Chapter 1

Bacterial Molecular Networks: Bridging the Gap Between Functional Genomics and Dynamical Modelling

Jacques van Helden, Ariane Toussaint, and Denis Thieffry

Abstract

This introductory review synthesizes the contents of the volume *Bacterial Molecular Networks* of the series *Methods in Molecular Biology*. This volume gathers 9 reviews and 16 method chapters describing computational protocols for the analysis of metabolic pathways, protein interaction networks, and regulatory networks. Each protocol is documented by concrete case studies dedicated to model bacteria or interacting populations. Altogether, the chapters provide a representative overview of state-of-the-art methods for data integration and retrieval, network visualization, graph analysis, and dynamical modelling.

Key words: Protein interactions, Metabolic pathways, Regulatory networks, Graph motifs, Clustering, Dynamical modelling, Bacteria, Metagenomics, Computational biology, Bioinformatics, Systems biology

1. Bacterial Models for Molecular Interaction Networks

Bacteria are everywhere, from the most hospitable to the most hostile environment. They are an important component of the microfauna that supports all biogeochemical (or nutrient) cycles. Besides their well-known role as infectious agents, bacteria are essential for the nutrition of animals (gut microbiome) and plants (nitrogen fixing and other rhizosphere bacteria) and more generally, contribute to the well being of water, soil, and air ecosystems.

For over a century, bacteria have also been fruitful model systems for deciphering fundamental biological mechanisms. In particular, *Escherichia coli*, *Bacillus subtilis*, and a few others have been extensively used to define the basic principles of gene expression and its regulation (1–6).

Jacques van Helden et al. (eds.), *Bacterial Molecular Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 804, DOI 10.1007/978-1-61779-361-5_1,

Nowadays, with over a thousand bacterial genomes sequenced, even greater opportunities have opened for experimental or computational global analysis of metabolism, physiology, and evolution. In particular, the access to comprehensive sets of molecular components (genes, proteins, regulatory signals) is at the basis of the development of novel integrative approaches, aiming at understanding the function of specific sets of these components (operons, regulons, metabolic pathways, protein complexes, etc.) in the context of the whole organism. Biological processes rely on the combined activities of molecules interconnected to form a precisely wired network. The integration of all the cellular processes results in a complex network, comprising several thousands of molecules and interactions. A schematic picture of such networks can nowadays be obtained from the huge data sets gathered from high-throughput technologies ("omics"). These biological networks are usually represented as graphs in the mathematical sense of the term: a set of nodes (representing biomolecules) and edges (describing their interactions). The topological structure of these graphs shapes their dynamical properties, as demonstrated by the pioneering work of René Thomas on the role of positive and negative feedback circuits (7, 8).

The availability of genome-scale molecular networks instigated the swift development of sophisticated software tools to store, query, and visualize the datasets, analyze network structures, identify components, analyze dynamical properties, generate relevant, and testable functional predictions. Today, biologists are confronted to several dozens of software tools that can be combined to address complementary questions about networks of interest. However, using these tools requires to understand not only the biological concepts underlying the datasets, but also the related algorithmic and statistical aspects. It is easy to get lost when confronted to the multiple possibilities for analyzing a given data set: which algorithm should be chosen to answer a particular question? How should the parameters be tuned? How should the results be interpreted? In order to put computer tools in the hands of life sciences researchers, there is a crucial need for well-documented protocols, illustrated by relevant study cases, providing general guidelines as well as more detailed comments about the impact of parametric choices and the potential traps.

This volume targets students and researchers working in the field of experimental or computational biology. Firstly, it offers an overview of the state-of-the-art in network biology and its applications to decipher various types of molecular networks in bacteria. Furthermore, it is a practical guide for computational methods and software tools that can be used to build, retrieve, visualize, analyze, and model biological networks, with concrete microbial applications (Table 1).

Table 1
Software tools covered by the method chapters

Software tool	Website	Interface	Chapter(s)
BioCyc	http://biocyc.org/	Web browser	Latendresse et al.
COLOMBOS	http://ibiza.biw.kuleuven.be/colombos/	Web browser	Fu et al.
DISTILLER	http://ibiza.biw.kuleuven.be/DISTILLER/	Web browser	Fu et al.
GINsim	http://gin.univ-mrs.fr/GINsim/	Java application	Chaouiya et al.
GNA	http://www-helix.inrialpes.fr/gna	Java application	Batt et al.
MAVisto	http://mavisto.ipk-gatersleben.de	Java webstart application	Schwobbermeyer and Wunschiers
MCL	http://micans.org/mcl/	Unix shell prompt	van Dongen and Abreu-Goodger
NeAT	http://rsat.bigre.ulb.ac.be/neat/	Web browser	Brohée Lima-Mendez Faust and van Helden
netZ	http://www3.imperial.ac.uk/theoreticalsystemsbiology/data-software	R package	Kelly et al.
PRODISTIN	http://crfb.univ-mrs.fr/webdistin/	Web browser	Baudot et al.
R graph methods	http://www.r-project.org/ http://www.bioconductor.org/ http://www.graphviz.org/ http://www.boost.org/	R packages	Le Meur and Gentleman
RegulonDB	http://regulondb.ccg.unam.mx	Web browser or MySQL or SOAP	Salgado et al.
RNSC	http://www.cs.utoronto.ca/~juris/data/rnsc/	Unix shell prompt	King et al.
SmolDyn	http://www.smoldyn.org	Unix shell prompt	Andrews
Snoopy	http://www-dssz.informatik.tu-cottbus.de/index.html?/software/snoopy.html	Stand alone application for Mac OSX, Windows and Linux	Marwan et al.

2. Experimental and *In Silico* Approaches to Unravel Interactions, Infer Pathways, and Build Networks (Part I)

The first section of the book provides a general overview of currently available experimental and in silico approaches used to characterize molecular interactions, which constitute the components of biomolecular pathways and networks.

This section starts with a review by *Bouveret and Brun* (Chap. 2), covering the techniques available to microbiologists for studying protein–protein interactions in bacteria, and the relevance of these interactions for high-throughput studies, including the deciphering of various bacterial interactomes.

Beyond description at the level of the organism, recent developments of high-throughput technologies allow for the direct sequencing of DNA from bacterial communities, providing access to the diversity and versatility of microbial populations in various environments and in function of various physicochemical parameters. *Williamson and Yooseph* (Chap. 3) elaborate upon the different steps required for such metagenomic analyses, including DNA extraction, sequencing, and computational analyses.

Microorganisms not only flirt and compete in their natural niches, but they also constantly exchange biomolecules, including DNA. *Toussaint and Chandler* (Chap. 4) summarize the reticulated relationships between prokaryotic species resulting from the exchanges of mobile genetic elements, and propose routes to improve their annotation and to decipher their evolutionary relationships. Directly related to this review, the method chapter by *Lima-Mendez* (Chap. 5) explains how to build a reticulate representation of the relationships between microbial genomes, where genetic material is transmitted by both vertical transmission and horizontal exchanges.

The next chapters deal with in silico resources (software tools and databases) to build, store, and analyze molecular networks. As discussed by *Arita* (Chap. 6), in the case of metabolic networks, the mapping of molecular interactions onto a graph usually implies some loss of information, which can be limited by selecting specific topological transformations that depend on the biological questions addressed. This is further detailed in the following chapter, where *Faust and van Helden* (Chap. 7) present a pragmatic utilization of graph-based representations of metabolic reactions for inferring metabolic pathways from sets of functionally related enzyme-coding genes. The method is illustrated with a study case directly relevant to bacteria, the inferring of metabolic pathways from operons. The same approach can also be applied to sets of functionally related enzyme-coding genes obtained from other sources such as co-expression clusters or phylogenetic profiles. The protocol by *Fu et al.* (Chap. 8) introduces COLUMBUS

and DISTILLER, two algorithms dedicated to the mining of public resources for transcriptome data and to infer transcriptional networks by a combined analysis of transcriptional profiles and *cis*-regulatory motifs. *Pellegrini et al.* (Chap. 9) summarize the concepts and methods at the basis of phylogenetic profiling. The approach relies on the analysis of presence/absence of orthologs across a set of phylogenetically related species, in order to predict sets of genes that are likely to be involved in related functions.

The section ends with a presentation of novel interfaces to the most popular databases on bacterial gene regulation (RegulonDB) and metabolism (BioCyc). The protocol by *Salgado et al.* (Chap. 10), describes three ways to retrieve transcriptional networks from RegulonDB, adapted to various types of utilization: simple download, programmatic access via Web services, and direct access to the SQL database management system. The resulting networks can then be used as input to perform different types of computational analyses, e.g., detect over-represented network motifs, identify regulatory modules, or characterize degree distributions. The protocol by *Latendresse et al.* (Chap. 11) introduces a set of flexible tools specifically designed to visualize bacterial metabolic and regulatory networks annotated in the BioCyc database. Although *E. coli* serves as the main reference, the tools described should apply to many other bacterial species, which will be covered by the MetaCyc database once sufficient annotation becomes available about their regulatory networks.

3. Topological Analysis of Bacterial Networks (Part II)

After having collected the components of a given molecular network, biologists soon realize that it is impossible to intuitively grasp its global properties. Irrespective of the method used to obtain the network (case-by-case experiments, high-throughput technologies, computer-based inference), large-scale biomolecular networks are complex, intricate, and often noisy. This complexity stimulated the development of a wide variety of theoretical approaches to analyze network topology (see, e.g., ref. 9, 10) and extract relevant modules. This section presents a set of reviews and method chapters describing a variety of topological approaches: analysis of degree distributions, detection of network motifs, network-based clustering. The chapter by *Geraci et al.* (Chap. 12) introduces the key concepts and describes standard and novel approaches to extract significant network characteristics (e.g., node degree properties, over-represented motifs) in bacterial molecular networks. The following method chapters document specific tools to perform different types of topological analyses. *Kelly et al.* (Chap. 13) present an R library enabling to fit alternative

models onto the degree distribution of a given network and to select the most likely model. This type of analysis challenges simplistic models that do not adequately fit the data (see also ref. 11). *Schwoebbermeyer and Wuenschiers* (Chap. 14) describe how to use the software MAVisto to mine networks for recurrent motifs by combining a versatile motif search algorithm with interactive exploration methods and specialized visualization techniques.

The three following chapters present graph-based clustering algorithms that can be used to extract functional modules from biological networks. *van Dongen et al.* (Chap. 15) propose several protocols and study cases where the algorithm MCL is applied to partition interaction networks derived from protein sequence similarities or correlations between gene expression profiles. *King et al.* (Chap. 16) apply the algorithm RNSC to identify densely connected subgraphs that might reveal protein complexes in protein–protein interaction networks. *Baudot et al.* (Chap. 17) illustrate the use of the PRODISTIN tool, which combines hierarchical clustering and analysis of annotated protein functions in order to extract consistent clusters of interacting proteins associated to particular biological processes (Gene Ontology annotations), thereby enabling novel functional predictions.

The two last chapters of this section present generic software packages enabling various types of network analysis. The protocol by *Brohée* (Chap. 18) explains the use of the Network Analysis Tools (NeAT), a Web-based software suite combining a variety of tools to analyze, compare, and classify biological networks (co-expression, transcriptional networks, GO annotations, etc.), with a special emphasis on the definition of robust metrics and control procedures (generation of random networks and network alterations). The chapter by *Le Meur and Gentlemen* (Chap. 19) demonstrates the power of R statistical programming language and packages to analyze statistical and topological properties of graph-based representations of biological networks.

4. Dynamical Modelling of Bacterial Networks (Part III)

A proper understanding of biological processes requires an appreciation of their temporal and spatial behaviour. Although mathematical biologists have been working for decades on the development of dynamical models that recapitulate in silico some essential aspects of the behaviour of living organisms (e.g., cell cycle), the correlation of such models with functional genomic data sets is seldom made. Here again, bacteria offer unmatched opportunities because of their limited complexity compared to multicellular organisms, their amenability to experiments, and the existence of rich data sets (in particular for *E. coli*). The last part of

the book presents a series of qualitative and quantitative approaches to model the temporal behaviour of biological networks in bacteria.

Focusing on metabolism, *Behre et al.* (Chap. 20) review established methods to detect and analyze structural invariants, pointing to conserved moieties or elementary fluxes at the basis of asymptotical stationary behaviour. *Marwan et al.* (Chap. 21) present the Petri net formalism, a framework enabling the development of discrete, continuous, or stochastic models, or even combination thereof. Petri nets are particularly suited to the modelling of metabolism and have been recently applied to signal transduction pathways, and, to a lesser extent, to transcriptional regulatory networks. This is illustrated by their case study: the modelling of phosphate regulation in enteric bacteria. *Batt et al.* (Chap. 22) focus on the regulatory network controlling *E. coli* response to carbon starvation, and describe the progressive building of a predictive qualitative dynamical model for this response and its parameterization. The protocol proposed by *Chaouiya et al.* (Chap. 23) uses a logical formalism, that allows for the simulation of bacteriophage λ development, while hinting at the specific roles of feedback circuits found at the core of the phage regulatory network.

Most detailed mechanistic dynamical models are limited to the level of single organisms. In contrast, mathematical ecology and evolutionary studies emphasize the relationships between variants within a population, or between different species. Bacteria are involved in a wide spectrum of interspecies relationships with a wide variety of organisms, from commensalism to mutualism and from symbiosis to parasitism. Analysis and modelling of these relationships is an important emerging field of research illustrated here by two chapters. *Friesen and Jones* (Chap. 24) review eco-evolutionary studies of adaptive dynamics, inclusive fitness, and population genetic models, providing insight into the strengths and weaknesses of each approach and into how current evolutionary methods can contribute to the deciphering of the mechanistic basis of host–symbiont interactions. Focusing on bacterial virulence, *Kepseu et al.* (Chap. 25) review the main mechanisms used by bacteria to trigger the production of virulence factors. The dynamical modelling of these mechanisms leads to the identification of two qualitatively distinct infectious transitions, while an irreversible switch behaviour is at the basis of an efficient onset of virulence.

The studies mentioned above refer to coarse-grain dynamical models of bacterial molecular networks, where variables represent the mean behaviour of population of biomolecules, cells, or species. Justified for large numbers of components, this approximation does not reflect the variety of behaviour that can occur at the microscopic level, in particular when the number of interacting components is small. This consideration has led to the development of stochastic methods for the modelling of the

temporal behaviour of individual components (e.g., molecules or cells) in spatially distributed interacting networks. *Andrews* (Chap. 26) presents Smoldyn, a computer program for modelling cellular systems with spatial and stochastic detail. Thanks to a novel rule-based modelling, chemical species and reactions are automatically generated as they arise in simulations, enabling computationally efficient simulations of bacterial systems.

5. Outlook: Bridging the Gap Between Bioinformatics and Dynamical Modelling

During the recent years, sustained by the progress in high-throughput "omics" approaches, our perception of the bacterial world, its dynamics and evolution have undergone a revolution.

Various microbiomes and viromes, sampled from niches as diverse as mammalian gut, oceans or polluted environments, can now be analyzed to answer questions about diseases and environmental equilibrium, which could not even have been formulated a decade ago. Several recently published studies (12–14) illustrate how several high-throughput methodologies can altogether provide a comprehensive picture of transcriptional activity, protein expression and interactions, along with metabolic activity in bacteria such as *Mycoplasma pneumoniae*. The multiplication of such investigations raises a pressing, increasing need to combine methods and expertise in mathematics, informatics, and biology to extract significant insights from the flood of molecular data.

By combining the efficient computational tools such as those presented in this and other volumes of the MiMB series (15–18), researchers will be able to organize, analyze, and visualize large-scale molecular datasets and, most importantly, to assess the statistical relevance of the features of interest. Such tools are essential to infer the molecular behaviour of bacteria in specific conditions (i.e., response to available nutriments, hosts for symbiosis, or pathogenesis, etc.). By and large, high-throughput technologies are providing us with a snapshot of the network state, but a real understanding of the biological systems' behaviour would require a proper formalization of the underlying causal relationships.

This problem is at the core of the domain of mathematical biology since half a century. In the recent years, sophisticated methods and software tools have been developed to model the spatial–temporal behaviour of bacterial networks at different scales and with different levels of details. Arguably, qualitative approaches are better adapted to cope with the current lack of quantitative data. However, with the raise of large-scale quantitative measurements, we can expect more and more opportunities for the application of detailed kinetic or stochastic approaches to the predictive dynamical modelling of bacterial molecular

networks. Irrespective of the mathematical formalism, it remains challenging to explicitly articulate functional genomics data analysis and dynamical modelling, even for a single cell. Another recurrent difficulty lies in relating components, interactions, and behaviours, from molecules to microbial populations or ecosystems and that occur at different temporal and spatial scales.

Although microbes have recurrently proved to be much more complex than expected, bacteria still offer excellent opportunities to develop novel concepts, methods, and tools to bridge the gaps between computational biology and dynamical modelling. This is a necessary step to reach the level of understanding commended by "systems biology" and, even more, for the development of synthetic biology, which beyond the characterization and predictive modelling of existing organisms aims at engineering new living systems. Among these, bacteria may be engineered with properties to cope with economic, environmental, or medical challenges (for a recent review on synthetic bacteriology, see ref. 19). Here again, efficient design requires a proper blend of computational and experimental expertise (see e.g., ref. 20).

A first important step in the synthesis of a new bacterial cell was recently reported (21). A Mycoplasma chromosome was synthesized, assembled, and introduced into cells of a different Mycoplasma species, generating cells capable of continuous genome replication and cell division with properties characteristic of the sole synthetic genome. This amazing technical accomplishment does not imply that we can master bacterial reproduction and evolution. Indeed, a systematic exploration of the consequences of rewiring regulatory interactions among global regulators in the best known bacteria, *E. coli*, led to puzzling and unpredictable results in terms of gene expression profiles and global fitness (cf. 22; Isalan, personal communication). Explicit dynamical modelling of the underlying networks may hopefully allow for the understanding of such non intuitive behaviour.

On a longer range, further methodological developments along the lines outlined in this volume will no doubt contribute to go from the understanding and possible manipulation of individual bacterial cells and cellular networks to the management of bacterial populations, of their interactions, as well of their response to surrounding viral and other predators populations.

Acknowledgments

The collaboration between the TAGC (DT) and BiGRe (JvH and AT) laboratories is supported by the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal

Science Policy Office, project P6/25 (BioMaGNet). The BiGRe laboratory is further supported by the MICROME Collaborative Project funded by the European Commission within its FP7 Programme, under the thematic area "BIO-INFORMATICS – Microbial genomics and bio-informatics" (contract number 222886–2).

References

1. Jacob F, Monod J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol, 3:318–356.
2. Monod J, Jacob F. (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harb Symp Quant Biol, 26:389–401.
3. Brock TD. (1990) *The Emergence of Bacterial Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
4. Beckwith JR, Silhavy TJ. (1992) *The Power of Bacterial Genetics: A Literature-Based Course*. Cold Spring Harbor Laboratory Press, Plainview, NY.
5. Thieffry D. (1996) *Escherichia coli* as a model system with which to study cell differentiation. Hist Philos Life Sci, 18:163–193.
6. Morange M. (1998) *A History of Molecular Biology*. Harvard University Press, Cambridge, MA.
7. Thomas R, D'Ari R. (1990) *Biological Feedback*. CRC Press, Boca Raton, FL.
8. Thieffry D. (2007) Dynamical roles of biological regulatory circuits. Brief Bioinform, 8:220–225.
9. Alon U. (2007) *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, Boca Raton, FL.
10. Martinez-Antonio A, Janga SC, Thieffry D. (2008) Functional organisation of *Escherichia coli* transcriptional regulatory network. J Mol Biol, 381:238–247.
11. Lima-Mendez G, van Helden J. (2009) The powerful law of the power law and other myths in network biology. Mol Biosyst, 5:1482–1493.
12. Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L. (2009) Transcriptome complexity in a genome-reduced bacterium. Science, 326:1268–1271.
13. Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castano-Diez D, Chen WH, Devos D, Guell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Bottcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin AC. (2009) Proteome organization in a genome-reduced bacterium. Science, 326:1235–1240.
14. Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen WH, Wodke JA, Guell M, Martinez S, Bourgeois R, Kuhner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutierrez-Gallego R, Russell RB, Gavin AC, Bork P, Serrano L. (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science, 326:1263–1268.
15. Fenyö D. (2010) *Computational Biology, Methods in Molecular Biology. Methods in Molecular Biology*, Vol. 673. Springer.
16. Ladunga I. (2010) *Computational Biology of Transcription Factor Binding. Methods in Molecular Biology*, Vol. 674. Springer.
17. Wu CH, Chen C. (2011) *Bioinformatics for Comparative Proteomics. Methods in Molecular Biology*, Vol. 694. Springer.
18. Hamacher M, Eisenacher M, Stephan C. (2011) *Data Mining in Proteomics. Methods in Molecular Biology*, Vol. 696. Springer.
19. Michalodimitrakis K, Isalan M. (2009) Engineering prokaryotic gene circuits. FEMS Microbiol Rev, 33:27–37.
20. Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J. (2008) A fast, robust and tunable synthetic gene oscillator. Nature, 456:516–519.
21. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA,

Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, 3rd, Smith HO, Venter JC. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science, 329:52–56.

22. Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, Garriga-Canut M, Serrano L. (2008) Evolvability and hierarchy in rewired bacterial gene networks. Nature, 452:840–845.

http://www.springer.com/978-1-61779-360-8

Bacterial Molecular Networks
Methods and Protocols
van Helden, J.; Toussaint, A.; Thieffry, D. (Eds.)
2012, XI, 546 p., Hardcover
ISBN: 978-1-61779-360-8
A product of Humana Press