Progress in Drug Research

Founded by Ernst Jucker

Progress in Drug Research

Systems Biological Approaches in Infectious Diseases

Vol. 64

Edited by
Helena I. Boshoff and Clifton E. Barry III

Birkhäuser Verlag
Basel · Boston · Berlin

Editors

Helena I. Boshoff
Clifton E. Barry III
Tuberculosis Research Section
Laboratory of Immunogenetics
National Institute of Allergy and Infectious Diseases
Rockville, MD 20852
USA

Library of Congress Cataloging-in-Publication Data

Systems biological approaches in infectious dieseases / edited by Helena J. Boshoff and Clifton E. Barry, III.
p. cm. – (Progress in drug research vol. 64)
Includes bibliographical references and index.
ISBN-13: 978-3-7643-7566-9 (alk. paper)
ISBN-10: 3-7643-7566-3 (alk. paper)
I. Communicable diseases–Epidemiology–Mathematical models. 2. Drugs–Design–Mathematical models. I. Boshoff, Helena J., 1968- II. Barry, Clifton E., 1963-

ISBN 10: 3-7643-7566-3 Birkhäuser Verlag, Basel – Boston – Berlin
ISBN 13: 978-3-7643-7566-9

Part of Springer Science+Business Media
Printed on acid-free paper produced from chlorine-free pulp. TCF
Cover design and layout: Micha Lotrovsky, CH-4106 Therwil, Switzerland
Printed in Germany

ISBN 10: 3-7643-7566-3 e-ISBN: 3-7643-7567-1
ISBN 13: 978-3-7643-7566-9

9 8 7 6 5 4 3 2 1 www.birkhauser.ch

Contents

Foreword

The much-lamented "innovation gap" often referenced by current authors with respect to drug discovery in the pharmaceutical industry is a sure sign that an era has passed. The reductionist view of disease as the direct consequence of isolated errors of metabolism that could be explained and understood as simple enzyme alterations is a thing of the past. Likewise the naïve view that the system-wide consequences of small molecule interventions could be predicted through simple *in vitro* assays has become obsolete. Infectious diseases, representing an evolved and complex evolutionary conflict between two life-forms, have been at the vanguard of embracing systems biology concepts due to the obvious failure to cure such diseases by simply studying an invading parasite's physiology in a test tube. Driven by a virtual renaissance in technology the simple approaches of the previous era have given way to a vast new array of integrative sciences aimed at modeling and understanding the complex and dynamic interactions that characterize real human diseases. Although still struggling for granularity these integrative sciences share a common vision – erasing the differences between disciplines and embracing complexity in tools that offer glimpses of whole biological systems and mesh seamlessly with infinite chemical space.

Rather than focus this book on the tools, approaches, successes and failures of the old era we challenged our contributors to look forward and project the tools that will become indispensable to the new era – the tools that would turn this "innovation gap" into an "innovation leap". The "omic" sciences are one prime example of the integrative approach to infectious disease. With hundreds of genome sequences of organisms from all branches of the tree of life literally at our fingertips, transcriptomics, proteomics and metabolomics are proving to be only the first wave of large, complex datasets that are now being augmented by protein interaction networks, reverse protein arrays, the protein-DNA interactome, etc. The magnitude of these datasets has challenged experimental, mathematical and computational scientists who are banding together around the

emerging discipline of "Systems Biology". Systems biology aims towards nothing less than the complete reconstruction of the biological complexity of living organisms in chemically and mathematically defined terms. Complete models for simple prokaryotes are within our grasp and models of complex multi-cellular organisms will emerge within our scientific careers and these models will have a profound impact on drug discovery.

Systems biology at present is defined by the tools employed to generate large-scale datasets. There remains a gap between those tools that have been reduced to practice and give reproducible, reliable datasets with information that allows us to model part of the system, for example transcriptomics, and tools that have critical information but cannot currently provide robust datasets such as metabolomics. Transcriptomics has been applied widely in infectious disease research and has already resulted in significant insights with therapeutic consequences. Metabolomics, however, is the frontier between analytical chemistry and biology, and the tools required for the simultaneous identification and quantitation of all the relevant small molecules in even a simple prokaryote are still being developed. Metabolomic analyses, however, have the potential to inform many aspects of the drug discovery pipeline from target identification to biomarkers of response to therapy. As the complexity of the link between transcription, translation and metabolic flux has expanded, so too have the models required to explain and interpret such data.

The information emerging from measurements and models of host-pathogen systems also requires bridging another gap between chemists with a desire for simple isolated enzyme assays and biologists with a desire for complex whole-cell based assays. Chemical genetics is one element of such a bridge and is on the verge of becoming a core large-scale technology. "Reverse chemical genetics" is perhaps the more intuitive approach where a candidate target is screened for small molecule ligands that are then used to examine the influence of target interruption in a whole-cell context. "Forward chemical genetics", however, is arguably a more powerful approach for target identification in anti-infectives programs. In this approach small molecules are directly screened for a desired phenotypic effect followed by identification of the relevant protein target in the pathogen or in the host – an exercise that minimizes the "biological uncertainty" associated with target selection. More and more often decreasing biologi-

cal uncertainty involves an intense integration of the full suite of "omics" technologies. The approach is a natural complement of traditional genetic approaches since it directly asks the therapeutically relevant interruption of protein function question in an appropriately complex system.

In a sense what all of these large-scale biology approaches are pushing towards is accurate information in highly disease-relevant environments in an effort to choose smarter targets and minimize the risk of drug development. While this is a direction that the pharma industry has been evolving towards in many ways, systems biology is pushing the fringe of what is possible. The future of many development compounds is dramatically affected by their performance at a systems level. Nowhere is this more acute than in the area of predictive toxicology where current guidelines specify increasing numbers of standard assays. The number of *in vitro* toxicology examinations that are mandatory is increasing and this trend is likely to continue. As these tests grow increasingly sophisticated (e.g. whole rabbit heart screening for cardiac toxicology assessment) they are increasingly being informed by systems biology data, and in the future toxicogenomics is likely to play a large role in preclinical development.

We think that the impending "innovation leap" in anti-infectives therapeutics development lies squarely within the sort of interdisciplinary, integrative efforts described within the systems biology framework in this book. Every step of the drug-development pathway will benefit directly from assays and models that do not make reductionistic assumptions to make predictions but rather are based upon embracing biological complexity to gain true insight into the consequences of therapeutic strategies as early as possible.

July, 2006

Helena I. Boshoff
Clifton E. Barry III

Progress in Drug Research, Vol. 64
(Helena I. Boshoff and Clifton E. Barry III, Eds)

Systems biology and its impact on anti-infective drug development

By Michael P. Stumpf[1], Brian D. Robertson[2], Ken Duncan[2] and Douglas B. Young[2]

Imperial College London, UK

[1]Centre for Integrative Systems Biology at Imperial College (CISBIC), Division of Molecular Biosciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

[2]Centre for Integrative Systems Biology at Imperial College (CISBIC), Department of Molecular Microbiology and Infection, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
<d.young@imperial.ac.uk>

Abstract

Systems biology offers the potential for more effective selection of novel targets for anti-infective drugs. In contrast to conventional reductionist biology, a systems approach allows targets to be viewed in a wider context of the entire physiology of the cell, with the potential to identify key susceptible nodes and to predict synergistic effects of blocking multiple pathways. In addition to the holistic perspective provided by systems biology, the emphasis on quantitative analysis is likely to add further rigour to the process of target selection. Systems biology also offers the potential to incorporate different levels of information into the selection process. Consideration of data from microbial population biology may be important in the context of predicting future drug-resistance profiles associated with targeting a particular pathway, for example. This chapter provides an overview of major themes in the developing field of systems biology, summarising the core technologies and the strategies used to translate datasets into useful quantitative models capable of predicting complex biological behaviour.

Keywords: imaging, integrative systems biology, mathematical models, metabolic networks, protein interaction network, targets for anti-infective drugs, transciptional networks

1 Introduction

The current approach of target-driven drug discovery is underpinned by dramatic progress that has been achieved in molecular and structural biology within a framework provided by the revolution in genome sequencing. Sequences are available for most of the major pathogens and straightforward procedures are in place for the production of recombinant proteins required for drug discovery efforts based on high-throughput screens and structure-based compound optimisation. The challenge for the future of anti-infective drug development lies in target selection. Can we develop a rational approach to target identification that will allow us to produce new drugs and drug combinations that act faster than existing compounds, that are effective against the range of adaptive microbial phenotypes generated during infection, and that reduce the evolution of drug-resistant strains? To address this challenge we have to be able to evaluate potential targets within the context of the overall physiology of both pathogen and host with a level of predictive accuracy that matches the precision that we currently apply when working with the isolated targets (see Chapters 10 and 12). This will involve taking a step back from conventional reduc-

tionist approaches and entering the domain that is commonly referred to as systems biology, a conclusion also reached by the US Food and Drug Administration (FDA) in its report *Challenge and Opportunity on the Critical Path to New Medical Products* [1].

Investigation of intact biological systems is a relatively recent concept in the molecular biosciences. In ecology and epidemiology system-level descriptions of biological processes – often coupled with a rigorous quantitative framework – have a longer history, reaching back certainly to the first half of the 20^{th} century. Advances in molecular biosciences have been achieved by a predominantly reductionist approach, based on isolation and analysis of individual components in preference to study of the system as a whole. As a result we now have rich and detailed data about the function of many genes and their protein products in an increasing number of species spanning all three kingdoms of life, and often a good understanding of how these are organised into local modules responsible for a range of cellular processes and signalling. The fledgling discipline of systems biology now aims to provide a global framework for the integrative, coherent and consistent analysis of all of the available data, moving beyond the purely descriptive towards a quantitative and predictive level of understanding.

From the perspective of infectious disease biology, an important goal of a systems-based approach will be to integrate information across a spectrum of biological complexity, with the 'system' ranging from an isolated microbe, to an individual infected host, and on to microbial and host populations. The evolution and spread of antibiotic resistance clearly involves a complex feedback between processes at the molecular and population levels for example, and an ability to link the molecular information emerging from functional genomics with the rich literature addressing host–pathogen (or host–vector–pathogen) systems from a population perspective will be essential for understanding and ultimately controlling infectious diseases.

Here we provide an outline of some of the experimental, theoretical and conceptual approaches that are involved in integrative systems biology and are considered in detail in subsequent chapters.

2 Data for systems biology: 'Omics, images and chemistry

A major impetus for the development of systems biology derives from technical advances associated with high-throughput sequencing [2] and chip-based systems [3, 4]. With the widespread availability of microarray formats for expression profiling, biologists whose primary focus was largely on the study of individual molecules or pathways were deluged with vast datasets comprising information on the simultaneous level of expression of every single gene in a cell or organism (the transcriptome) (see Chapter 2). While some simple clustering algorithms [5] provide an approach to analysing such datasets, it is clear that they contain a wealth of information that is not interpretable by conventional reductionist techniques. Analogous study of the total complement of proteins at a whole system level presents a greater technical challenge on account of the heterogeneity in their chemical and physical properties, but progress has been achieved by combining fractionation techniques such as two-dimensional gel electrophoresis with increasingly sophisticated mass spectrometry analysis [6, 7] (see Chapter 4). The ability to identify protein–protein interactions using yeast two-hybrid [8] and tandem affinity [9] purification systems has been particularly informative in mapping proteome networks (see Chapter 8). Analysis of protein-nucleic acid interactions [10, 11] at the level of transcriptional regulation generates an additional source of data that begins to link proteome and transcriptome information (see Chapter 4). Glycomic analysis based on mass spectrometry and nuclear magnetic resonance (NMR) techniques has provided insights into the further diversity generated by post-translational modification of proteins [12], and the same tools derived from physical chemistry allow quantitative analysis of the repertoire of small molecules that represent the cellular metabolome [13, 14] (see Chapter 5). At a higher level of complexity, metabonomic analysis provides an overview of metabolites in multicellular organisms, including the sharing of metabolite pools between host and microbe that is central both to commensal colonisation and to pathogen infection [14, 15] (see Chapter 10). Taken together, these 'omics datasets represent the starting material for the systems biologist, who faces the challenge of finding ways of maximising their integration and translation into usable information.

A second key source of data derives from imaging techniques. High-thoughput 'omics datasets are derived from analysis of biological systems at a population level, with differences between individual members of the population subsumed within an overall average. Technologies that derive data from single cells demonstrate that there is a significant underlying stochastic heterogeneity in the level of expression and in the spatial distribution of molecules within individual members of genetically clonal populations. In some cases these stochastic variations have been shown to be crucial in determining biological functions of the system [16], and an understanding at this level is a major component of systems biology.

Recent advances in fluorescent microscopy [17] have revealed an unprecedented degree of organisation and complexity in bacterial cells, despite their lack of membrane-bound cellular compartments. During the cell cycle many bacterial proteins localise to particular sites at specific times; understanding how such topological specificity is achieved is a fundamental question in cell biology. A recent example of proteins displaying previously unexplained dynamic protein localisation are the Spo0J/Soj proteins of *B. subtilis*, which are involved in chromosome segregation and transcriptional regulation. Using fluorescence microscopy Howard and colleagues [18] showed that Spo0J organises into compact foci associated with the nucleoid, while Soj undergoes irregular relocations from pole to pole or nucleoid to nucleoid. They propose that these irregularities are due in part to low copy number fluctuations: the relatively low numbers of the Spo0J/Soj proteins in a cell, together with the intrinsic probabilistic nature of their interactions, leading to large fluctuations in their dynamic behaviour. Stochasticity is vital for capturing the observed irregularity of the spatiotemporal protein dynamics for the Spo0J/Soj system.

The phenotypic tolerance to antimicrobial drugs associated with particular growth states of many microorganisms has also been shown to have a stochastic element [19, 20]. Integrating spatio-temporal information derived from single cell imaging with the type of information provided by high-throughput analysis of bulk populations is another central challenge for the systems biologist.

Biological systems are dynamic and observations recorded over time – particularly in response to some defined perturbation – provide critical information that is missing from a static analysis. Techniques for in-

duction of relatively simple perturbations include changes to the cellular environment, induction or repression of selected genes, and addition of small molecule inhibitors (see also Chapter 3) [21, 22]. The use of chemical modulators is particularly informative in the context of anti-infectives. Changes in bacterial gene expression profiles induced by exposure to known drugs allows mapping of characteristic response networks [23], facilitating screening for compounds with novel mechanisms of action (see Chapter 2). Advances in genome re-sequencing technologies [24–26] present exciting opportunities for a chemical genomics approach to rapid target identification based on an initial chemical lead (see Chapter 3). Starting with a compound (of known or unknown structure) which has activity against a whole microbe, the target can be identified by isolating resistant mutants and identifying the corresponding genetic changes. This represents a very attractive approach to integration of chemistry, functional biology, and genetics.

3 Making models

When describing a biological system we have to determine first the level at which we wish to study the constituent processes and interactions. Often this will be determined by the nature and quality of the experimental data: if the data are plagued by high error levels it may not be possible or even desirable to formulate a detailed mathematical or conceptual model. In practice, most biological models are hybrids containing qualitative and quantitative elements.

3.1 Qualitative systems approach

Biologists have always relied on models to conceptualise how organisms work. Such models can be purely verbal models or descriptions of biological structures or processes. In a qualitative approach one uses only the most coarse-grained information about the constituents of a biological system. In the context of the Krebs cycle, for example, we do not care about the three dimensional structure of the enzymes or substrate molecules and their molecular interactions. In general no attempt is be-

ing made at predicting quantitative responses of a system or at quantifying results [27]. Qualitative (including verbal) models of the same system are very difficult to compare; if different researchers propose their own verbal models for a biological process it can be extremely difficult to decide to what extent these models are similar or not. Moreover they make almost exclusively linear assumptions: i.e., they make statements of the type "if A increases then B decreases". Incorporating feedback into a verbal model, for example, can become enormously cumbersome.

3.2 Quantitative systems approach

In a quantitative approach, as many details of a system are ignored as is possible (generally by trial and error). Again, for example, molecular structures may be ignored, but instead of a purely qualitative description of interactions and processes a mathematical description or function is now chosen to represent the entities making up the system and the interactions among them [28]. The mathematical model now requires us to specify our assumptions explicitly and from the outset and, once these have been determined, mathematical or computational analysis of the model will allow us to study its change over time (see Chapters 7 and 11). This is then compared to experimental data. Depending on the question at hand or the experimental data available the mathematical models can be very abstract and generic, or directly targeted at a particular biological problem. In the former case it may be possible, for example, to investigate systematically the expected behaviour of a certain type of theoretical model. This can then be compared qualitatively against experimental data. Especially when there is little data available such an approach has been very popular. This type of approach has also been used extensively in theoretical physics where, for example, highly simplified models of magnetic materials have been studied to qualitatively reproduce experimental results [29].

If more detailed data are available, and if a statistical approach can be devised which allows us to estimate the parameters of a mathematical model from such data, then more detailed predictive modelling approaches become possible. Such approaches have been highly successful, for example, in modelling the immune response to human immunodeficiency virus (HIV) [30], or in developing very detailed models of the human

heart which can be used [31, 32], with some success, to model the effect of certain cardiovascular drugs (see Chapter 9).

Models of biological systems must, however, be understood not as realistic descriptions but as simplified representations of much more complicated entities. Almost all models will eventually be superseded by more sophisticated and more powerful models. In some areas – including for didactic purposes – even simple models retain their usefulness even if their limitations are known.

4 Networks

Molecular networks – in particular, metabolic, transcription regulation, and protein-interaction networks – offer the possibility of a coherent and consistent framework for the description of the whole complement of biological processes inside a cellular system [33] (see also Chapters 6 and 11). These networks have taken on a central role in computational systems biology. Statistical inference of networks [34–36], in particular co-expression networks estimated from microarray data, and the analysis of network structures have become important fields of research.

Networks can be described mathematically in terms of graphs (Fig. 1). Graphs occur in many different settings and as a result the theoretical description of graphs/networks has progressed independently in disciplines as varied as mathematics [37], computer science, statistical physics [38, 39], engineering and sociology. Integrating the different techniques developed in these disciplines and adapting them for the use in the modern life-sciences will allow us to analyse the increasing amount of network data currently being generated in systems biology [33].

One of the central features of natural networks is that they are highly heterogeneous: some nodes (whether genes, proteins or metabolites) have a large number of interaction partners, while most nodes interact with only a few other nodes in the network [38, 40]. This reflects some biologically intuitive relationships: we now know that some proteins are involved in many different processes and take an almost pivotal position in an organism's functional organisation (just like some highly promiscuous individuals – so-called super-spreaders – contribute to the spread of sexual

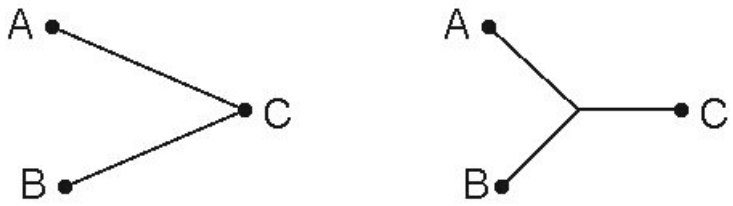

Figure 1.
Edges (left) connect nodes A, B and C. In this case there are direct pairwise interactions between nodes A and C and between B and C, but not between A and B. In the right part of the figure we show a hyper-edge which connects all three nodes. Interaction data collected from mass spectrometry surveys generally only allows us to construct such hyper-edges but not to determine pairwise interactions reliably.

transmitted diseases) [41]. This heterogeneity is further exacerbated by the modular architecture of biological processes: hierarchies and modules appear to be natural attributes of biological (and evolving) systems [42–44] (see Chapter 6). This, however, also poses considerable challenges to the simple models which have been so successful in the past. The complexity (and evolutionary contingency) of such detailed data pose considerable statistical challenges [45, 46] (see Chapter 10).

4.1 Protein interaction networks

Yeast two-hybrid (Y2H) [8], tandem affinity purification and mass spectrometry (MS) [9] have been used to map interactions among proteins (see also Chapter 8). We now have fairly extensive protein interaction data for *S. cerevisiae* [47–49] and partial data for *D. melanogaster* [50], *C. elegans* [51] and, more recently, two partial datasets for humans [52, 53]. There is also interactome data for three pathogens, *E. coli* [54], *H. pylori* and *P. falciparum* [55], with more data becoming available all the time (Fig. 2).This data has to be considered with great care, however: it is prone to false-positive and false-negative results (error rates of 40% have been suggested). Moreover, these networks are biased or skewed because of the methods used to detect them. Y2H appears to be the noisiest experimental technique while MS data are subject to bias in favour of interactions among highly expressed proteins and, if complexes are formed, cannot tell us which pair-wise in-

teractions exist within the clusters [56, 57]. These techniques provide mostly qualitative descriptions of what interacts with what, but can include quantitative data on the frequency of interactions or the strength of interactions. In terms of networks, they do not provide directional information about whether one or other partner is driving the interaction.

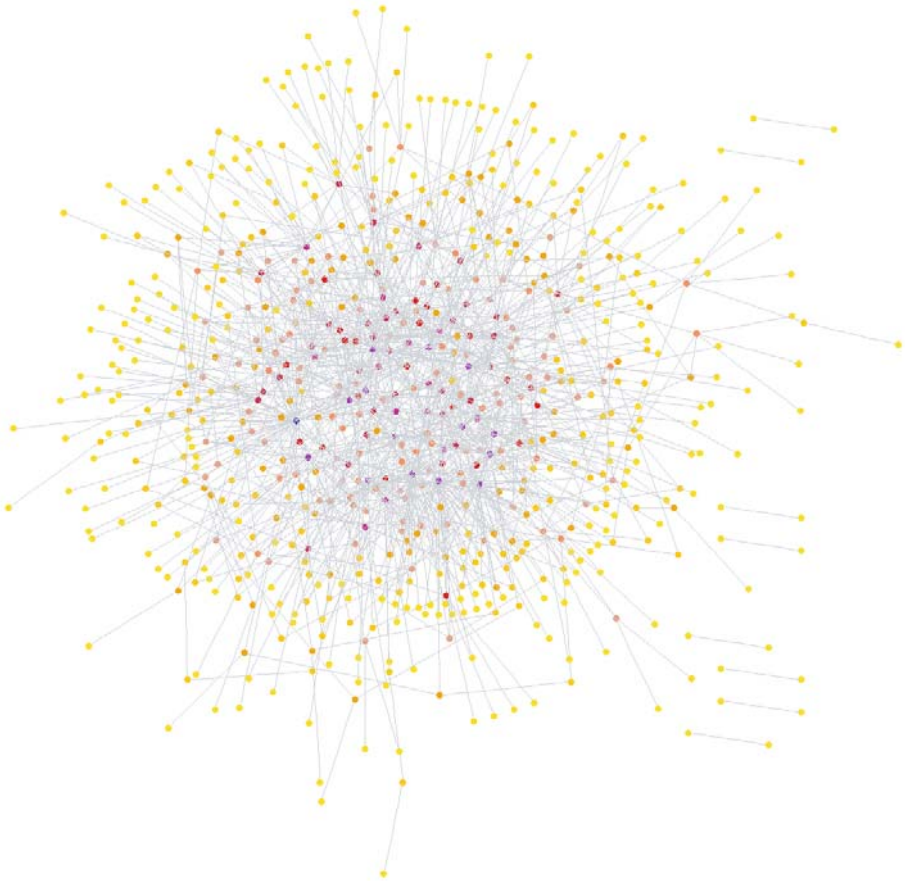

Figure 2.
Protein Interaction network (PIN) of *H. pylori*. This network is based on the available data in the database of interacting proteins (DIP) and thus does not represent the complete PIN. The heterogenous nature is however already apparent with most nodes having only one or two interaction partners, whereas a small number of nodes (so-called hubs) have many interaction partners.

Protein interaction network data (just as the other network data discussed below) offer a highly idealised and partial representation of cellular processes. They will change over three different time-scales: changes will occur at the evolutionary (between species), developmental and physiological levels. At the moment the data will at most allow us to resolve differences between species. This, as well as the fact that present experimental techniques may only capture a subset of the interactions has to be kept in mind.

4.2 Transcriptional networks

Initiation and regulation of gene expression is currently best understood at the level of transcriptional gene regulation. Transcription factors bind to regulatory elements upstream of the genes they regulate and these relationships can be depicted using directed graphs [58]. In addition to experimental and labour-intensive validation of transcription factors and their binding sites in genomes a growing number of *in silico* approaches are being developed and applied across all domains of life [59]. These use either co-expression patterns of genes to identify those that are presumably regulated by the same (or a similar) transcription-factor; or they employ linguistic/evolutionary arguments to find regulatory elements in sequenced genomes [60]. At present our data on transcriptional networks is also incomplete and suffers probably from ascertainment problems (i.e., researchers have focussed on their 'favourite' genes and mapped them with great care without gaining a global overview). For other processes of gene regulation there is even more rudimentary understanding of the involved mechanisms/molecules and the structure of the underlying networks. Transcriptional networks include both qualitative descriptions and quantitative data in terms of fold changes in gene expression, as well as information about direction of the interaction: e.g., there is a difference between gene A coding for a transcription factor which initiates transcription on gene B or gene B controlling expression of A. It is still frequently overlooked that transcriptional regulation encompasses only a tiny fraction of gene expression regulation. Incorporation of post-transcriptional and post-translational processes is only starting to be considered.

4.3 Metabolic networks

The whole complement of enzymes and substrates inside cellular systems (or whole organisms) are increasingly described in terms of metabolic networks [61, 62] (see Chapter 5). These are a straightforward conceptual development from the notion of individual biochemical pathways (such as the Krebs cycle) towards a more integrative perspective (see Chapter 7). To a certain extent the integrative analysis of metabolic networks has progressed furthest as biochemical pathways are relatively straightforwardly described quantitatively using the familiar Michaelis-Menten theory of enzyme kinetics [27, 28]. Metabolic networks contain both qualitative and quantitative information.

In metabolic networks we can choose whether we want the enzymes or the substrates to be the nodes in the network. Over the past few years the view to denote enzymes as nodes of the metabolic network has prevailed.

Considerable work has gone into characterising the structure, evolution and functional organisation of these networks (see Chapters 6 and 11). Very simple mathematical models of network growth give rise to networks with structural properties similar to those observed in molecular networks [38, 63–65]. These networks offer an attractive perspective on biological systems but it is important to keep in mind their present limitations: (i) present network data are incomplete [66] and it is difficult to extrapolate from incomplete network data to the true network; (ii) experimental – in particular high-throughput – methodologies are notoriously noisy and data may be unreliable; (iii) some interactions may be too short-lived or weak to be observed experimentally but nevertheless have profound physiological importance; (iv) molecular networks are generally described in terms of (necessarily) simplified mathematical models, such as static graphs. In reality, however, they are highly dynamic and responsive objects. Simple models are slowly but steadily becoming too simplistic to capture the complexity of biological processes [67].

5 Integrative systems biology

While networks generated by different techniques are currently viewed independently, linking these together in integrated models is a central goal of systems biology (see Chapters 10 and 11). Clearly protein interactions depend in the first instance on genes being transcribed and translated; initiation of transcription in turn requires transcription factors which are themselves proteins. Enzymes, of course, are also proteins and are required for the metabolism inside a cell just as metabolic products are necessary to keep the protein synthesis going. By integrating the different forms of data, it should ultimately be possible, for example, to predict the proteome from knowledge of the genome, and to use knowledge of the transcriptome to derive insights into the metabolome.

Two examples serve to illustrate some of the challenges that need to be addressed in moving towards these ambitious goals. One hypothesis put forward in the context of linking genome to proteome, is that proteins involved in interactions with multiple other proteins (highly connected 'nodes') will be subject to increased pressure in favour of evolutionary conservation. While this is intuitively attractive, statistical analysis of data on protein interaction networks and genome conservation in *S. cerevisiae* and *C. elegans* showed that it was not the case [45]. An association was identified, however, between the degree of evolutionary conservation of a protein and its level of expression within the cell. A second example concerns the relationship between transcriptomic data and essential function. The adaptive responses that pathogens undergo during infection are most readily studied in terms of changes in gene expression (see Chapter 12). It would seem reasonable to infer that the induction of a gene in response to a particular environment will relate in some way to its required function but a simple comparison of list of genes that are upregulated – for example, in the case of a mycobacterial pathogen entering a host phagocyte [68] – displays little or no overlap with a list of genes identified as essential for survival. In a recent study of the factors underlying fungal virulence (using *S. cerevisae* as a model system), we have found that inclusion of protein interaction data does allow us to begin to link expression and essentiality datasets (M. Stumpf, unpublished observations). The usefulness of molecular network data has now been demonstrated for a number of

different phenotypes, especially in *S. cerevisiae*; in light of such successes it seems natural to further explore whether it is possible to detect associations between network structures – rather than individual genes – and complex phenotypes. This would mean that rather than looking at individual genes or their protein products we would shift focus to the interactions directly. Given the lack of tangible success in mapping human genes underlying complex (disease) phenotypes, such a network centred approach ought to be worth considering.

6 New targets for anti-infective drug development

The initial impact of wide-scale pathogen genome sequencing has been to allow conventional charts of biochemical pathways to be annotated with gene names. Saturation mutagenesis tools have provided information on genes that are essential in particular growth media and, in some cases, under infection conditions. Systems biology aims to convert this static and informationally sparse framework into a dynamic network of nodes and fluxes. Quantitative models will highlight bottlenecks and nodes that are crucial for microbial viability and will distinguish between those at which a small or a large reduction in activity would be required for significant biological impact (see Chapter 7). The ability to input different types of data will allow models to be customised using information from genotypic data and from *in vivo* expression profiling to optimise for selection of targets that are appropriate in the context of existing drug resistance or in the context of phenotypic drug tolerance associated with latent tuberculosis and treatment of biofilm infections, for example. It can be anticipated that a systems biology framework will allow a rational approach to identification of synergistic drug combinations that will result in more rapid action and perhaps reduction in the evolution of resistance. Genetic experiments have shown that combining mutations which independently have no detectable impact on survival can result in 'synthetic lethality' [69, 70]. Similarly, it may be possible to identify drug combinations which result in a novel enhanced lethality by hitting two or more independent targets.

Systems biology may also help us in understanding infection processes in more detail. An illustrative outlook on what may be to come in the

future is provided by a recent study by Uetz et al. [71] who studied interactions among human proteins and herpes-virus proteins. If or when the enormous experimental problems can be overcome – there is as yet no reliable experimental technique which allows us to test for transient or weak interactions – then such studies give much more detailed insights into infection biology at the molecular level with a distinct focus on the physical interaction *per se*. If we are willing to speculate for a moment then such approaches harbour a host of exciting possibilities waiting to be explored: we may for example be able to study why different species have different susceptibilities to different infectious agents – Simian Immunodeficiency Virus (SIV) and HIV are good examples for the subtle impact of cross-species effects – or we may study whether the molecular interactions between *P. falciparum* and their human hosts and fly vectors, respectively, can be exploited for clinical purposes.

As models evolve, they will integrate increasingly diverse sources of data. This could include information from structural biology and functional biochemistry that relate to the 'drugability' of targets. Pathogen–host systems biology comes with an additional component as infectious disease biology can only really be understood in an ecological and evolutionary framework: pathogens compete for a potentially limited host population, while hosts in turn mount an immune response against pathogens and may even develop suitable strategies against pathogens. There are a host of beautiful examples of apparent host–pathogen co-evolutionary dynamics (for example between lizards and some species of *Plasmodium*) [72]. In addition we must consider the interaction between the host and the drug (see Chapter 9); host metabolism or modification of the drug will also influence the way it interacts with its target and the system as a whole. Every effect we study at the molecular or cellular levels may lead to complicated (and long-term) feedback processes at the population level. Thus host–pathogen systems biology has to be even more immodest than other branches of the fledgling discipline of systems biology: it encompasses all levels from molecules all the way up to epidemiological dynamics at the eco-system level.

References

1 FDA (2004) The Critical Path to New Medical Products. http://www.fda.gov/oc/initiatives/criticalpath/

2 Kan CW, Fredlake CP, Doherty EA, Barron AE (2004) DNA sequencing and genotyping in miniaturized electrophoresis systems. *Electrophoresis* 25: 3564–3588

3 McGall GH, Christians FC (2002) High-density genechip oligonucleotide probe arrays. *Adv Biochem Eng Biotechnol* 77: 21–42

4 van Steensel B (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat Genet* 37 Suppl: S18–24

5 Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 7: 55–65

6 Hernandez P, Muller M, Appel RD (2006) Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrom Rev* 25: 235–254

7 Liebler DC (2004) Shotgun mass spec goes independent. *Nat Methods* 1: 16–17

8 Fields S (2005) High-throughput two-hybrid analysis. The promise and the peril. *Febs J* 272: 5391–5399

9 Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* 24: 218–229

10 Ren B, Dynlacht BD (2004) Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol* 376: 304–315

11 Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E et al (2000) Genome-wide location and function of DNA binding proteins. *Science* 290: 2306–2309

12 Haslam SM, Gems D, Morris HR, Dell A (2002) The glycomes of *Caenorhabditis elegans* and other model organisms. *Biochem Soc Symp*: 117–134

13 Doherty MK, Beynon RJ (2006) Protein turnover on the scale of the proteome. *Expert Rev Proteomics* 3: 97–110

14 Nicholson JK, Holmes E, Wilson ID (2005) Gut microorganisms, mammalian metabolism and personalized health care. *Nat Rev Microbiol* 3: 431–438

15 Kitano H, Oda K (2006) Robustness trade-offs and host-microbial symbiosis in the immune system. *Mol Syst Biol* 2 doi: 10.1038/msb4100039

16 Howard M, Kruse K (2005) Cellular organization by self-organization: mechanisms and models for Min protein dynamics. *J Cell Biol* 168: 533–536

17 Yuste R (2005) Fluorescence microscopy today. *Nat Methods* 2: 902–904

18 Doubrovinski K, Howard M (2005) Stochastic model for Soj relocation dynamics in *Bacillus subtilis*. *Proc Natl Acad Sci USA* 102: 9808–9813

19 Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S (2004) Bacterial persistence as a phenotypic switch. *Science* 305: 1622–1625

20 Kussell E, Kishony R, Balaban NQ, Leibler S (2005) Bacterial persistence: a model of survival in changing environments. *Genetics* 169: 1807–1814

21 Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E et al (2005) A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* 307: 223–227
22 Manjunatha UH, Boshoff H, Dowd CS, Zhang L, Albert TJ, Norton JE, Daniels L, Dick T, Pang SS, Barry CE 3rd (2006) Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 103: 431–436
23 Boshoff HI, Myers TG, Copp BR, McNeil MR, Wilson MA, Barry CE 3rd (2004) The transcriptional responses of *Mycobacterium tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action. *J Biol Chem* 279: 40174–40184
24 Albert TJ, Dailidiene D, Dailide G, Norton JE, Kalia A, Richmond TA, Molla M, Singh J, Green RD, Berg DE (2005) Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nat Methods* 2: 951–953
25 Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380
26 Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J et al (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* 12: 1749–1755
27 Fell D (1996) *Understanding the Control of Metabolism*. Portland Press Ltd, UK
28 Murray JD (2001) *Mathematical Biology: An Introduction: Pts. 1 & 2*. Springer-Verlag New York Inc, USA
29 Fulde P (1995) *Electron Correlations in Molecules and Solids*. Springer-Verlag Berlin and Heidelberg GmbH & Co, Germany
30 Nowak M, May R (2000) *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press, UK
31 Crampin EJ, Halstead M, Hunter P, Nielsen P, Noble D, Smith N, Tawhai M (2004) Computational physiology and the Physiome Project. *Exp Physiol* 89: 1–26
32 Noble D (2006) Systems biology and the heart. *Biosystems* 83: 75–80
33 de Silva E, Stumpf MPH (2005) Complex networks and simple models in biology. *J Royal Soc Interface* 2: 419–430
34 Dobra A, Hans C, Jones B, Nevins JR, West M (2004) Sparse graphical models for exploring gene expression data. *J Multiv Analysis* 90: 196–212
35 Schafer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21: 754–764
36 Schäfer J, Strimmer K (2005) Learning large-scale graphical Gaussian models from genomic data. In: JF Mendes (ed.): *Science of Complex Networks: from Biology to the Internet and WWW (CNET 2004)*, The American Institute of Physics, USA
37 Bollobas B (1985) *Random Graphs*. Academic Press Inc (London) Ltd, UK
38 Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, UK

39 Newman MEJ (2003) The structure and function of complex networks. *SIAM Review* 45: 167–256

40 Alm E, Arkin AP (2003) Biological networks. *Curr Opin Struct Biol* 13: 193–202

41 May RM, Lloyd AL (2001) Infection dynamics on scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 64: 066112

42 Gagneur J, Krause R, Bouwmeester T, Casari G (2004) Modular decomposition of protein–protein interaction networks. *Genome Biol* 5: R57

43 Hallinan J (2004) Gene duplication and hierarchical modularity in intracellular interaction networks. *Biosystems* 74: 51–62

44 Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. *Proteins* 54: 49–57

45 Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MP (2005) Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. *BMC Evol Biol* 5: 23

46 Mazurie A, Bottani S, Vergassola M (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biol* 6: R35

47 Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141–147

48 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K et al (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180–183

49 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 98: 4569–4574

50 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E et al (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302: 1727–1736

51 Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T et al (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303: 540–543

52 Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437: 1173–1178

53 Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S et al (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 122: 957–968

54 Fribourg S, Romier C, Werten S, Gangloff YG, Poterszman A, Moras D (2001) Dissecting the interaction network of multiprotein complexes by pairwise coexpression of subunits in *E. coli*. *J Mol Biol* 306: 363–373

55 LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C et al (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438: 103–107

56 Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 22: 78–85
57 von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417: 399–403
58 Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31: 64–68
59 Qiu P (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem Biophys Res Commun* 309: 495–501
60 Boffelli D, Nobrega MA, Rubin EM (2004) Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 5: 456–465
61 Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 427: 839–843
62 Fell DA, Wagner A (2000) The small world of metabolism. *Nat Biotechnol* 18: 1121–1122
63 Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286: 509–512
64 Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol* 4: 51
65 Burda Z, Correia JD, Krzywicki A (2001) Statistical ensemble of scale-free random graphs. *Phys Rev E Stat Nonlin Soft Matter Phys* 64: 046118
66 Stumpf MP, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA* 102: 4221–4224
67 Stumpf MPH, Ingram PJ, Nouvel I, Wiuf C (2005) Statistical model selection methods applied to biological networks. *Proc Comp Systems Biol* 3
68 Stewart GR, Patel J, Robertson BD, Rae A, Young DB (2005) Mycobacterial mutants with defective control of phagosomal acidification. *PLoS Pathog* 1: 269–278
69 Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S et al (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295: 321–324
70 Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H et al (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294: 2364–2368
71 Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, Berger B, Rajagopala SV, Roupelieva M, Rose D, Fossum E et al (2006) Herpesviral protein networks and their interaction with the human proteome. *Science* 311: 239–242
72 Perkins SL (2001) Phylogeography of Caribbean lizard malaria: tracing the history of vector-borne parasites. *J Evol Biol* 14: 34–45

Progress in Drug Research, Vol. 64
(Helena I. Boshoff and Clifton E. Barry III, Eds)

Applications of transcriptional profiling in antibiotics discovery and development

By Hans Peter Fischer[1] and Christoph Freiberg[2]

[1]Genedata AG,
Postfach 254,
4016 Basel, Switzerland
<Hans-Peter.Fischer@genedata.com>

[2]Bayer HealthCare AG,
Pharma Research & Development
(current address:)
Niederrhein University of Applied Sciences,
Department of Chemistry,
Adlerstraße 32,
47798 Krefeld, Germany
<Christoph.Freiberg@hs-niederrhein.de>

Abstract

This chapter will review specific applications of microarray technology and related data analysis strategies in antibacterial research and development. We present examples of microarray applications spanning the entire antibiotics research and development pipeline, from target discovery, assay development, pharmacological evaluation, to compound safety studies. This review emphasizes the utility of microarrays for a systematic evaluation of novel chemistry as antibiotic agents. Transcriptional profiling has revolutionized the process of target elucidation and has the potential to offer substantial guidance in the identification of new targets. Microarrays will continue to be a workhorse of anti-infectives discovery programs ranging from efficacy assessments of antibiotics ('forward pharmacology') to drug safety evaluations ('toxicogenomics').

1 Introduction

Since Fleming's discovery of the antibacterial activity of penicillin in 1928, discovery efforts in antibiotic research were mainly based on random cell-based screening and on the modification of already established chemical structures with antibacterial activity. However, the traditional approaches to antibiotic discovery are increasingly challenged by bacterial pathogens that rapidly develop resistance to established drugs. Although classical approaches to anti-infective drug discovery are still being used, new technologies show promise to significantly accelerate the discovery and development of novel drugs that are required to keep up with the increasing incidence of drug resistance [1]. In this context, molecular profiling technologies that enable the highly parallel quantification of mRNA, proteins or metabolites in a bacterial cell have attracted significant attention. In this review, we focus on applications of mRNA profiling technologies, sometimes referred to as microarray or DNA chip technologies.

Microarray technologies have greatly benefited from the availability of whole genome sequence data. In 1995, the genomic DNA sequence of the bacterium *Haemophilus influenzae* was deciphered as the first genome of a cellular organism [2]. In the decade since then, the complete genomic information of the majority of medically relevant bacterial species has been made available. Today, hundreds of microbial genomes are publicly available and can be used for developing specialized expression profiling technologies. In parallel, microarray technology has advanced tremendously