Available online at www.sciencedirect.com
SCIENCE DIRECT®

Biological networks

Eric Alm and Adam P Arkin*

Recent advances in high-throughput methods have provided us with a first glimpse of the overall structure of molecular interaction networks in biological systems. Ultimately, we expect that such information will change how we think about biological systems in a fundamental way. Instead of viewing the genetic parts list of an organism as a loose collection of biochemical activities, in the best case, we anticipate discrete networks of function to bridge the gap between genotype and phenotype, and to do so in a more profound way than the current qualitative classification of linked reactions into familiar pathways, such as glycolysis and the MAPK signal transduction cascades. At the present time, however, we are still far from a complete answer to the most basic question: what can we learn about biology by studying networks? Promising steps in this direction have come from such diverse approaches as mathematical analysis of global network structure, partitioning networks into functionally related modules and motifs, and even *de novo* design of networks. A complete picture will probably require integrating the data obtained from all of these approaches with modeling efforts at many different levels of detail.

Addresses
Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS Calvin, Berkeley, CA 94720, USA
*e-mail: aparkin@lbl.gov

Current Opinion in Structural Biology 2003, **13**:193–202

This review comes from a themed issue on
Theory and simulation
Edited by Charles L Brooks III and David A Case

0959-440X/03/$ – see front matter

DOI 10.1016/S0959-440X(03)00031-9

Abbreviations

ChIP	chromatin immunoprecipitation
FRET	fluorescence resonance energy transfer
MS	mass spectrometry
PDB	Protein Data Bank

Introduction

Biological interactions at many different levels of detail, from the atomic interactions in a folded protein structure to the relationship of organisms in a population or ecosystem, can be modeled as networks. We focus on molecular interaction networks, which we define as a set of nodes, representing metabolites, genes or gene products, and a set of directed or undirected edges, representing the interactions between them (either direct physical interactions or functional associations). In particular, three nondisjoint molecular interaction networks have been the focus of most recent theoretical studies: the protein–protein interaction network; the transcriptional regulatory network; and small-molecule metabolism. Interaction trapping, FRET and MS methods make the determination of protein–protein interactions perhaps the easiest and most direct of the three to assay for. As a result, new data are arriving at an unprecedented rate. On the other hand, the available data are both incomplete and distorted by a large fraction of false positives, as well as false negatives [1••]. As for transcriptional regulatory networks, a comprehensive (but still incomplete) picture is available for a small number of organisms (*Escherichia coli* and *Saccharomyces cerevisiae*), constructed primarily through years of genetics and biochemistry rather than by high-throughput approaches. Because the methods used to elucidate these networks were focused, values can sometimes be placed on edges indicating whether and to what extent the link represses or activates transcription. A good deal of biochemistry is often elided from these networks, however, such as how the activities of transcription factors are controlled by post-translational modification, complexation and degradation. Metabolic networks, again deduced through countless person-hours of studying individual enzymes, are perhaps the most complete of the three, but the complexity of these networks is greatly increased by the presence of ubiquitous feedback loops between enzymes and metabolites, which makes direct stoichiometric analysis more difficult. Such feedback loops play a role in increasing the complexity of protein–protein and protein–DNA interaction networks as well, demonstrating what is turning out to be a recurring theme in biological modeling: the more we know, the more complex it is — even as some overarching principles become clearer.

New experimental techniques expedite network elucidation

High-throughput genome sequencing projects have made available nearly complete genetic parts lists for almost 100 organisms; however, they don't tell us (directly anyway) which metabolites and micrometabolites are present, which molecular structures/organelles are likely to form, the possible binding states of DNA regulatory regions, or all the possible splicing variants and post-translational modifications of proteins. Moreover, relationships between these components must be identified biochemically. Progress toward this goal has come in the form of high-throughput approaches to identifying the connections within the protein–protein and protein–DNA interaction networks. In this section, we briefly review some of these experimental approaches, the data from which are used in many of the theoretical studies discussed later.

Four independent studies provided the first glimpse of the global protein–protein interaction network in *S. cerevisiae* [2,3,4•,5•]. The first two utilized yeast two-hybrid technology, in which a bait protein fused to a DNA-binding domain is used to attract a prey protein fused to a transcriptional activation domain, resulting in expression of a reporter gene [2,3]. Because output of a reporter gene is measured instead of direct binding, transient associations can be detected. However, yeast two-hybrid screens are particularly ill suited to identifying multiprotein complexes, especially when binding is highly cooperative. The second two studies used MS-based screens designed to detect such protein complexes [4•,5•]. In the MS-based assays, bait proteins are tagged and potential complexes are purified from cellular lysate using affinity chromatography. Individual components are then isolated by SDS-PAGE and identified by MS. A surprising result from all four of these studies is that there is less overlap between the methods than perhaps anticipated and far from complete coverage of previously documented protein interactions, suggesting that our knowledge of the protein–protein interaction network in yeast is not yet saturated. To address these issues, Bork and colleagues [1••] compared these different high-throughput methods with several other sources of interaction data: correlated mRNA expression; *in silico* methods based on co-occurrence of related genes in operons, gene fusions and phylogenetic profiles [6–10]; and genetic interaction data (systematic genetic analysis, discussed below) [11]. They noted that, even for filtered yeast two-hybrid data, a technique achieving reasonable specificity in their test, probably about 50% of the predicted interactions are false positives. To improve the accuracy of predictions from these sources, they examined interactions predicted by more than one technique and found that a higher degree of accuracy comes at the expense of coverage; only about 3% (2400) of the 80 000 predicted interactions are predicted by more than one method. The same study estimates at least 30 000 total interactions in yeast.

Although the yeast two-hybrid and MS studies provide the largest currently available protein–protein interaction data sets (apart from literature-culled databases such as MIPS [12], YPD [13], BIND [14] and DIP [15]), two emerging technologies promise large complementary data sets, each with specific advantages over those previously discussed. The first new technology, proteome chips, reported preliminary results identifying new calmodulin-binding proteins, as well as proteins that interact with various phospholipids [16]. In this protein array technology, tagged proteins are immobilized on a chip and can then be probed for binding activity or, potentially, for any other biochemical activity of interest. The second method, called systematic genetic analysis (SGA), involves crossing two comprehensive libraries of nonlethal single-gene knockout strains of *S. cerevisiae* and identifying double-gene knockout strains that display a 'synthetic lethal' or no-growth phenotype [11]. Although the initial study considered only eight query genes, a network of over 200 genes was uncovered, many of which had previously unknown function. Because the yeast deletion strains used in this study were constructed by replacing the deleted genes with molecular 'bar codes', it may be feasible to study every possible double mutant using high-throughput parallel growth assays. The synthetic lethal phenotype described here is strong evidence of a physiologically relevant interaction between the two genes, but the trade-off is that the mode of interaction (simple binding or more indirect functional interaction) cannot be determined without further experimentation.

Whereas the techniques mentioned above primarily probe the protein–protein interaction network, the protein–DNA interaction network has been the subject of large-scale investigation using chromatin immunoprecipitation (ChIP) chips. In a groundbreaking study by Young and co-workers [17••], ChIP chips were used to identify likely binding sites for most of the known transcriptional regulators in yeast. By constructing strains in which specific regulatory proteins were tagged with an epitope for screening in chromatin immunoprecipitation assays, they recovered promoter sites bound to each of the 106 regulators after growth in rich media. These sites were then identified using DNA microarrays constructed from noncoding regions of the yeast genome. In addition to providing an outline of global transcriptional regulation in yeast, the results from Young and co-workers present a quantitative estimate of the amount of combinatorial regulation in yeast, a comprehensive set of DNA regulatory motifs and new insights into several well-studied biological processes. Of course, there are still gaps in our understanding of this network (presumably to be filled by more traditional genetic studies), for example, to what extent do the protein–DNA interactions activate or repress transcription, and under what conditions?

Biological networks are similar in structure to other complex networks

The experimental techniques discussed above lead to collections of 'interactions' between various biomolecules. These interactions have few, if any, quantitative labels on them, a high error rate and, in most cases, little cellular context. They do, however, provide an overview of the global structure or topology of these important cellular networks. Studies of complex networks as diverse as the World Wide Web and the scientific co-authorship network have uncovered unexpected nonrandom global organizational patterns. More recently, similar topological features have been reported for biological networks, specifically metabolism. Metabolic networks, at least for several model organisms, represent the most complete picture available of any molecular network and therefore are an ideal subject for the study of large-scale network

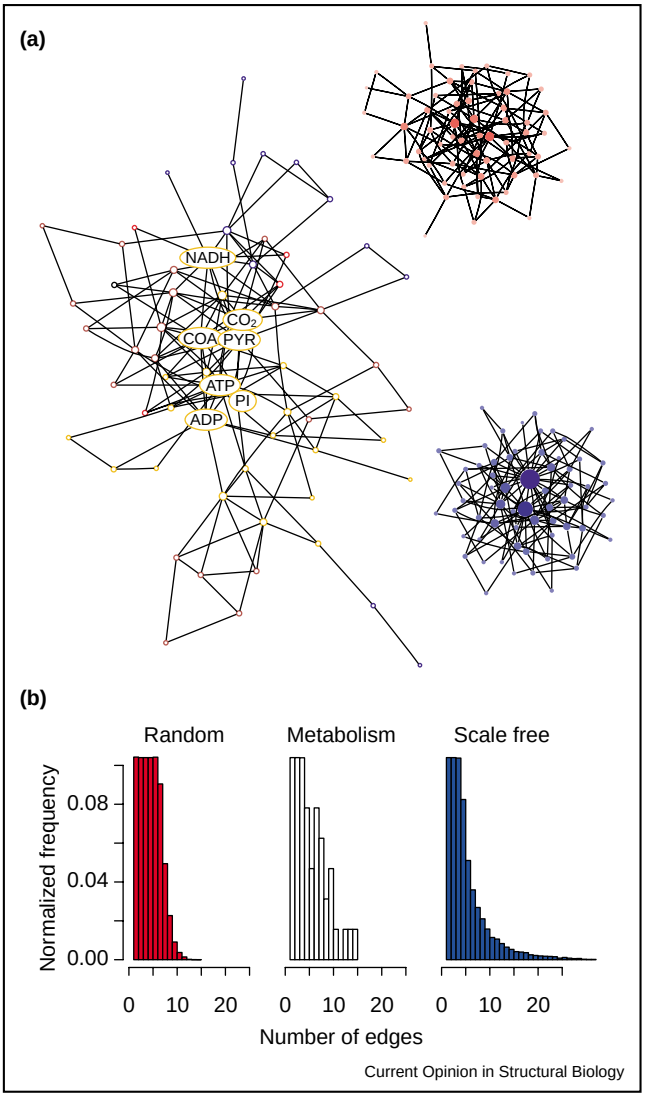
properties. In the studies described below, however, they are rarely represented in their full form, which requires a detailed account of the stoichiometry of reactions and feedback between enzymes and metabolites. In these networks, nodes represent metabolites, and the edges between nodes are directed and represent enzymatic reactions that utilize one of the nodes as a substrate and produce the other as a product. Empirical studies of metabolism in 43 organisms (not completely independent, as most of these networks were deduced by sequence comparison to *E. coli* and yeast, and are highly conserved evolutionarily) by Barabasi and co-workers [18–20] reveal that the metabolic networks in these organisms have several features in common with other nonrandom networks: metabolic networks are scale free; metabolism is a small-world network of fixed diameter; and metabolic networks are modular. Each of these findings is discussed below and some example networks are shown in Figure 1.

In the terminology of critical phenomena, scale free describes a system at the point of phase transition, whereby correlation functions of the order parameter have the form of a power law. In the context of scale-free metabolic networks, the probability of finding a node with k edges can be approximated by $P(k) \propto k^{-a}$, compared to the Poisson distribution expected for a random network. Essentially, this means that there are a small but finite number of highly connected nodes in the system. It has been pointed out that the occurrence of many genomic components follows a power-law distribution, including protein families, superfamilies, folds, short DNA words and even pseudogene families [21]. A surprisingly simple model of network growth proposed by Rzhetsky and Gomez [22] is sufficient to explain the abundance of this type of distribution in biological systems and is discussed in the following section.

In small-world networks, any two nodes can be connected through a much shorter path than would be expected in a random network of similar size and number of connections. Popular examples of small-world networks include the network of actors co-starring in films, the majority of whom can be connected to Kevin Bacon in fewer than four steps, and the mathematical co-authorship network centered around Paul Erdös. Such networks are characterized by their diameter, defined as either the maximum or average number of edges separating any two nodes. Metabolic networks were found to be small-world networks and, additionally, the network diameter does not appear to vary between different organisms.

Modularity in complex networks such as metabolism is somewhat more difficult to define precisely. Conceptually, modules refer to groups of genes that perform a discrete function separable from the rest of the system [23]. In a recent paper, Barabasi and co-workers [20]

Figure 1



Metabolism is a scale-free network. Central metabolism in *E. coli* is represented as an undirected graph for simplicity. Nodes represent metabolites, and two nodes are connected by an edge if there is a reaction in which one is consumed as a substrate and the other yielded as a product. Stoichiometry data for *E. coli* metabolism were obtained from the web site of B Palsson, University of California, San Diego (http://gcrg.ucsd.edu/). **(a)** Central metabolism is shown as an undirected graph (open circles); the most highly connected nodes (ADP, ATP, CO_2, coenzyme A [COA], NADH, phosphate (inorganic; PI), pyruvate [PYR]) are labeled. Randomly connected (red circles) and scale-free (blue circles) networks of similar size and number of connections are shown for comparison. The number of edges connected to each node is indicated by the node size (larger means more edges). **(b)** The fraction of nodes with a given number of edges is shown for three types of networks: randomly connected networks, scale-free networks and the metabolic network shown in (a). The frequencies for random and scale-free networks are computed by averaging over an ensemble of 500 networks of similar size and topology. Both the scale-free and metabolic networks have a small, but finite number of nodes with over ten edges, whereas the randomly connected graphs have almost none.

address the apparent paradox that metabolic networks appear to have characteristics of both modular networks (in which each node should have about the same number of links) and scale-free networks (in which most pathways are linked through a few highly connected 'hubs'). Using the clustering coefficient, defined as $C_i = 2n_i/k_i(k_i - 1)$, where n_i is the number of links between the k_i nearest neighbors of node i, as a measure of modularity, they find that metabolic networks in all organisms examined are modular in addition to being scale free. To resolve the paradox, they define a new class of networks called 'hierarchical networks', which possess both scale-free and modular properties, and provide a simple algorithm for generating such networks that may resemble the way networks are constructed during the evolutionary process.

Such empirical findings are intriguing because they provide hope that there are common design principles shared between biological systems and engineered systems, in the latter of which network topology is intended to provide special properties, such as robustness or quick communication between nodes. However, several key differences between some living and nonliving networks suggest that caution should be taken when making generalizations between these distinct classes of systems. First, purely topological studies neglect the identity of each node. In some networks, such as the Internet, all of the nodes perform the same function (i.e. to route IP packets), but, in metabolic networks, each node represents a distinct chemical species. Whereas the Internet might function similarly if individual nodes were rewired while keeping the same overall topology, metabolic reactions are highly specific and edges cannot, in general, be swapped because of additional constraints such as conservation of mass (in fact, a more realistic model of the Internet that takes account of differences in router connectivity and speed suggests that similar constraints may apply). Doyle has proposed that such networks are not actually free of a characteristic scale, but are composed of many nodes of different types organized into modular and hierarchical structures, and has suggested the term 'scale rich' to describe them (J Doyle, personal communication). From a biochemical perspective, the connectivity of the most highly linked nodes, water and ADP/ATP, can be understood in terms of the large number of hydrolysis and energy-utilizing reactions, and the convenience of having a common energy carrier and biosynthetic building block, and may have little to do with the benefits endowed on the cell by maintaining a specifically scale-free network topology. Second, the concept of small-world networks tends to overlook the stoichiometry inherent to biochemical reactions. Watts and Strogatz [24] demonstrated that a relatively small number of randomly connected edges can reduce a mostly linear, large-world network to one with small-world properties. In metabolic networks, these path-length-reducing edges can come in the form of cofactors that connect seemingly unrelated reactions. For example, we can reduce the glycolytic pathway distance from glucose to pyruvate to two steps by allowing links through cofactors such as ADP, although at least nine distinct enzymes are needed to produce pyruvate from glucose. Although we expect such studies will ultimately reveal deep connections between evolved and engineered networks, taken together, these differences advise against overgeneralizing between these two distinct classes of networks.

Biology from a network perspective

There are different philosophies as to how network information can be used to increase our understanding of biological systems. In the above studies, network topology data were used to generate new hypotheses about how systems are organized. A complementary approach involves taking existing hypotheses and using the extensive network data to either support or reject them. The second approach, reformulation of old questions from a network perspective, is the focus of this section and the next. In particular, the relationship between the evolution of genes and the networks they constitute has long been the subject of speculation by theorists. With the availability of large-scale quantitative data on the structure of molecular networks, it is possible to pose specific questions about the role of network structure in the evolutionary process and the role of evolution in shaping network structure. Several recent studies highlight the power of this new approach to understanding biology.

Rzhetsky and Gomez [22] examined the scale-free (power-law) distribution of protein domain types by constructing a simple, biologically plausible model of network evolution that reproduces this distribution. In their model, each gene (or gene product) can have an upstream or downstream domain or both, and upstream domains interact directly with downstream domains of the same class. Domains of existing classes grow by duplication, whereas innovation of new domain classes occurs at a constant rate. Their simple model successfully reproduces the scale-free distribution of domain types observed in the genome sequences of *E. coli* and yeast. Parameterizing their model with data from these two organisms allowed them to address an outstanding question in genomics: what is the total number of distinct domains in an organism's genome? Based on their model, they estimate at least 4600 domains for *E. coli* and over 12 900 domains for yeast.

Combining complete genome sequence data with an outline of protein interaction networks, it is possible to examine the effect of network structure on the rate of protein evolution. Feldman and co-workers [25] examined the rates of evolution of a large set of genes shared by two evolutionarily distant eukaryotes (*S. cerevisiae* and *C. elegans*) by constructing sequence alignments and comparing the average number of substitutions per site

for each gene with that of its interacting partners. They found that proteins that interact (as predicted by yeast two-hybrid and MS methods) tend to evolve at similar rates. After evaluating several alternative hypotheses, they conclude that interacting proteins evolve at similar rates as a result of co-evolution, providing the first quantitative picture of the overall rate and frequency of this important evolutionary process. In addition, they found only a weak correlation between the number of interactions made by a protein and its rate of mutation. This weak correlation may reflect a deficiency in the accuracy of the interaction data or suggest that other factors play a role in determining the rate of evolution.

A role for structure?

To what extent can structural information help clarify the role of individual nodes and the relationships between them in these networks? A recent study by Park *et al.* [26] used protein structure information to add value to nodes in the protein–protein interaction network. After taking a census of interactions between protein domains in the PDB, they conclude that most domain families interact with only a few other families, whereas a small number interact less specifically. Based on their findings, it may be possible to predict likely interaction targets for proteins of known structure, add confidence to predictions from other methods or even to predict structural assignments for proteins that interact with partners of known structure. Edwards *et al.* [27] used data from solved protein complexes to estimate the accuracy and even improve the quality of predictions of protein–protein interactions obtained via high-throughput techniques.

Studies focused on small-molecule metabolism demonstrate how structural information can be used to directly test hypotheses concerning the origin of metabolic pathways. Initial studies have focused primarily on the structural composition of metabolic pathways [28,29]. Because evolution of conserved metabolic pathways presumably occurred very early in the history of life, structural information is important because sequence similarity may be difficult to detect. These studies set the stage for more recent work that aims to use network information together with sequence and structural homology to select between two leading models of pathway evolution [30,31].

Modules and motifs

That biological systems are modular is not a new idea. In fact, examples of modularity in biological systems were recognized at least as early as the cell theory proposed over 150 years ago by Schleiden and Schwann. Examples of biological systems with modular components include the organization of the bacterial gene regulatory network into operons, and the modular organization evident during plant and animal development. We have thus come to expect a similar modular organization at the molecular level and many individual examples of such modules (by some definition) have already been documented.

In the past, sequence homology allowed us to deduce the molecular functions of a gene based on its inclusion in a larger protein family. Recent research has been directed at understanding the systems-level function of a gene by identifying the module or modules to which it belongs. An obstacle to elucidating the modular structure of molecular networks is the lack of a precise definition of what constitutes a module in this context. The clustering coefficient described above presents a mathematically precise definition of modularity in metabolic networks, but it is unclear how this definition helps a biochemist to understand the function of these networks. By contrast, Hartwell *et al.* [23] present a biologically relevant definition of modules as discrete units of function separable from the rest of the system, but this definition lacks the precision to unambiguously partition a network of genes into modules. As a result, the studies below each use a slightly different operational definition of modularity, although all are based on the co-regulation of gene expression within modules.

Identifying modules

Systematic high-throughput data acquisition provides an opportunity to unravel gene regulation networks in eukarya, and in bacteria and archaea. In yeast, functional relationships can be inferred from the numerous available collections of gene expression data. In bacteria and archaea, the conservation of co-regulated genes in operons, together with the large number of complete genome sequences, provides a similar opportunity to infer functional relationships, provided orthologous genes can be identified across several species.

Barkai and co-workers [32•] tackle the problem of identifying co-regulated genes in yeast, in which combinatorial regulation can make expression patterns cryptic. At the heart of their 'signature' approach is a two-step procedure. First, a set of input genes that are known to participate in the same process (or, in some cases, are chosen randomly) is selected. Then, a set of experiments is chosen over which the expression patterns of the input genes are significantly correlated. This step is critical, because genes sharing only a subset of regulatory motifs in common may not have significantly correlated expression patterns when all experimental conditions are considered. Next, a set of genes is chosen whose expression is significantly correlated considering only those experiments selected in the previous step. As a result, genes under complex combinatorial control can be identified as part of the same transcriptional module, even though they may show little correlation in expression over all experimental conditions. This method was applied to a comprehensive set of gene clusters obtained by grouping together genes with shared DNA motifs in their 5′ upstream regions. In

this way, every possible six, seven or eight base pair DNA sequence was considered as a possible regulatory motif, resulting in the identification of 86 transcriptional modules comprising over 2200 genes.

A complementary approach to unraveling combinatorial gene regulation was taken by Pilpel *et al.* [33]. In their computational study, Pilpel *et al.* first identified a set of 356 DNA motifs believed to represent transcription factor binding sites, using both known sites and sites obtained after applying the motif-finding Gibbs sampler AlignACE [34] to genes in related functional categories. They found that genes sharing a single regulatory motif in their promoter region showed little correlation in expression patterns, confirming results from other studies, but when they considered genes sharing the exact same subset of regulatory motifs, they found significant correlation in expression patterns. As most of their DNA motifs (92%) were identified using computational methods on partially annotated genomic sequence information, their method should be applicable to any organism with a complete genome sequence and a large collection of gene expression data, such as human.

In contrast to the large amount of gene expression data currently available for yeast, there are only about 50 different whole genome microarray data sets available for all bacterial species (although this number is increasing rapidly). Nonetheless, the large number of genome sequences completed or currently in the pipeline provides an equally exciting opportunity to unravel bacterial and archaeal regulatory networks. In particular, two approaches are beginning to show considerable progress: using conservation of gene order in operons across unrelated genomes to infer co-regulation and phylogenetic footprinting of *cis*-regulatory motifs upstream of several orthologous transcriptional units.

Snel *et al.* [35•] used genomic data to identify transcriptional modules, or regulons (defined as the set of genes regulated by a common transcription factor), by grouping together genes whose orthologs occur within the same operon in several unrelated species. Several recent advances in operon prediction [36–40], as well as the large number of sequenced microbial genomes, promise to further improve methods that take advantage of operon-based gene order conservation. As more genome sequences become available, however, (sometimes spurious) interactions between modules tend to interconnect many functionally unrelated pathways. Their solution is to identify and remove connections from several 'linker' proteins, which tend to connect otherwise distinct subnetworks. Using this graph-based heuristic, they recovered a set of nearly 800 functionally homogenous modules, including many proteins of previously unknown function. Although this study and the study by Barkai and co-workers [32•] focused on different organisms, the large discrepancy between the numbers of modules identified (86 modules in 2200 genes versus 800 modules in 3000 genes) underscores the lack of a universally accepted operational definition of modularity.

Another approach to parsing bacterial and archaeal transcriptional regulatory networks into discrete modules is the use of phylogenetic footprinting [41–44]. In a footprinting study, several related genomes are compared by aligning the upstream regions of orthologous genes. Assuming that orthologous genes are under similar transcriptional control in related species (not always true), the most highly conserved sequences should represent functionally important DNA motifs, such as transcription factor binding sites. Preliminary studies have focused on proteobacterial species, but the refinement of algorithms and the many more related families of bacteria in the genome sequencing pipeline (such as cyanobacteria and lactic acid bacteria) make this one of the most exciting research areas in microbial genomics.

Motifs

In the previous section, we presented a loose definition of modules as discrete units of function separable from the whole. Here, we define motifs as a set of genes or gene products with specific molecular functions arranged together such that they perform some 'useful' behavior. In contrast to modules, the behaviors of motifs are not, in general, separable from the rest of the system and they generally constitute only part of a recognizable systems-level function (such as a feedback loop or a logical operation). From an engineering perspective, it is satisfying to think that modules of genes, such as those being elucidated by methods already described, comprise common motifs arranged in new ways to produce different phenotypes. Even more satisfying would be a set of design principles common both to circuits of human design and to those that result from evolution. In two separate computational studies, Alon and co-workers [45,46•] make the claim that, in fact, both of these statements are true. In their initial study, they explore the network of transcriptional regulation in *E. coli* by searching RegulonDB (a database of *E. coli* operons, transcriptional regulators and promoters [47]) for three types of motifs: feed-forward loops, in which a transcription factor and its downstream target both regulate a third target; single-input modules, in which a group of operons is controlled by a single transcription factor; and dense overlapping regulons, in which the target operons for a group of transcription factors are highly overlapping. These three motifs were found to occur more frequently in the *E. coli* transcriptional network than in random networks, supporting the idea that they represent basic building blocks of transcriptional circuits. In a more recent study, Alon and co-workers found a set of over-represented motifs shared between both biological circuits and circuits of human design by enumerating all

three- and four-node subnetworks of several biological and technological systems, including transcriptional regulation networks, food webs, neuronal connections in *C. elegans*, electronic circuits and a subset of the World Wide Web. By studying these vastly different types of networks, they were able to make generalizations about the types of motifs suited to different networks, and found distinct differences between networks whose primary function is to carry out energy flow and those that perform information processing.

Interactions between different networks can complicate analysis

Although much initial progress has been made toward understanding individual networks, such as protein–DNA interactions and metabolism, much work remains to be done before we have a clear picture of how information is passed between these different types of networks. The importance of drawing these connections can be demonstrated by considering some specific examples from the exhaustive search for circuit motifs discussed in the previous section. Surprisingly, Alon and co-workers found no negative-feedback loops (ignoring autoinhibitory regulators) in the transcriptional network of *E. coli*, whereas a similar study from the same group found that both three- and four-node feedback loops are highly overrepresented in electronic circuits. This raises the question: do biological systems use negative feedback? The answer is certainly yes, even the *E. coli* transcriptional network. One clear example of negative feedback is the regulation of the Trp operon, in which the metabolic end product, tryptophan, represses transcription of the Trp enzymes by binding directly to the Trp repressor [48].

Consider also the Ara operon, cited by Alon and co-workers as an example of a feed-forward loop. Because CAP regulates AraC, binding upstream near the AraC promoter, and both AraC and CAP bind upstream of the AraBAD operon, this system seems to be a clear example of the feed-forward motif described in their recent study. However, this simple model is only strictly correct in the limit where cAMP and arabinose concentrations are high. When interactions between proteins and metabolites are taken into account, the picture becomes somewhat more complex. In the absence of arabinose, AraC represses transcription of AraBAD through a DNA-looping mechanism, whereas in the presence of arabinose, AraC activates transcription of AraBAD, but negatively regulates its own transcription [49]. As a result, the behavior of this circuit *in vivo* may be qualitatively different from its apparent behavior from a purely protein/DNA-centric point of view.

Initial efforts to tie these distinct types of data together are already underway. Several recent examples include the use of yeast two-hybrid and protein–DNA interaction data to better understand gene expression patterns [50–53], and the use of protein–protein interaction data together with gene expression data to uncover signal transduction pathways [54].

Engineering networks

A critical test of our understanding of gene networks is the *de novo* design of genetic circuits with novel behavior. Recent attempts to design novel circuits from existing genetic parts have received enormous interest because they represent proof of principle that the *de novo* design of useful motifs, which could ultimately be used as components in larger synthetic networks, is within our grasp [55-62,63••]. Of particular interest is a recent study by Guet *et al.* [63••] that employed combinatorial libraries to sample the topology/parameter space available to a simple three-node regulatory network containing LacI, TetR and lambda cI, together with five promoters of varying strengths and specificities regulated by these proteins. From these simple components, numerous networks were identified that performed computational functions on the input signals (presence or absence of anhydrotetracycline and IPTG), such as NAND, NOR and NOT IF. These and other related studies have brought to light several important design considerations for genetic networks. First, network topology alone is not sufficient to determine network behavior. In the study by Guet *et al.*, pairs of networks were found that have the same topology but different logical behaviors, as well as pairs of networks with different topologies but the same logical behavior. Second, the noise inherent to genetic systems requires that reliable circuits produce behavior that is robust to fluctuations in both the choice of parameters and the initial state of the system. As a result, biological networks such as the *Drosophila* segment polarity module have been shown to display a robustness to both the exact parameter values and the initial state of the system that is almost unheard of in networks of human design [64]. Notably, nearly all of the above studies combined analytical or numerical simulation of networks with experimental validation, evidence that our concept of how to study complex biological systems is shifting from brute-force genetics toward an engineering approach.

Although we are far from constructing synthetic networks as complex as those seen in nature, Ideker *et al.* [52] developed a systematic engineering-based approach to studying complex regulatory networks that met with considerable success. Underlying their approach is a cycle of model building, systematic experimental perturbation and model refinement. Their recent study represents a first pass through this cycle applied to the well-studied yeast GAL (galactose utilization) network. After constructing a model that reflects our current understanding of the regulation of galactose utilization (obtained from the literature), they systematically constructed knockout strains for nine genes in the regulation of the GAL pathway. Next, they identified yeast genes responding

to these perturbations using DNA microarrays and isotope-coded affinity tag (ICAT) tandem MS. For genes that responded to these perturbations, they found specific connections at the protein interaction level (by searching databases of published protein–protein and protein–DNA interactions), as well as at the genomic level (by identifying *cis*-regulatory sequences upstream of these genes), that might explain their observed behavior. Finally, they revisited the original model with several observations that suggest specific refinements, such as a role for the metabolite galactose-1-phosphate in galactose regulation.

Conclusions

When we first entered the era of high-throughput biology several years ago, there was considerable debate in the editorial columns of numerous journals over what role hypotheses would play in this new biology [65–67]. A particularly contentious point was whether data collection should be driven by the desire to test specific hypotheses or whether it is better to collect as broad a sample as possible. On the one hand, technological advances promised a comprehensive biological data set faster and cheaper than could be obtained through a less systematic, hypothesis-driven effort. The strongest technology advocates even suggested that, to discover unexpected patterns, we must approach data collection free of the bias of hypotheses. On the other hand, critics of this approach questioned the accuracy of data that were collected without the kinds of well-reasoned controls that accompany a more focused research plan, as well as the premise that data-mining alone can produce understanding.

In this review, we have tried to build the case that the era of high-throughput biology is well underway and that we are entering what has been called the 'era of pattern detection' [66]. At this point, it is useful to step back and consider to what extent the arguments for and against hypothesis-free data collection have held true. First, there is no question that high-throughput techniques have produced data faster and cheaper even than originally anticipated. Moreover, these data have profoundly changed our perception of biology and given rise to entirely new fields of study, such as comparative genomics, that would not have been possible otherwise. At the same time, however, our ability to turn systems-level data into systems-level understanding has been limited.

Throughout our review of theoretical approaches to understanding systems-level networks, we have cited complicating factors that can limit their scope or relevance. We have also endeavored to point out that the limitations of current theoretical modeling efforts stem not from a lack of sufficiently clever approaches, but rather from the difficulties of modeling complex biological systems. In our own modeling experience, directed at understanding protein folding and the lysis/lyogeny decision in lambda phage, we have found that 'simple' models, which incorporate data from numerous different experimental approaches, often take tens of thousands of lines of computer code to state precisely. Understanding the systems-level behavior of entire cellular networks will probably be much more complicated. Difficulties arise from interactions between many different cellular subnetworks, ubiquitous feedback loops and the fact that network behavior depends not only on its topology, which can be deduced from high-throughput interaction assays, but also on the details of specific kinetic parameters, which generally require directed biochemical studies to obtain.

So, to what extent do our current limitations stem from the trade-off between the breadth of data available from high-throughput methods and the depth of knowledge that more traditional approaches provide? As an example, consider again the regulation of the arabinose operon. Given perfect (error-free) data, including the key proteins involved and all protein–protein, protein–DNA and protein–metabolite interactions, we might be able to form a rough sketch of how the basic regulation of this operon works. However, it is unlikely that we would be able to build an accurate model of the dynamical behavior of this simple system without further mechanistic information. Moreover, the early acceptance of such a simplified model may have precluded the years of genetics, biochemistry and structural biology that produced such profound insights as the discovery of DNA looping, which helped to explain the mechanism of eukaryotic enhancers, and the role of arm–domain interactions in protein function [49].

What, then, is the role of theoretical modeling efforts in relation to the -omic scale data sets being generated? At present, there is already a wealth of data deposited in public databases that might yield insight into currently unresolved biological problems. Although efforts to model the systems-level 'behavior' of cellular interaction networks based solely on these data may not even be feasible, we certainly expect more from our investment than just individual facts deposited in a database. New algorithmic approaches to parsing networks into modules and motifs represent exciting first steps toward adding more value and biological relevance to these data. Furthermore, engineering-based approaches, including the *de novo* design of simple networks, coupled with modeling bring us closer to the ultimate goal of building realistic large-scale models of biological systems.

At the beginning of this review, we presented the question: what can we learn about biology by studying networks? Though we are still a long way from a complete answer to this important question, we can offer two partial answers. First, network-based approaches to uncovering patterns will help to organize this vast collection of data in a way that makes it more accessible and valuable to traditional biologists. Second, reformulating existing

biological questions from a network perspective (as discussed in a previous section) has the potential to take full advantage of the wealth of available data and answer questions that could not be addressed otherwise. Although these complex biological networks are proving to be more difficult to model the more we learn about them, we fully expect that our efforts will be rewarded with a detailed picture of the process by which living systems derive phenotype from genotype.

Acknowledgements

We are grateful to the Howard Hughes Medical Institute and the Department of Energy for the support of research particular to this review. We also thank Chris Rao and the other members of the Arkin laboratory for helpful comments and discussion on this manuscript.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**:399-403.
The authors describe in detail the advantages and disadvantages of several high-throughput methods for detecting protein–protein interactions. They also compare predictions from each method with a high confidence reference set to estimate accuracy and coverage, and investigate possible sources of bias.

2. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae***. *Nature* 2000, **403**:623-627.

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

4. • Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J, Michon A, Cruciat C *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**:141-147.
See annotation to [5•].

5. • Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, Adams S, Millar A, Taylor P, Bennett K, Boutilier K *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry**. *Nature* 2002, **415**:180-183.
These papers [4•,5•] describe two similar MS-based approaches to identifying components of multiprotein complexes. These approaches offer advantages over yeast two-hybrid screening because cooperative binding between three or more proteins is allowed and interactions need not occur in the nucleus. On the other hand, transient interactions may be difficult to detect using these methods.

6. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends Biochem Sci* 1998, **23**:324-328.

7. Overbeek R, Fonstein M, D'Souza M, Pusch G, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.

8. Pellegrini M, Marcotte E, Thompson M, Eisenberg D, Yeates T: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.

9. Enright A, Iliopoulos I, Kyrpides N, Ouzounis C: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature* 1999, **402**:86-90.

10. Marcotte E, Pellegrini M, Thompson M, Yeates T, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function**. *Nature* 1999, **402**:83-86.

11. Tong A, Evangelista M, Parsons A, Xu H, Bader G, Pag N, Robinson M, Raghibizadeh S, Hogue C, Bussey H *et al.*: **Systematic genetic analysis with ordered arrays of yeast deletion mutants**. *Science* 2001, **294**:2364-2368.

12. Mewes H, Frishman D, Guldener A, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 2002, **30**:31-34.

13. Csank C, Costanzo M, Hirschman J, Hodges P, Kranz J, Mangan M, O'Neill K, Robertson L, Skrzypek M, Brooks J, Garrels J: **Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD)**. *Methods Enzymol* 2002, **350**:347-373.

14. Bader G, Hogue C: **BIND–a data specification for storing and describing biomolecular interactions, molecular complexes and pathways**. *Bioinformatics* 2000, **16**:465-477.

15. Xenarios I, Salwinski L, Duan X, Higney P, Kim S, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Res* 2002, **30**:303-305.

16. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al.*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**:2101-2105.

17. •• Lee T, Rinaldi N, Robert F, Odom D, Bar-Joseph Z, Gerber G, Hannett N, Harbison C, Thompson C, Simon I *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799-804.
A map of protein–DNA interactions in yeast was built using high-throughput ChIP arrays. These interactions provide an overall picture of the transcriptional regulatory network in yeast and should prove an invaluable resource for future theoretical studies.

18. Barabasi A, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**:509-512.

19. Albert R, Jeong H, Barabasi A: **Error and attack tolerance of complex networks**. *Nature* 2000, **406**:378-382.

20. Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**:1551-1555.

21. Luscombe N, Qian J, Zhang Z, Johnson T, Gerstein M: **The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties**. *Genome Biol* 2002, **3**:RESEARCH0040.

22. Rzhetsky A, Gomez S: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome**. *Bioinformatics* 2001, **17**:988-996.

23. Hartwell L, Hopfield J, Leibler S, Murray A: **From molecular to modular cell biology**. *Nature* 1999, **402**:C47-C52.

24. Watts D, Strogatz S: **Collective dynamics of 'small-world' networks**. *Nature* 1998, **393**:440-442.

25. Fraser H, Hirsh A, Steinmetz L, Scharfe C, Feldman M: **Evolutionary rate in the protein interaction network**. *Science* 2002, **296**:750-752.

26. Park J, Lappe M, Teichmann S: **Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast**. *J Mol Biol* 2001, **307**:929-938.

27. Edwards A, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes**. *Trends Genet* 2002, **18**:529-536.

28. Teichmann S, Rison S, Thornton J, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli***. *J Mol Biol* 2001, **311**:693-708.

29. Saqi M, Sternberg M: **A structural census of metabolic networks for *E. coli***. *J Mol Biol* 2001, **313**:1195-1206.

30. Rison S, Teichmann S, Thornton J: **Homology, pathway distance and chromosomal localization of the small molecule**

metabolism enzymes in *Escherichia coli*. *J Mol Biol* 2002, **318**:911-932.

31. Alves R, Chaleil R, Sternberg M: **Evolution of enzymes in metabolism: a network perspective**. *J Mol Biol* 2002, **320**:751-770.

32. • Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**:370-377.
A new method for analyzing genome-wide expression data is applied to a large yeast data collection. A two-step approach is described that can overcome some problems associated with combinatorial regulation of yeast gene expression. In addition, the approach allows individual genes to be included in more than one functional module. Applying their algorithm to a comprehensive set of putative gene clusters, the authors are able to assign about 2200 yeast genes to 86 functional modules.

33. Pilpel Y, Sudarsanam P, Church G: **Identifying regulatory networks by combinatorial analysis of promoter elements**. *Nat Genet* 2001, **29**:153-159.

34. Hughes J, Estep P, Tavazoie S, Church G: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae***. *J Mol Biol* 2000, **296**:1205-1214.

35. • Snel B, Bork P, Huynen M: **The identification of functional modules from the genomic association of genes**. *Proc Natl Acad Sci USA* 2002, **99**:5890-5895.
The authors analyzed protein interaction networks derived from the co-occurrence of genes in operons by partitioning them into functional modules. In doing so, they identify 'linker' proteins, which tend to connect functionally distinct clusters into one large component, and remove connections from these proteins to preserve the functional homogeneity within modules.

36. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes**. *Bioinformatics* 2002, **18(suppl 1)**:S329-S336.

37. Zheng Y, Szustakowski J, Fortnow L, Roberts R, Kasif S: **Computational identification of operons in microbial genomes**. *Genome Res* 2002, **12**:1221-1230.

38. Sabatti C, Rohlin L, Oh M, Liao J: **Co-expression pattern from DNA microarray experiments as a tool for operon prediction**. *Nucleic Acids Res* 2002, **30**:2886-2893.

39. Ermolaeva M, White O, Salzberg S: **Prediction of operons in microbial genomes**. *Nucleic Acids Res* 2001, **29**:1216-1221.

40. Craven M, Page D, Shavlik J, Bockhorst J, Glasner J: **A probabilistic learning approach to whole-genome operon prediction**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:116-127.

41. Rajewsky N, Socci N, Zapotocky M, Siggia E: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons**. *Genome Res* 2002, **12**:298-308.

42. McCue L, Thompson W, Carmack C, Ryan M, Liu J, Derbyshire V, Lawrence C: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes**. *Nucleic Acids Res* 2001, **29**:774-782.

43. McGuire A, Hughes J, Church G: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes**. *Genome Res* 2000, **10**:744-757.

44. Gelfand M, Koonin E, Mironov A: **Prediction of transcription regulatory sites in archaea by a comparative genomic approach**. *Nucleic Acids Res* 2000, **28**:695-705.

45. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**:824-827.

46. • Shen-Orr S, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of *Escherichia coli***. *Nat Genet* 2002, **31**:64-68.
The authors extend the notion of motifs to the genetic regulatory network of *E. coli* and find that much of the protein–DNA interaction network in this organism is built from recurring patterns of connections. In particular, they describe and analyze three such patterns that occur much more often than would be expected in a random network.

47. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12**. *Nucleic Acids Res* 2001, **29**:72-74.

48. Hawkins J: *Gene Structure and Expression*. Cambridge, UK: Cambridge University Press; 1996.

49. Schleif R: **Regulation of the L-arabinose operon of *Escherichia coli***. *Trends Genet* 2000, **16**:559-565.

50. Ge H, Liu Z, Church G, Vidal M: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae***. *Nat Genet* 2001, **29**:482-486.

51. Ideker T, Ozier O, Schwikowski B, Siegel A: **Discovering regulatory and signalling circuits in molecular interaction networks**. *Bioinformatics* 2002, **18(suppl 1)**:S233-S240.

52. Ideker T, Thorsson V, Ranish J, Christmas R, Buhler J, Eng J, Bumgarner R, Goodlett D, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**:929-934.

53. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae***. *Nucleic Acids Res* 2001, **29**:3513-3519.

54. Steffen M, Petti A, Aach J, D'haeseleer P, Church G: **Automated modelling of signal transduction networks**. *BMC Bioinformatics* 2002, **3**:34.

55. Elowitz M, Leibler S: **A synthetic oscillatory network of transcriptional regulators**. *Nature* 2000, **403**:335-338.

56. Gardner T, Cantor C, Collins J: **Construction of a genetic toggle switch in *Escherichia coli***. *Nature* 2000, **403**:339-342.

57. Becskei A, Serrano L: **Engineering stability in gene networks by autoregulation**. *Nature* 2000, **405**:590-593.

58. Becskei A, Seraphin B, Serrano L: **Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion**. *EMBO J* 2001, **20**:2528-2535.

59. Elowitz M, Levine A, Siggia E, Swain P: **Stochastic gene expression in a single cell**. *Science* 2002, **297**:1183-1186.

60. Thattai M, van Oudenaarden A: **Attenuation of noise in ultrasensitive signaling cascades**. *Biophys J* 2002, **82**:2943-2950.

61. Ozbudak E, Thattai M, Kurtser I, Grossman A, van Oudenaarden A: **Regulation of noise in the expression of a single gene**. *Nat Genet* 2002, **31**:69-73.

62. Arkin A, Ross J, McAdams H: **Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells**. *Genetics* 1998, **149**:1633-1648.

63. •• Guet C, Elowitz M, Hsing W, Leibler S: **Combinatorial synthesis of genetic networks**. *Science* 2002, **296**:1466-1470.
The authors surveyed the phenotypic behaviors available to a three-node genetic regulatory network by constructing a combinatorial library of promoters and transcriptional regulators.

64. von Dassow G, Meir E, Munro E, Odell G: **The segment polarity network is a robust developmental module**. *Nature* 2000, **406**:188-192.

65. Brown P, Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nat Genet* 1999, **21**:33-37.

66. Goodman L: **Hypothesis-limited research**. *Genome Res* 1999, **9**:673-674.

67. Allen J: **Bioinformatics and discovery: induction beckons again**. *Bioessays* 2001, **23**:104-107.