Molecular Microbiology (2006) **62**(5), 1239–1250 doi:10.1111/j.1365-2958.2006.05453.x
First published online 16 October 2006

MicroReview

Integration of omics data: how well does it work for bacteria?

Sigrid C. J. De Keersmaecker,[1] Inge M. V. Thijs,[1] Jos Vanderleyden[1] and Kathleen Marchal[1,2]*
[1]*Centre of Microbial and Plant Genetics (CMPG) Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, and* [2]*ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.*

Summary

In the current omics era, innovative high-throughput technologies allow measuring temporal and conditional changes at various cellular levels. Although individual analysis of each of these omics data undoubtedly results into interesting findings, it is only by integrating them that gaining a global insight into cellular behaviour can be aimed at. A systems approach thus is predicated on data integration. However, because of the complexity of biological systems and the specificities of the data-generating technologies (noisiness, heterogeneity, etc.), integrating omics data in an attempt to reconstruct signalling networks is not trivial. Developing its methodologies constitutes a major research challenge. Besides for their intrinsic value towards health care, environment and industry, prokaryotes are ideal model systems to further develop these methods because of their lower regulatory complexity compared with eukaryotes, and the ease with which they can be manipulated. Several successful examples outlined in this review already show the potential of the systems approach for both fundamental and industrial applications, which would be time-consuming or impossible to develop solely through traditional reductionist approaches.

From omics data to networks, the seed of systems biology

Recent technological advances have dramatically changed our views on molecular biology. Whereas a few years ago each gene or protein was studied as a single entity, new, so-called 'omics' technologies [e.g. genomics, transcriptomics, proteomics, metabolomics and interactomics, recently reviewed by Joyce and Palsson (2006)] allow one to analyse large numbers of genes or proteins simultaneously (Greenbaum *et al*., 2001). As a result, a gene is no longer studied as an isolated entity but as being part of a complex network (see Box 3). In a systems biology approach, a cell is considered as a system that executes a genetic programme. Moreover, this system receives dynamically changing environmental cues and transduces these signals into the observed behaviour (i.e. change of phenotype or change of physiological response). The network mediates this signal transduction. A complete network thus consists of all components (e.g. DNA, RNA, proteins, metabolites) in a cell interacting with each other. Modelling the dynamic action of these networks to predict cellular behaviour is the ultimate goal of systems biology (Kitano, 2002; Arita *et al*., 2005). This requires at first gaining insight in the causal interactions between the molecular entities (see Box 3) by the reconstruction of the basic network structure (also referred to as 'scaffolds'; Joyce and Palsson, 2006) from large amounts of data (defined as top-down systems biology, see Box 1a). Such a basic structure is usually represented graphically, with nodes indicating the main players in the network (proteins, genes) and edges referring to the interactions between them (see Box 3). Based on such basic network structures, more detailed mechanistic models can then be compiled taking into account dynamic behaviour (referred to as bottom-up systems biology, see Box 1a). Although ideally a complete systems biology flow combines both approaches, in this review we emphasize the top-down approach as it forms the analytical framework for the more detailed bottom-up approach.

Top-down systems biology is increasingly relying on the integration of heterogeneous omics data. Indeed, using distinct data sources instead of a single high-throughput data set to reconstruct networks has several advantages. First, different omics data (e.g. genome sequence, transcriptome, proteome, interactome, metabolome) unveil distinct aspects of networks and

Accepted 20 September, 2006. *For correspondence. E-mail Kathleen.Marchal@biw.kuleuven.be; Tel. (+32) 16 32 16 31; Fax (+32) 16 32 19 66.

Box 1. Tools for data integration are based on either one of the following principles.

a. Top-down versus bottom-up inference (Palsson, 2002)

Bottom-up inference starts from a comprehensive model of known interactions between molecular entities as described in literature and curated databases. Such models can be used to simulate cellular behaviour or to predict the outcome of a perturbation experiment. Inconsistencies between observed data and simulations point towards deficiencies in the current network structure and outline hypotheses of novel interactions that can better explain the observations.

Top-down inference aims at reconstructing the interactions between molecular entities based on data only, hence without using prior knowledge on the network structure. Top-down inference is a data-demanding approach and is, given the current data availability, often underdetermined (i.e. the network that is reconstructed from the data is not unique, many equally likely solutions can explain the observations). However, the top-down inference can be made increasingly tractable by integrating data from different sources.

b. Sensitivity versus false discovery rate

High-throughput data are inherently noisy (low signal to noise ratio). When searching for biological relevant information from such noisy data, trading off between reducing the number of false positives (FP) and detecting a sufficient number of true positives (TP) is essential. The false discovery rate (FDR) is defined as the number of false positives (FP) on the total number of predictions (TP+FP), and should be kept as low as possible. Indeed, the higher the number of false positives, the more laborious downstream experimental validation will be. Sensitivity, on the other hand, determines the number of predicted true positives (TP) on the total number of positives [i.e. true positives and false negatives (FN)] and is an indication of the positives missed by the prediction method. When integrating different data sets, either a very conservative or a non-conservative prediction can be chosen depending on the final goal. A conservative prediction (low FDR, low sensitivity) requires that all data sources agree on this specific prediction (for instance, by taking the intersection of all predictions of the individual data sources). A non-conservative prediction considers all predictions supported by at least one data source (for instance, by taking the union of the predictions of the individual data sets). Most implementations reconcile both extremes.

c. Global versus query-specific driven analyses

Analyses in systems biology can be both global or query driven. With global analysis, we refer to a complete analysis of all data available without focus on any specific pathway. Such analyses are meant to discover global patterns or to gain a holistic view on the behaviour of an organism. Besides this, it is of utmost importance for molecular biologists to query data sets, usually a combination of public and own data, about their particular processes of interest. Query-based pattern discovery tools, for example, tools based on the concept of using prior knowledge in the Bayesian framework, make use of this principle.

d. Supervised versus unsupervised methods

Supervised learning requires a target variable (dependent variable) that is causally dependent on other variables (explanatory variable). Classification is an example of a supervised learning technique that is often used. For classification, the target variable is a binary class label which can be either positive (class 1, a set of proteins known to interact) or negative (class 2, a set of proteins most likely not interacting).

Based on a set of training points for which both the target variable and the explanatory variable are known, a mathematical model can be built to optimally predict the class membership (class 1 versus class 2) of these data points. In case of predicting protein interactions, the explanatory variable can be data on, for example, coexpression, and the target variable consists of predicting whether two proteins interact or not. Training themathematical model implies choosing the parameters of the model such that a maximum number of true predictions is made for the training set, while avoiding over-fitting (optimization). The more accurate the training set, the better the model will be. Examples of models commonly used for data integration in systems biology are support vector machines, Bayesian networks, neural networks, etc.

In unsupervised learning, all variables are treated in the same way, and no distinction is made between explanatory and dependent variables. There is no need for a target variable and unsupervised methods (such as PCA, NCA, clustering) typically discover patterns, such as clusters in the data sets.

When sufficiently experimentally verified data are available (e.g. the experimentally verified interactions between proteins), a target variable can be defined. In other cases, target variables are tedious to construct. In such cases, unsupervised methods can be used. In addition, they prevent biasing new discoveries towards prior knowledge.

e. In a sequential versus a concurrent way

In sequential analysis, one data source is used after the other, e.g. microarray data are first clustered and over-represented motifs are subsequently searched for in the clusters of coexpressed genes using motif detection.

Concurrent analysis implies searching both data sets simultaneously. Considering the previous example, this would entail the simultaneous search for clusters of genes and regulatory motifs. Indeed, the degree of coexpression determines the quality of the detected motifs but the presence of motifs also confirms the reliability of the detected cluster. As both data sets mutually confirm each other, concurrent analysis is inherently more powerful, but often requires more computational resources.

integrating them leads to a more complete insight into these networks. Second, experimental and biological noise in the individual data measurements can be so prohibitive that each data type alone has a limited utility. The integration of data from different sources provides an effective means to deal with the high noise level in the individual data sources by lowering the false discovery rate (FDR) and increasing sensitivity (see Box 1b). Despite the advantages of data integration, inferring networks through integration of these heterogeneous data remains mathematically non-trivial, as the number of independently acquired omics experiments is usually

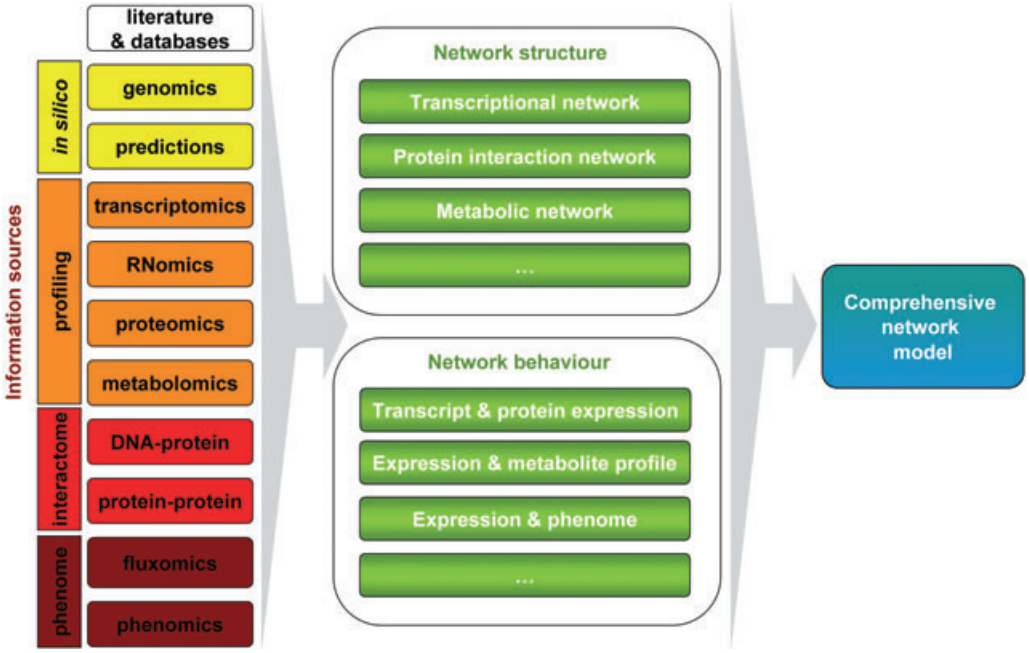

Fig. 1. Integration of omics data. Different information sources, i.e. omics data, literature and computational predictions, can be integrated to infer the structure of the transcriptional network, the protein interaction network or the metabolic network (network structure). Besides unravelling structures, omics data allow analysing the responses (mRNA, protein and metabolite profiles) triggered by each of these networks and studying their mutual relation (network behaviour). The ultimate goal in data integration will be to combine the transcriptional, protein interaction and metabolic network to construct comprehensive network models.

much smaller than the number of interactions to be inferred. This problem of under-determination is aggravated by the low signal to noise level of high-throughput data and the inherent stochasticity (see Box 3) of biological systems (Elowitz *et al*., 2002).

Although most of the network reconstruction approaches are generally applicable, many of the technological developments in this area of research are based on model organisms such as yeast (Ge *et al*., 2003; Hohmann, 2005) and *Caenorhabditis elegans* (Ge *et al*., 2003) while considerably less efforts focus on prokaryotes. Therefore, in this microReview we focus on prokaryotic omics data integration for network reconstruction (for the outline see Fig. 1). We give an overview of diverse methods which have been applied to infer from various omics data, structures of either the transcriptional, protein interaction or metabolic network and of studies that analyse the observed expression behaviour resulting from the concerted action of these networks (Fig. 1). These examples illustrate that systems microbiology is a challenging research domain, but with great potential for both fundamental and industrial applications, ranging from understanding bacterial adaptation and evolution to improved management of bacterial infections and designing better performing industrial strains.

Reconstruction of the network structure

Inference of the transcriptional network

Here, the causal interactions between transcriptional regulators and their target genes are considered at the transcriptional level. Expression-profiling experiments, usually performed by genome-wide microarray technology (Schena *et al*., 1995), measure changes in mRNA levels, upon mutation or in response to environmental changes. Initially, global methods (see Box 1c) were developed to reconstruct networks based only on microarray data (de Jong, 2002). Some of these methods, although conceptually interesting, are very data demanding [e.g. Bayesian networks and Dynamic Bayesian networks (Ong *et al*., 2002), see Box 3]. To reduce complexity, Gardner *et al*. (2003) focused on a few genes only (query-specific analysis, see Box 1c) instead of on the whole gene set. Using multiple linear regression (see Box 3) they explained the influences of perturbing the *Escherichia coli* DNA-damage response pathway (SOS response) on the steady-state mRNA levels and/or the dynamic (i.e. multiple time points) expression changes (Gardner *et al*., 2003; Bansal *et al*., 2006) of these query-specific genes. Within this framework, the authors could unravel the molecular interaction pathway of the SOS response mediated by RecA and

LexA, and elucidate the mode of action of the used perturbing pharmacological compounds.

The most promising methods for inferring transcriptional networks, however, combine microarray data with other data sources such as literature, DNA–protein interaction data (through ChIP-chip technology; Buck and Lieb, 2004) and/or *in silico* data on regulator binding sites (motif data) (Marchal *et al*., 2003; Tompa *et al*., 2005).

A first group of such integrative methods starts by compiling a model based on all known evidence. Kao *et al*. (2005), for instance, defined a linear static model for which the network connectivity information (interaction network between genes) was primarily extracted from the regulatory motif database RegulonDB (Huerta *et al*., 1998) and literature. They used network component analysis (NCA) (see Box 3) to deconvolute the relative contributions of the *E. coli* transcription factors in each condition using microarray data (Kao *et al*., 2005). This allowed the authors to demonstrate how metabolic enzyme activity exerts feedback regulation on the cAMP receptor protein (CRP) during carbon source transition.

A second group of integrative methods is data driven. In contrast to the previous class of methods which are model based, data-driven approaches are not confined to what is previously known. Data-driven integration of ChIP-chip and genome-wide expression data in bacteria thus far relies on intuitive approaches. One such approach identifies the relevant features (genes in this case) in each data set independently, and subsequently looks for features in the cross-section of these multiple data sets. Laub *et al*. (2002) were the first to use this approach to study the role of the *Caulobacter* master regulator CtrA in the dynamic expression changes during the cell cycle. In another approach, transcriptionally related data sets are combined in a sequential way (see Box 1e): *in silico* motif detection is applied to the targets identified by bacterial ChIP-chip assays analogous to the use of motif detection algorithms on clusters of coexpressed genes. In this way, hypotheses are generated that make the mode of regulation of the newly identified target genes more amenable to detailed molecular analysis. This was shown for the regulon of *Rhodobacter sphaeroides* photosynthesis regulators (Mao *et al*., 2005) and that of LexA, a master regulator of the SOS response in *E. coli* (Wade *et al*., 2005). Moreover, these studies rendered new global insights in prokaryotic transcription. For instance, the study on LexA binding revealed how in *E. coli*, unlike in eukaryotes, the transcription factor association with DNA is not controlled by DNA accessibility (Wade *et al*., 2005). Additionally, several non-consensus target sites seemed to be truly bound by the regulator *in vivo*, but not *in vitro* (Laub *et al*., 2002; Wade *et al*., 2005).

A more advanced method was developed by Zwir *et al*. (2005) who used an unsupervised method (see Box 1d) that groups co-regulated promoters in *E. coli* and *Salmonella enterica* based on gene expression data and a number of promoter-related features derived from amongst others RegulonDB, such as location and orientation of binding sites for a regulatory protein. Application of this method uncovered new members of the PhoP regulon and interactions of the PhoPQ regulon with other regulatory systems that were not discovered before.

Other methods, called module detection methods, aim at detecting from large genome-wide data compendia, gene sets which are coexpressed under a particular set of experimental conditions together with their corresponding regulators and transcription factor binding sites (Bar-Joseph *et al*., 2003; Segal *et al*., 2003; Tanay *et al*., 2004; Lemmens *et al*., 2006). Such so-called modules can be seen as a simplified representation of the transcriptional network. Most module detection methods are based on the concurrent analysis (see Box 1e) of diverse transcriptionally related data sources such as ChIP-chip, microarray and motif data and are currently being applied to prokaryotic data.

Studies on genome-wide transcriptional regulatory networks already revealed several intriguing structural and dynamic features of gene expression at a systems level (see Box 2).

Inference of the protein interaction network

Protein complexes can have a structural or an enzymatic function, or they can form regulator complexes involved in transcriptional regulation and, as such, link the protein interaction network to the transcriptional network. Defining the interactions among proteins is essential, because they play a role in virtually all biological processes. For example, exterior signals are transduced to the inside of a cell by protein–protein interactions of the signalling molecules. Exploring protein–protein interaction maps not only allows predicting protein function, based on the homology of the interacting proteins with characterized proteins in another species (Rain *et al*., 2001), but also revealing biological pathways. For instance, Noirot-Gros *et al*. (2002) could relate DNA replication with recombination/repair functions and signalling pathways. Moreover, protein interaction maps of human pathogens reveal clues on putative drug targets by providing insight into proteins functioning together during pathogenesis. Finally, genome-wide mapping of protein–protein interactions proved to be powerful in providing insight into the overall topology of a microbial network (Butland *et al*., 2005). The *E. coli* protein interaction network seemed to be ‘scale-free’ with a few highly connected ‘hubs’ but with most of the proteins having few interacting partners. Protein connectivity of a hub was positively correlated with the number

Box 2. Integration of omics data to gain insight in network topologies.

Below we give an illustrative example of how network reconstruction can be used to gain insight into fundamental biological questions. It shows how, based on the statistical analysis of inferred network topologies, basics of signal processing in bacteria and evolution of regulatory networks can be explained. Shen-Orr *et al*. (2002) were the first to reconstruct the topology of the *E. coli* transcriptional network based on data available in RegulonDB. According to their results, the global network structure has a modular composition, decomposable into several basic network motifs [feed-forward loops, single input module (SIM), etc.]. These network motifs are topologically distinct regulatory interaction patterns that are present more frequently in true biological networks than in random networks. These motifs are postulated to be the basic signal transduction elements, each with their own characteristic properties. A coherent feed-forward loop would be involved in filtering input signals by rejecting transient signals (Shen-Orr *et al*., 2002). A SIM is defined by a set of operons regulated by the same transcription factor that either function stoichiometrically or that are involved in the same metabolic pathway. Interestingly, small differences in the activation thresholds of the gene transcription by the common regulator can trigger a co-ordinated temporal response in the SIM output (Kalir and Alon, 2004). However, rather than ascribing the dynamic properties of the whole network to individual motifs, Dobrin *et al*. (2004) suggest a more complex model: single motifs would almost never occur in isolation but aggregate into homologous motif clusters that largely overlap with biological functions. In turn, these motif clusters coalesce into superclusters that define the global statistical properties of the whole network. Ma *et al*. (2004), who extended the transcriptional regulatory network of *E. coli* generated by Shen-Orr *et al*. (2002), found, however, that most of the motifs are connected to form a giant motif cluster instead of forming several small disconnected clusters.

By relating the network topology with microarray data obtained in several conditions, Balazsi *et al*. (2005) were able to propose a model for transcriptional signal transduction in bacteria. Complex environmental signals would be decomposed by the cell into elementary perturbations processed by individual origons. These represent regulatory subnetworks that originate at distinct classes of sensor transcription factors, i.e. an environmental signal-affected transcriptional subnetwork consisting of a set of operons regulated directly or indirectly by a single transcription factor that is not regulated transcriptionally by any other transcription factor. The final response develops by reassembling the elementary perturbations near the output of the network.

These regulatory networks and their topological properties have been the subject of different evolutionary studies. Babu *et al*. (2006), for instance, analysed conservation patterns of the *E. coli* transcriptional network (Shen-Orr *et al*., 2002) across 175 prokaryotic genomes. Natural selection appears to modify individual interactions to arrive at an optimal design for a given organism, rather than preserving whole blocks of pre-existing transcriptional interactions (Babu *et al*., 2006). Orthologous genes can become part of a different motif type in the regulatory network when a specific transcription factor in the species of origin is lost or gained, often in adaptation to a new environment. This might explain why distantly related organisms, but with similar lifestyle, tend to conserve network motifs.

of phylogenetic lineages in which its orthologue was still present (Butland *et al*., 2005).

Experimental protein interaction data are mainly based on two distinct technologies each of which detects complementary interaction types (Uetz and Finley, 2005). Yeast two-hybrid systems unveil physical interactions while mass spectrometry (MS) identification of proteins that coaffinity purify (co-AP) with a bait protein identifies stable complexes (functional interactions).

Besides these technologies, which give direct evidence, potential protein interactions can be predicted by using supervised techniques (see Box 1d) from other data sets as well, for instance, from microarray expression experiments (interacting proteins appear to be highly coexpressed), genome context [phylogenetic profiles (Wu *et al*., 2003), gene fusion, gene proximity], and the involvement of proteins in the same pathway, etc. (Snel *et al*., 2000; Janga *et al*., 2005) (as is shown by the many examples in yeast; Lu *et al*., 2005). However, so far only a fraction of the experimentally detected interactions could be readily predicted in *E. coli* using phylogenetic profiles or gene proximity criteria, e.g. only 3.4% of the interacting proteins were encoded by genes located within 500 bp of each other on the *E. coli* chromosome (Butland *et al*., 2005). Including other relevant features and integrating them in a single predictive model rather than using each of them separately could increase the predictions' reliability and coverage, though.

Inference of the metabolic network

Metabolic pathways have mainly been described by detailed deterministic or stochastic models consisting of non-linear ordinary differential equations (Michaelis Menten, S-systems, see Box 3) of which the structure is determined by previous knowledge on the pathway of interest (bottom-up models, see Box 1a). However, such models require knowledge of many parameters, such as enzyme kinetics, intracellular substrate concentrations, etc. The lack of experimental techniques to measure all these parameters at high-throughput level and the mathematical complexity of the problem restrict such models to small gene sets (Arita *et al*., 2005). Genome-scale models of microbial metabolism can be constructed by reducing these highly non-linear models to linear static systems (Covert *et al*., 2001a). *In silico* analysis of such metabolic networks uses a constraints-based approach. These constraints correspond to properties that limit the possible behaviours of the network and, thus, restrict the solution space. Examples of such constraints are thermodynamic constraints (irreversibility of reactions), capacities (such as maximum uptake rate of a transporter) and stoichiometries. Experimental data on flux levels, based on ^{13}C-labelling of substrates and isotopomer distribution analysis by two-dimensional nuclear magnetic resonance (2D NMR), gas chromatography-mass spectrometry (GC-MS) or liquid chromatography-mass spectrometry

Box 3. Glossary.

Interaction: In the context of this review an interaction refers to the connection between molecular entities, and corresponds to an edge in the graph-based representation. Depending on which molecular entities are referred to and the experimental procedures used to measure them (see Appendix S1), the interaction can either refer to a physical interaction (DNA–protein interactions in the transcriptional network, protein–protein interactions in the protein interaction network), or refer to a potential causal relation in the absence of direct physical measurements (for instance, the feedback of a metabolite on transcriptional regulation or the sequence of metabolic enzymes in a metabolic pathway in which the output of one enzyme serves as input for the subsequent enzyme). For an overview of different technologies to measure direct physical interactions between molecular entities we refer to Ge *et al.* (2003), Buck and Lieb (2004), Uetz and Finley (2005), Joyce and Palsson (2006) and *Supplementary material*.

Molecular entities: In the context of this review, mainly referred to as genes, proteins, metabolites. However, other molecular entities exist, such as small regulatory RNA molecules.

Multiple linear regression: Multiple linear regression attempts to model the relationship between two or more explanatory variables and a dependent variable by fitting a linear equation to the observed data.

Network: In the context of this review, it refers to the structural scaffold of a biological pathway or regulatory network. It consists of edges and nodes. Depending on which network is referred to, the nodes can be either proteins (protein interaction network), enzymes (metabolic pathway such as in KEGG) or regulators and targets genes (transcriptional network). Edges represent the interactions between the nodes (see also definition 'interaction'). A comprehensive network model refers to a structural network in which the protein interaction, transcriptional and metabolic network are combined into one global representation.

Network, Bayesian: A Bayesian network is a probabilistic graphical model that generally specifies the likelihood of an observation occurring, on the basis of the presence of various characteristics that are known or assumed to be associated with the observation according to prior information (Joyce and Palsson, 2006). It consists of a directed acyclic graph of nodes which indicate variables and edges which indicate probabilistic dependency relations among the variables. In the context of network inference, the nodes of a Bayesian network often represent the molecular entities and the edges the interdependencies between them (see also definition network).

Network component analysis (***NCA):*** Network component analysis (NCA) is a method to reconstruct transcription factor activities and connectivity strengths based on microarray data and partial network connectivity knowledge. In contrast to methods such as PCA (see also definition PCA), NCA takes advantage of partial network connectivity knowledge and is therefore able to better predict biologically meaningful signals. For example, if a regulatory node or factor is known from experimental evidence to have negligible or no effect on an output signal, then the corresponding edge may be removed or, equivalently, its weight may be set to zero. Such qualitative knowledge for a number of large biological systems is becoming available through high-throughput experiments. In contrast, traditional methods such as PCA depend purely on statistical assumptions and the resulting decomposition does not necessarily contain physically or biologically meaningful signals (Liao *et al.*, 2003).

PCA: Principal components analysis (PCA) is a linear transformation that transforms the data to a new co-ordinate system such that the greatest variance by any projection of the data comes to lie on the first co-ordinate (called the first principal component), the second greatest variance on the second co-ordinate, and so on. PCA can be used for dimensionality reduction in a data set while retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones.

Stochastic: From the Greek 'stochos'. A stochastic process is a process in which a transition depends on a previous state with a certain probability. There exists a chance that the transition does not occur. Probabilities and randomness are thus involved. This is in contrast to a deterministic process where the transition will most certainly occur.

S-system: S-systems (synergistic and saturable systems) are a set of non-linear ordinary differential equations which are in the context of this review used to represent biochemical interactions between genes. S-systems have unique mathematical properties that permit the investigation of rather large, realistic phenomena such as large networks. S-systems are derived from generalized mass balance equations, in which aggregates of all inputs and outputs are approximated by products of power-law functions. Justified by their mathematical derivation, S-system models can be designed directly from the topology of a network by including those and only those constituents in a power-law term that have a direct influence on a particular influx or efflux (i.e. the network needs to be known *a priori*) (Voit, 1992).

(LC-MS) (Emmerling *et al.*, 2002; Sauer, 2004; Wang *et al.*, 2006), can also be used to set these constraints. The optimal steady-state solution of the metabolic flux given a certain growth state, metabolite production or yield (defined as 'objective function') should lie in this constrained solution space and is detected by linear optimization (see Box 1d) (called flux balance analysis, FBA). Such models can successfully predict the effects of gene knockouts and allow quantitative dynamic simulation of substrate uptake, cell growth, etc. (Covert *et al.*, 2001a; Price *et al.*, 2004). These models and simulations hold great promise for future biotechnology applications. The benefits of systems biology for metabolic engineering, which uses knowledge about metabolic networks to rationally improve industrial useful strains (i.e. higher yield with lower fabrication costs), were recently exemplified by Teusink and Smid (2006) for lactic acid bacteria.

Analysis of the observed behaviour triggered by the different networks

Cellular behaviour results from the action of and interplay between the distinct networks. Therefore, besides inferring network structures at each of the individual molecular levels, studying and comparing the responses triggered by these different networks (e.g. transcriptome, proteome, metabolome) and their interrelation is of major interest.

Several studies correlate transcriptome and proteome, either on a limited (e.g. analysis of an L-threonine overproducing *E. coli* strain; Lee *et al*., 2003) or a more exhaustive scale. Corbin *et al*. (2003) for instance identified one-fourth of all *E. coli* proteins and discovered a positive relationship between protein and transcript abundance during exponential growth on glycerol.

Integration of expression data (at mRNA or protein level) with the metabolome often implies relating results from expression analysis to pathway databases. Becker *et al*. (2006) integrated *in vivo* generated phenotypic data on metabolic gene mutants and proteome data to search in a rational way for essential metabolic enzymes that represent potential antibiotic targets in *S. enterica*. Metabolic enzymes essential for virulence were identified using published metabolic mutant phenotypes (if the inactivation of an enzyme resulted in a significant attenuation of the wild-type virulence phenotype, it was considered essential), and comparative genomics (i.e. if metabolic genes appeared non-functional in other serovars, e.g. pseudogenes or missing, they were considered dispensable for virulence). Indirect information regarding additional enzymes with related biochemical functions was obtained by assuming that all non-redundant enzymes involved in pathways containing essential enzymes were also essential. To this end a graphical metabolic network model for *Salmonella* was constructed, based on the genome sequence (McClelland *et al*., 2001) and the EcoCyc and EcoSal databases (see Table S1). When combining the compiled list of enzymes with *in vivo* protein expression data that profile enzymes effectively present during *in vivo* infection, it appeared that most of these enzymes were non-essential during *in vivo* virulence. Moreover, the essential enzymes which were expressed were almost exclusively associated with a small subgroup of pathways. This drastically reduces the pool of possible new antibiotic targets. It is a nice example of how integrating different data sources can size down the number of candidate target enzymes to a more manageable number for further testing.

An increasing number of studies correlate patterns of gene expression with the production of specific metabolites. Measurement of metabolites gives information on how functional proteins act to transform energy and process materials. Krömer *et al*. (2004) studying lysine production in *Corynebacterium glutamicum* or Lafaye *et al*. (2005) studying the yeast sulphur pathway integrated metabolite profiles and metabolic fluxes with high-throughput transcriptomics or proteomics data, respectively, to identify correlations between expression and *in vivo* enzyme activity. These studies revealed that, depending on the growth conditions, the maximum flux of some enzymes correlated either positively (Krömer *et al*., 2004; Lafaye *et al*., 2005) or negatively (Lafaye *et al*., 2005) with maximum gene expression. For other enzymes, gene expression remained unaffected, despite increased fluxes, indicating that the metabolic capacities of these enzymes were not limited by their mRNA levels (Krömer *et al*., 2004). These studies demonstrate that the combination of different profiling techniques (metabolome, fluxome, transcriptome, proteome) provides detailed quantitative information on a biological network and can therefore help identify the key genes for strain improvement.

Need for a comprehensive network

Because of the computational complexity, efforts to infer the complete molecular network at all its levels in a concurrent way (see Box 1e) and to represent it in a comprehensive way are still under development. Covert *et al*. (2004) constructed the first integrated genome-scale model of the transcriptional and metabolic network in *E. coli* based on a compilation of known interactions. Metabolic interactions are based on a linear static model and regulatory interactions are imposed by logic statements that simulate the effect of regulatory processes over time. The model contains 1010 genes, of which 479 are regulators. Simulation parameters were estimated based on FBA (Covert *et al*., 2001b). Covert *et al*. used high-throughput experimental growth phenotyping [ASAP data (Bochner, 2003; Glasner *et al*., 2006), see Table S1 in *Supplementary material*] and gene expression experiments (see Table S1 in *Supplementary material*) to assess the biological relevance of their model. Despite the linearity of their model, the predicted growth phenotypes agreed with the experimental ones in 79% of the cases and the model could predict the gene expression data with 49% accuracy and 15% coverage (Covert *et al*., 2004; Barrett *et al*., 2005). The model also identified previously unknown components and interactions in the networks, systematically generating hypotheses to be tested in newly designed informative experiments.

Barrett *et al*. (2005) used the comprehensive integrated metabolic and transcriptional *E. coli* model of Covert *et al*. (2004) to assess the global characteristics of all functional states computed in the 15 580 growth conditions the network can exhibit. Each functional state was described by an activity profile that contains both the calculated expression state of each gene and the logical interactions between all transcription factors and the genes they are known to regulate. They showed by clustering and dimensionality reduction (PCA, see Box 3) of these profiles that the set of all possible network states has only a few dominant modes that are organized according to the terminal electron acceptor

and the availability of glucose or gluconate as carbon sources. Relatively few transcription factors (FNR, Arc, i.e. the global regulators or hubs) are required to explain these dominant modes.

Research challenges for data integration

Top-down systems biology aims at inferring network structures from high-throughput omics data. It offers the structural backbone of a more far-going systems biology approach in which a full mechanistic network model is aimed at that is able to predict the overall cellular behaviour.

A first prerequisite of successful systems biology is the generation of the data. Developments in high-throughput technology are still ongoing at increasing pace. Solving technical bottlenecks, among which the measuring of biochemical parameters at single cell level, constitutes a research challenge on its own as is outlined in the recent report from the American Academy of Microbiology (AAM) (Buckley, 2005).

Despite these technological issues, the amount of omics data is steadily increasing. Therefore, the need for computational platforms that generate comprehensive biological insights from these data becomes urgent. Lately much effort has been put into developing algorithms and tools for integrative network reconstruction. Available methods differ from each other in the way they approach the data [supervised versus unsupervised (Box 1d); global versus query specific (Box 1c); sequential versus concurrent (Box 1e)]. Developing such biologically relevant data integration tools remains one of the major future challenges in bioinformatics research: accounting for both the biological aspects and data-related aspects is a non-trivial task. At first, each tool relies on an underlying model that represents biological reality. Designing this model implies including sufficient complexity to capture in a comprehensive and intuitive way different aspects of the biological system (such as the different molecular levels and interactions between them), but it requires at the same time imposing sufficient conceptual simplifications to keep the solution computationally tractable. The probabilistic methods of Segal *et al*. (2003) are a nice example of such well-designed models.

Second, the algorithms that fit these models also have to take into account the specificities of high-throughput data. These data are inherently noisy. Although some consistent sources of variation can be removed in advance by proper pre-processing (van Hijum *et al*., 2003; Leung and Cavalieri, 2003; Engelen *et al*., 2006), a residual noise level due to biological stochasticity or experimental error will always remain. Algorithms should take into account estimates of the noise levels on each of the individual data sources and trace how they influence the reliability of the final predictions. Experiment design tools should exploit the information on how the noise propagates from the initial data source to the final result in order to predict which experiment can further improve reliability of certain predictions (Ideker *et al*., 2000).

Also, most high-throughput data are condition dependent and usually only give a snap shot of the molecular entities being present and/or interacting at a specific moment. High-throughput data are thus incomplete and exhibit many missing or unobserved data. As a result, only when data sources are perfectly synchronized in time or condition, they are expected to agree with each other. When extrapolating the information to other conditions, agreement between data sources can confirm an interaction but observing a conflict does not indicate a flaw. Most of the algorithms developed so far do not take into account such data directionalities during inference. These and many other data properties determine the specificities of the models and algorithms to choose. Therefore, close interaction with biologists and awareness of the intricacies of their data will be crucial to guide bioinformaticians in building appropriate and useful tools.

The downside of trying to build more and better tools is that they become less accessible to biologists. Biologists are overwhelmed with an ever-increasing number of tools with a lack of guidance on which one performs best. For instance, it appears that distinct methods dealing with the same problem give different results with few or no overlap (Lemmens *et al*., 2006). This does not necessarily implicate that (one of) the used methods do(es) not work but rather that depending on the properties of the tool, other aspects of the data sets are emphasized. Benchmarking methods before applying them on data to be analysed helps understanding the limits and possibilities of a tool. However, this is not trivial as no single biological network has yet been completely characterized. Comparison with a standard of known interactions allows one to assess the sensitivity of a method (recovery rate of true positives, see Box 1b) but does not penalize the presence of false positives. Moreover, as tuning algorithmic parameters usually comes down to trading off between sensitivity and detection of false positives (FDR), most algorithms have parameter settings in which the sensitivity is 100% at the expense of a very high false detection rate (FDR). The use of simulated data, although only a weak reflection of real biological data, remains important to get familiar with a tool (Albers *et al*., 2006; Van den Bulcke *et al*., 2006). So despite all the efforts in providing user-friendly platforms for data analysis (e.g. Thijs *et al*., 2002; Xia *et al*., 2005) extracting useful information by applying complex tools usually requires, besides some basic knowledge of the underlying statistics, lots of user-experience.

Another critical need for data integration is a centralized, curated database to enable rapid submission and retrieval of data and to expedite access to information from diverse areas of research (Buckley, 2005). As exemplified by Table S1 (*Supplementary material*), repositories of public prokaryotic high-throughput data, at the different molecular levels, are scarce. This in contrast to yeast, where already from the very beginning, data were publicly shared with the community. However, the current policy of most journals to release high-throughput data sets will get prokaryotes abreast. Another drawback is that most databases are biased towards well-studied pathways. The potential of systems biology lies in concerted research efforts that focus on all pathways in one particular organism (e.g. the International *E. coli* Alliance; Mori, 2004). Besides this, some important molecular entities are not yet sufficiently covered with high-throughput data. Recording the expression behaviour of regulatory small non-coding RNAs (sRNAs) (Masse *et al*., 2003; Vogel *et al*., 2003; Gottesman, 2005), for instance, will become as important as recording mRNA expression profiles in understanding how cells modulate gene expression. Densely tiled arrays having probes covering all intergenic and antisense regions of a genome, in combination with alternative labelling methods to cope with their small size and structure, hold great promise for small RNA detection (Hu *et al*., 2006). The same goes for profiling post-translational modifications. Although some methods allow detecting post-translational modifications (Van den Bergh and Arckens, 2005), these are not as widely adopted for prokaryotes as for yeast (Ptacek *et al*., 2005). Other omics data, such as metabolomics data, are not exploited to their full potential yet. Metabolome databases are still under development (Schauer *et al*., 2005) and available metabolite definitions are still non-reconciled with metabolite profiles (Kopka, 2006).

Once these data are generated, they should be stored in the proper layout for easy access. To this end, a set of quality standards and rules for cataloguing information are needed. Besides MIAME (Brazma *et al*., 2001), an established set of recommendations for the submission of microarray data, other standardizations are being developed, e.g. the Systems Biology Markup Language (SBML). For a complete overview on standards for systems biology we refer to Brazma *et al*. (2006) and the AAM report (Buckley, 2005).

Prokaryotes as inciters for data integration

Aiming at reconstructing a complete structural model of regulatory networks, top-down systems biology is based on the integration of data gained at different molecular levels. However, the complexity of biological systems in combination with the characteristics of omics data renders this data integration non-trivial. Currently developed methodologies are not yet able to take into account all of the intricacies of true biology. As prokaryotes are considered 'simpler' from the viewpoint of regulatory complexity, they are, as is illustrated by the examples in this review, powerful model systems to further develop systems biology tools, with impact beyond the microbial life. More importantly, systems microbiology will also have a value on its own, by paving the way in unravelling molecular mechanisms and with many applications in, for example, metabolic engineering and health care.

Acknowledgements

We apologize to those colleagues whose work we have been forced to omit through considerations of space. This work is partially supported by: (i) IWT project GBOU-SQUAD-20160, (ii) Research Council KULeuven: GOA-Ambiorics, IDO genetic networks; EF/05/007 SymBioSys, and (iii) FWO projects: G.0413.03 and G.0241.04. I.M.V.T. is aspirant of the Fund for Scientific Research, Flanders (FWO-Vlaanderen). We thank T. Van den Bulcke for critical reading and M. Fauvart for graphical assistance.

References

Albers, C.J., Jansen, R.C., Kok, J., Kuipers, O.P., and van Hijum, S.A. (2006) SIMAGE: SImulation of DNA-MicroArray Gene Expression data. *BMC Bioinformatics* **7:** 205.

Arita, M., Robert, M., and Tomita, M. (2005) All systems go: launching cell simulation fueled by integrated experimental biology data. *Curr Opin Biotechnol* **16:** 344–349.

Babu, M.M., Teichmann, S.A., and Aravind, L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* **358:** 614–633.

Balazsi, G., Barabasi, A.L., and Oltvai, Z.N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc Natl Acad Sci USA* **102:** 7841–7846.

Bansal, M., Gatta, G.D., and di Bernardo, D. (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* **22:** 815–822.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21:** 1337–1342.

Barrett, C.L., Herring, C.D., Reed, J.L., and Palsson, B.O. (2005) The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *Proc Natl Acad Sci USA* **102:** 19103–19108.

Becker, D., Selbach, M., Rollenhagen, C., Ballmaier, M., Meyer, T.F., Mann, M., and Bumann, D. (2006) Robust *Salmonella* metabolism limits possibilities for new antimicrobials. *Nature* **440:** 303–307.

Bochner, B.R. (2003) New technologies to assess genotype-phenotype relationships. *Nat Rev Genet* **4:** 309–314.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29:** 365–371.

Brazma, A., Krestyaninova, M., and Sarkans, U. (2006) Standards for systems biology. *Nat Rev Genet* **7:** 593–605.

Buck, M.J., and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83:** 349–360.

Buckley, M.R. (2005) *Systems Microbiology: Beyond Microbial Genomics. American Society for Microbiology. ASM Report.* Washington, DC: American Society for Microbiology Press.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., *et al.* (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433:** 531–537.

Corbin, R.W., Paliy, O., Yang, F., Shabanowitz, J., Platt, M., Lyons, C.E., Jr, *et al.* (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc Natl Acad Sci USA* **100:** 9232–9237.

Covert, M.W., Schilling, C.H., Famili, I., Edwards, J.S., Goryanin, I.I., Selkov, E., and Palsson, B.O. (2001a) Metabolic modeling of microbial strains *in silico*. *Trends Biochem Sci* **26:** 179–186.

Covert, M.W., Schilling, C.H., and Palsson, B. (2001b) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213:** 73–88.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92–96.

Dobrin, R., Beg, Q.K., Barabasi, A.L., and Oltvai, Z.N. (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics* **5:** 10.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002) Stochastic gene expression in a single cell. *Science* **297:** 1183–1186.

Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., *et al.* (2002) Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *J Bacteriol* **184:** 152–164.

Engelen, K., Naudts, B., De Moor, B., and Marchal, K. (2006) A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics* **22:** 1251–1258.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102–105.

Ge, H., Walhout, A.J., and Vidal, M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **19:** 551–560.

Glasner, J.D., Rusch, M., Liss, P., Plunkett, G., III, Cabot, E.L., Darling, A., *et al.* (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic Acids Res* **34:** D41–D45.

Gottesman, S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet* **21:** 399–404.

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

van Hijum, S.A., Garcia de la Nava, J., Trelles, O., Kok, J., and Kuipers, O.P. (2003) MicroPreP: a cDNA microarray data pre-processing framework. *Appl Bioinformatics* **2:** 241–244.

Hohmann, S. (2005) The Yeast Systems Biology Network: mating communities. *Curr Opin Biotechnol* **16:** 356–360.

Hu, Z., Zhang, A., Storz, G., Gottesman, S., and Leppla, S.H. (2006) An antibody-based microarray assay for small RNA detection. *Nucleic Acids Res* **34:** e52.

Huerta, A.M., Salgado, H., Thieffry, D., and Collado-Vides, J. (1998) RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res* **26:** 55–59.

Ideker, T.E., Thorsson, V., and Karp, R.M. (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac Symp Biocomput* **5:** 305–316.

Janga, S.C., Collado-Vides, J., and Moreno-Hagelsieb, G. (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res* **33:** 2521–2530.

de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* **9:** 67–103.

Joyce, A.R., and Palsson, B.O. (2006) The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7:** 198–210.

Kalir, S., and Alon, U. (2004) Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* **117:** 713–720.

Kao, K.C., Tran, L.M., and Liao, J.C. (2005) A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. *J Biol Chem* **280:** 36079–36087.

Kitano, H. (2002) Computational systems biology. *Nature* **420:** 206–210.

Kopka, J. (2006) Current challenges and developments in GC-MS based metabolite profiling technology. *J Biotechnol* **124:** 312–322.

Krömer, J.O., Sorgenfrei, O., Klopprogge, K., Heinzle, E., and Wittmann, C. (2004) In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome. *J Bacteriol* **186:** 1769–1784.

Lafaye, A., Junot, C., Pereira, Y., Lagniel, G., Tabet, J.C., Ezan, E., and Labarre, J. (2005) Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *J Biol Chem* **280:** 24723–24730.

Laub, M.T., Chen, S.L., Shapiro, L., and McAdams, H.H. (2002) Genes directly controlled by CtrA, a master regulator of the *Caulobacter* cell cycle. *Proc Natl Acad Sci USA* **99:** 4632–4637.

Lee, J.H., Lee, D.E., Lee, B.U., and Kim, H.S. (2003) Global analyses of transcriptomes and proteomes of a parent strain and an L-threonine-overproducing mutant strain. *J Bacteriol* **185:** 5442–5451.

Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., *et al.* (2006) Inferring transcriptional module networks from ChIP-chip-, motif- and microarray data. *Genome Biol* **7:** R37.

Leung, Y.F., and Cavalieri, D. (2003) Fundamentals of cDNA microarray data analysis. *Trends Genet* **19:** 649–659.

Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., and Roychowdhury, V.P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci USA* **100:** 15522–15527.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* **15:** 945–953.

Ma, H.W., Kumar, B., Ditges, U., Gunzer, F., Buer, J., and Zeng, A.P. (2004) An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res* **32:** 6643–6649.

McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., *et al.* (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413:** 852–856.

Mao, L., Mackenzie, C., Roh, J.H., Eraso, J.M., Kaplan, S., and Resat, H. (2005) Combining microarray and genomic data to predict DNA binding motifs. *Microbiology* **151:** 3197–3213.

Marchal, K., Thijs, G., De Keersmaecker, S., Monsieurs, P., De Moor, B., and Vanderleyden, J. (2003) Genome-specific higher-order background models to improve motif detection. *Trends Microbiol* **11:** 61–66.

Masse, E., Majdalani, N., and Gottesman, S. (2003) Regulatory roles for small RNAs in bacteria. *Curr Opin Microbiol* **6:** 120–124.

Mori, H. (2004) From the sequence to cell modeling: comprehensive functional genomics in *Escherichia coli*. *J Biochem Mol Biol* **37:** 83–92.

Noirot-Gros, M.F., Dervyn, E., Wu, L.J., Mervelet, P., Errington, J., Ehrlich, S.D., and Noirot, P. (2002) An expanded view of bacterial DNA replication. *Proc Natl Acad Sci USA* **99:** 8342–8347.

Ong, I.M., Glasner, J.D., and Page, D. (2002) Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics* **18** (**Suppl. 1**)**:** S241–S248.

Palsson, B. (2002) *In silico* biology through 'omics'. *Nat Biotechnol* **20:** 649–650.

Price, N.D., Reed, J.L., and Palsson, B.O. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2:** 886–897.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., *et al.* (2005) Global analysis of protein phosphorylation in yeast. *Nature* **438:** 679–684.

Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409:** 211–215.

Sauer, U. (2004) High-throughput phenomics: experimental methods for mapping fluxomes. *Curr Opin Biotechnol* **15:** 58–63.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., *et al.* (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett* **579:** 1332–1337.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34:** 166–176.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31:** 64–68.

Snel, B., Lehmann, G., Bork, P., and Huynen, M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28:** 3442–3444.

Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* **101:** 2981–2986.

Teusink, B., and Smid, E.J. (2006) Modelling strategies for the industrial exploitation of lactic acid bacteria. *Nat Rev Microbiol* **4:** 46–56.

Thijs, G., Moreau, Y., De Smet, F., Mathys, J., Lescot, M., Rombauts, S., *et al.* (2002) INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics* **18:** 331–332.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137–144.

Uetz, P., and Finley, R.L., Jr (2005) From protein networks to biological systems. *FEBS Lett* **579:** 1821–1827.

Van den Bergh, G., and Arckens, L. (2005) Recent advances in 2D electrophoresis: an array of possibilities. *Expert Rev Proteomics* **2:** 243–252.

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* **7:** 43.

Vogel, J., Bartels, V., Tang, T.H., Churakov, G., Slagter-Jager, J.G., Huttenhofer, A., and Wagner, E.G. (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* **31:** 6435–6443.

Voit, E.O. (1992) Symmetries of S-systems. *Math Biosci* **109:** 19–37.

Wade, J.T., Reppas, N.B., Church, G.M., and Struhl, K. (2005) Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes Dev* **19:** 2619–2630.

Wang, Q.Z., Wu, C.Y., Chen, T., Chen, X., and Zhao, X.M. (2006) Integrating metabolomics into a systems biology framework to exploit metabolic complexity: strategies and applications in microorganisms. *Appl Microbiol Biotechnol* **70:** 151–161.

Wu, J., Kasif, S., and DeLisi, C. (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19:** 1524–1530.

Xia, X., McClelland, M., and Wang, Y. (2005) WebArray: an online platform for microarray data analysis. *BMC Bioinformatics* **6:** 306.

Zwir, I., Huang, H., and Groisman, E.A. (2005) Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation. *Bioinformatics* **21:** 4073–4083.

Supplementary material

The following supplementary material is available for this article online:

Table S1. An overview of some important microbial databases is given.

Appendix S1. Overview of omics data.

This material is available as part of the online article from http://www.blackwell-synergy.com