AEF), 89

Alzheimer's disease, 58

Amino acid contact energy networks (AACENs), 65

Amino acids, 43, 57, 91
  biosynthesis, 109
  evolution, 67
  metabolism, 109, 112, 113
  networks (AANs), 57, 59, 65
  sequences, 43, 192

Anabaena
  cylindrica, 78, 79
  variabilis, 80

Apoptosis, 41, 42, 62

Aquifex aeolicus, 115

Archaeoglobus fulgidus ATCC 49558, 115

Arthritis, 62

Arthrospira platensis, 78, 79, 84, 88

Atherosclerosis, 62

Autotrophy, 90, 91, 104, 112, 114

B

Basic-leucine zipper (bZIP) transcription factor family, 135

Bayesian inference, 192

BEAST, 194

Betweenness, 154

Bifurcating hydrogenase, 115

Binding sites, 6–7

Bioelectricity, 78

Bioethanol, 78

Biofuels, 75–102

Biogas, 78–80, 93

BioGRID, 34

Biohydrogen, 78–80, 93

Biological networks, 127
  analysis, 15

BioModels Database, 115

Biomolecular networks, 57

Biowastes, 105

Bray–Curtis Index, 148

2, 3-Butanediol, 78

1-Butanol, 78, 82

C

CAGE. *See* Cap Analysis Gene Expression (CAGE)

Caldicellulosiruptor
  bescii, 105
  saccharolyticus, 105

Canberra distance, 149

Cancer, 23, 26, 27, 42, 58, 133

Cap Analysis Gene Expression (CAGE), 9

Capaspora owczarzaki, 170, 172

Carotenoids, 109

CASOP, 93

Cassava, 135

CBP. *See* Consolidated bioprocessing (CBP)

Cellobiose, 111, 113
Cell signaling network, 129
Cellular regulatory landscape, 133
Cellulose, 105, 106, 113
Cellulosome, 106, 107
Chaperones, 62
Chemolithoautotroph, 110
Chemoorganotrophs, 107, 115
Chemotrophy, 104
ChIA-PET, 8
ChIP-Seq, 1
Chlorophyll, 89
  fluorescence, 91
Chromatin, 1–10
  immunoprecipitation (ChIP), 1–10, 20
Cistrome, 10
Clostridium thermocellum, 105–107
Clustering coefficient (C_i), 59, 64, 152–154, 161, 162, 178
Clustering comparison, 156, 158, 164, 174
Coefficient of Community, 146
Co-expression subnetworks, 26
Communication pathways, 57
Connectivity, 152
Consolidated bioprocessing (CBP), 105, 106, 111
Core transcriptional response (CTR), 82
Correlation metrics, 143
Cosine similarity, 147
Creutzfeld–Jacob disease, 58
Cryptococcus neoformans, 172
Cscan, 10
CTR. *See* Core transcriptional response (CTR)
CXC chemokine receptor, 67
Cyanobacteria, 75–102
Cyanothece sp., 78, 88
Cytochrome oxidase, 91
Czekanowski Index, 148, 159

D
Database of Interacting Proteins (DIP), 34
Data integration, 15
Day/night cycle, 92
2-Dehydro-3-deoxy-phosphogluconate aldolase, 108
Deoxyribinucleoside monophosphate, 126
Desulfurococcus
  kamchatkensis, 115
  mucosus, 115
DFBA. *See* Dynamic flux balance analysis (DFBA)
Dichotomy, 187
Differential expression subnetworks, 25
Digesters, 105
DIP. *See* Database of Interacting Proteins (DIP)
Directed graph, 22
Disease-gene network, 129
Diseases, 57, 60–63, 69
Draft reconstruction, 103
Drug-target network, 129
DRUM, 92
Dynamic flux balance analysis (DFBA), 92

E
ECH. *See* Energy-conserving hydrogenase (ECH)
EGFR. *See* Epidermal growth factor receptor (EGFR)
ENCODE, 6, 10
Energy-conserving hydrogenase (ECH), 115
Enrichr, 10
Enterococcus faecium, 174
Entner–Doudoroff (ED) pathway, 108
Epidermal growth factor receptor (EGFR), 42, 45
Epigenomics, 8, 126
EPS. *See* Exopolysaccharides (EPS)
Escherichia coli, 82, 109
Ethanol, 78, 82, 115, 127
Ethylene, 78
Euclidean Similarity, 159
Evolutionary trees, 185
Exopolysaccharides (EPS), 110

F
False discovery rates (FDRs), 6
Fatty acids, 78, 109
FBA. *See* Flux balance analysis (FBA)
FBPA. *See* Fructose bisphosphate aldolase (FBPA)
FDRs. *See* False discovery rates (FDRs)
Ferredoxin, 108
  plastoquinone reductase, 91
Ferredoxin hydrogenase (FRHD), 114, 115
Ferredoxin:NAD hydrogenase (HYDFDNr), 114
Ferredoxin:NADP oxidoreductase (FRNDPRr), 114
Ferroplasma thermophilum, 105
FHL. *See* Formate-hydrogen lyase (FHL)
Flux balance analysis (FBA), 87, 90, 103–110
Flux variability analysis (FVA), 90, 103, 110
Food waste, 105

Formate-hydrogen lyase (FHL), 114
FRHD. *See* Ferredoxin hydrogenase (FRHD)
FRNDPRr. *See* Ferredoxin:NADP oxidoreductase (FRNDPRr)
Fructose bisphosphate aldolase (FBPA), 108
Functional interaction networks, 21

G
Gap-filling, 103
GDH. *See* Glutamate dehydrogenase (GDH)
Gene
  expression analysis, 15
  regulatory network, 129
Gene-protein-reaction (GPR), 85
Gene Set Enrichment analysis (GSEA), 24
Genome-scale models, 76, 77, 103–116
Gluconeogenesis, 112
Glucosidases, 109
Glutamate dehydrogenase (GDH), 115
Glycerol, 108, 111
Glycolysis, 89, 91, 112
GPCRs. *See* G-protein-coupled receptors (GPCRs)
GPR. *See* Gene-protein-reaction (GPR)
G-protein-coupled receptors (GPCRs), 61–67
  dimerization pathways, 66
Graph concepts, 22
GREAT, 7
GSEA. *See* Gene Set Enrichment analysis (GSEA)
GTPase, 61

H
Heat shock protein 90 (Hsp90), 62
  client protein pathways, 67
Hemicellulose, 105
Heterotrophy, 91
Hexane, 82
High throughput sequencing, 121, 123
Histidine, biosynthesis, 114
HOMCOS, 37, 42
HOMER, 7
Hot spots, 57, 63
Hsp90. *See* Heat shock protein 90 (Hsp90)
Human Protein Reference Database (HPRD), 34
Huntington's disease, 62
HYDFDNr. *See* Ferredoxin:NAD hydrogenase (HYDFDNr)
Hydrogen, 108
Hydrogenase, 91, 115
  bifurcating, 115
Hydroxypropionate-hydroxybutyrate cycle, 110
Hyperthermophiles, 104
Hyperthermus butylicus, 115
Hypothermophiles, 104

I
IDEA. *See* Interactome Dysregulation Enrichment Analysis (IDEA)
IntAct, 34
Interactome, 15, 17, 37
Interactome Dysregulation Enrichment Analysis (IDEA), 28
Interleukin, 42
InterPreTS, 37
Isobutanol, 78, 93, 106, 107
Isobutyraldehyde, 78

J
Jaccard Index, 146, 147, 150, 159–167, 177, 178
Jaccardized Czekanowski Index, 150
Jaccard overlap, 156, 164, 179
JASPAR, 20

K
Kimura's two parameters (K80) model, 191

L
Lactate dehydrogenase (LDH), 108, 115
Lactobacillus acidophilus, 174
Lactococcus
  lactis, 114, 174
  planatarum, 114
LDH. *See* Lactate dehydrogenase (LDH)
Least square (LS), 191
Leukemia, 62
Lignin, 105
Lignocellulose, 76, 105, 106, 111
Lipases, 109
Livestock manure, 105
Lodderomyces elongisporus, 170
LS. *See* Least square (LS)

M
Machado–Joseph disease, 62
MACS, 6

MAPK signaling pathway, 42
MARi. *See* Maximum adjacency ratio (MARi)
Markov chain Monte Carlo (MCMC), 194
Matthew's correlation coefficient (MCC), 64
Maximum adjacency ratio (MARi), 152
Maximum information coefficient (MIC), 151
Maximum likelihood (ML), 192
Maximum parsimony (MP), 192
MCC. *See* Matthew's correlation coefficient (MCC)
MCMC. *See* Markov chain Monte Carlo (MCMC)
MD. *See* Molecular dynamics (MD)
ME. *See* Minimum evolution (ME)
MEGADOCK, 39
Metabolic networks, 21, 76, 129
Metabolism genome-scale, modeling, 105
Metallosphaera sedula, 105
Methanococcus jannaschii, 104, 115
Methanol, 112
2-Methyl-1-butanol, 78
Methylene-tetrahydrofolate reductase (MTHFR), 114
MIC. *See* Maximum information coefficient (MIC)
Minimum evolution (ME), 191
MINT. *See* Molecular Interaction Database (MINT)
Mixotrophy, 92
Molecular dynamics (MD), 67
Molecular epidemiology, 186
Molecular Interaction Database (MINT), 34
Molecular phylogenetics, 185
Monophyletic group, 188
Moorella thermoacetica, 112
Most parsimonious tree, 193
Most recent common ancestor (MRCA), 187
MrBayes, 194
MST merging, 181
MTHFR. *See* Methylene-tetrahydrofolate reductase (MTHFR)
Multiple sequence alignment (MSA), 191
Mutual information, 157

N
NADH2. *See* NADH hydrogenase (NADH2)
NADH hydrogenase (NADH2), 115
NADPH, 89, 91, 93
  dehydrogenase complexes, 91
Nearest neighbor rule, 7
Neighbor-joining (NJ), 191
Network
  biology, 15, 121
  comparison, 47, 144, 156, 178
  connectivity, 63, 64
  metamodeling, 143
  parameters, 63, 64, 67
  profile comparison, 158
  reconstruction, 75, 76, 87, 121, 127, 129
  topology, 144, 151, 170, 174–181
Next-generation sequencing (NGS), 1, 2, 10, 123, 186
NJ. *See* Neighbor-joining (NJ)
Non-photochemical quenching, 92

O
Omics, 17, 77, 80–82, 93, 116, 121, 134
Orthologous genes, 188
Outgroup, 187
2-Oxoglutarate, 111

P
Paralogous genes, 189
Paraphyletic group, 188
Parsimony, 191–193
  informative sites, 193
PCNA. *See* Proliferating cell nuclear antigen (PCNA)
PCNs. *See* Protein contact networks (PCNs)
PCR. *See* Polymerase chain reaction (PCR)
PeakSeq, 6
Pearson's Correlation Coefficient (*r*), 145, 159
Percentage accepted mutation (PAM) matrix, 192
Phenol, 110, 111
Phosphotransacetylase (PTA), 115
Photoacclimation, 92
Photoautotrophs, 90, 91
Photobioreactors, 82
Photoinhibition, 91, 92
Photon harvesting, 91, 92
Photophosphorylation, 89
Photosynthesis, 77, 89–91, 135
  efficiency, 78
Photosystem I/II, 89
Phototrophy, 75, 93, 104
PhyloBayes, 194
Phylogenetic
  analysis, 185
  footprinting, 131
  markers, 185
  reconstruction, 191
Phylogeny, 185

PIPER, 39
Plastoquinone oxidase, 91
Polymerase chain reaction (PCR), 125, 126
Polyphyletic group, 188
Polytomy, 187
Post-docking, 43
PrePPI, 42
Prion diseases, 62
PRISM, 38
Proliferating cell nuclear antigen (PCNA), 61
Proteases, 109
Protein contact networks (PCNs), 59
Protein–ligand docking, 41
Protein–protein docking, 37–47
Protein–protein interactions (PPI), 18, 33, 34, 57
Protein–protein interfaces, 57, 66
Proteins, 2, 18, 58, 90, 127
  DNA-binding, 2
  docking, 33–47
  regulatory, 129
  toroidal, 61, 62, 65
Protein structure networks (PSNs), 59
Proteomics, 82, 111
PTA. *See* Phosphotransacetylase (PTA)
p53 tumor suppressor protein, 62
Pyrococcus furiosus, 115
Pyrolobus fumarii, 115
Pyrosequencing, 123

R
Recent common ancestor (RCA), 186
Regulatory networks, 1–14, 17–22, 121–131
Residue interaction networks (RINs), 59
Reverse engineering, 131, 132, 135
Rhodopsin, 61
RNAseq, 1, 9, 126

S
Saccharomyces cerevisiae, 18, 34
Schizosaccharomyces
  japonicus, 170
  octosporus, 170
SDOCK, 42
SEED model, 111
Sequence
  alignment, 190
  polymorphisms, 5
Sequencing, 1, 121
  errors, 5
  genome, 21, 77
  high-throughput, 121, 123, 186
  next-generation (NGS), 1, 2, 10, 123, 186
  pyro-, 123
  RNA, 16
  Sanger, 123
  third-generation, 125
Similarity metrics, 144, 159
Sørensen Index, 147, 159
Spearman Correlation Coefficient (r_s), 146, 159
SPP, 6
Staphylothermus hellenicus, 115
Star tree, 187
Stoichiometric matrix, 103
Streptococcus thermophilus, 114
Subgraph, 23, 24
Subnetworks, 16–27, 60, 65, 92, 128, 168
Sulfobacillus thermosulfidooxidans, 105
Sulfolobus solfataricus, 105, 110
Supercomputing, 33, 34
Support vector machines (SVMs), 43, 64
Synechococcus elongatus, 79, 88
Syngas, 112
Systems
  metabolic engineering, 76
  regulation, 121, 133

T
TCDB. *See* Transport Classification Database (TCDB)
Template-based method, 130
Thermoanaerobacterium saccharolyticum, 115
Thermobifida fusca, 111
Thermophiles, 104
Thermothrix thiopara, 104
Thermotoga maritima, 107
Thermus aquaticus, 104
Thermus thermophilus, 109
Third-generation sequencing, 125
Tisochrysis lutea, 92
TNFα. *See* Tumor necrosis factor-α (TNFα)
Topology measures, node-based, 152
  topological overlap, 153
Toroidal proteins, 61
TPI. *See* Triose phosphate isomerase (TPI)
Transcriptional regulatory networks (TRN), 127–131
Transcription factors, 1, 62, 127–136
  binding sites (TFBS), 6, 129, 131
Transcription regulation, 1, 121, 128–136
Transcription start site (TSS), 7
Transcriptomics, 16

TRANSFAC, 20
Transport Classification Database (TCDB), 106
Triose phosphate isomerase (TPI), 108
TRN. *See* Transcriptional regulatory networks (TRN)
Tryptophan, 109
TSS. *See* Transcription start site (TSS)
Tumor necrosis factor-α (TNFα), 42

U
Undirected graph, 22
Universal PBM Resource for Oligonucleotide Binding Evaluation (UniPROBE), 20
Unweighted pair group method with arithmetic mean (UPGMA), 191

V
Voronoi polyhedra, 58

W
Weighted graph, 23
Weighted pair group method with arithmetic mean (WPGMA), 191
Wood–Ljungdahl (WLD) pathway, 112

X
Xenologous genes, 189
Xylan, 113
Xylanases, 105
Xylose, 112

Y
Yeast one-hybrid (Y1H), 20

Z
ZDOCK, 39