Rational drug discovery: what can we learn from regulatory networks?

Sui Huang

To enable the list of genes and proteins contained within genomic databases to be useful for drug discovery, we need to understand how the genome maps into the phenome. An essential, but not explicitly listed ingredient of the genome is the regulatory interactions between genes and proteins that form a genome-wide network. How can the concept of regulatory networks increase our understanding of living systems? Networks are more than just static 'wiring diagrams'. Gene interactions impose dynamic constraints, which, although obvious in emergent phenotypic properties, are not captured by traditional one-gene-one trait approaches. Understanding the nature of these constraints in gene-activation state space will pave the way to a holistic yet formal and genomics-based approach to rational drug development.

Sui Huang
Dept of Surgery
Children's Hospital
Harvard Medical School
300 Longwood Avenue
Boston, MA 02115, USA
tel: +1 617 355 7852
fax: +1 617 566 6467
e-mail: sui.huang@tch.harvard.edu

Who'll know aught living and describe it well,
Seeks first the spirit to expel.
He then has the component parts in hand
But lacks, alas! the spirit's band.

J.W. von Goethe, Faust I, 1808

▼ The sequencing of the human genome has been celebrated as a technological milestone of similar magnitude to the first moon landing. However, if reading through the genome sequence was a giant leap for biologists, then it certainly was a small step for mankind. Life scientists have quickly realized that an encyclopedia of genes does not explain how the genome maps into the phenome, nor does it provide a menu list for selecting drug targets. The next big step will be to address the daunting question of how the parts (genes and proteins) make up the whole organism. This question not only presents an old riddle of academic biology, but also has practical implications in drug therapy – the aim being to modulate the macroscopic, phenotypic behavior by interfering with the microscopic elements, the genes and proteins.

Although it is now obvious that there is no one-to-one mapping between a gene (or the protein it encodes) and a phenotypic trait, the deeply rooted intuition among life scientists is still that for every phenotype and every disease there is a potential molecular target. This paradigm explains the surprise and frustration in the field of genomics-driven drug discovery when the genome sequencing projects revealed that the number of genes in the human genome was as small as 30,000–40,000 [1]. That only 500 of the ~500,000 proteins encoded by these genes (a rough estimate allowing for post-transcriptional modifications, notably differential splicing, and proteolysis) are known drug targets should dispel the subconscious illusion that the genome database is a list of proteins whose individual functions directly translate into all aspects of the behavior of an organism [2]. It is noteworthy that many molecular targets of today's drugs were identified only after studying the mechanisms-of-action of successful drugs that were developed, based on a chain of serendipitous findings or naive mechanistic assumptions [3]. Such scenarios, then, often reveal that the drug in question acts on multiple targets that would not have been predicted previously.

Emergent properties and networks

A first step beyond the one-to-one mapping between genotype and phenotype is the proposal of a 'many-to-many' relationship, in which one gene can have many functions (pleiotropy), and

1359-6446/02/$ – see front matter PII: S1359-6446(02)02463-7

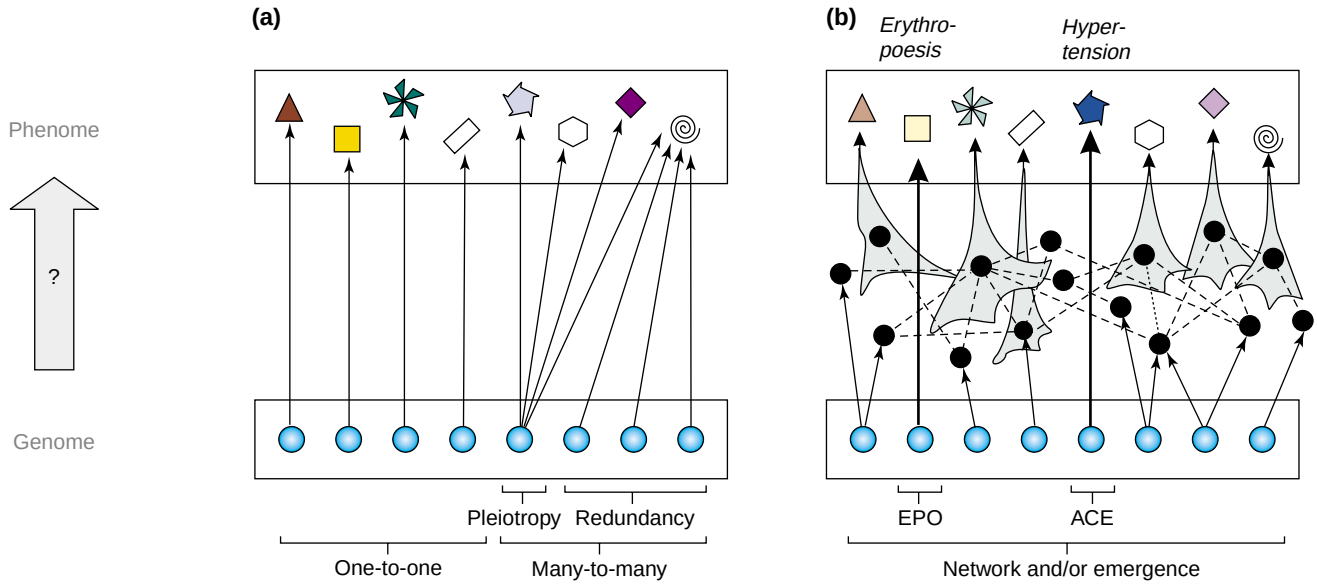

Figure 1. Genome–phenome relationship. (a) Traditional, linear gene-to-traits mapping. (b) Networks of interactions lead to emergent phenotypic properties that defy the one-gene-one trait notion. However, despite this complexity, there exist 'islands of linearity' with direct genotype-to-phenotype mapping: erythropoeitin (EPO) and angiotensin converting enzyme (ACE) have such properties and are obvious drug targets. Solid circles represent genes and proteins. Symbols of various shape represent phenotypic traits.

one function can be exerted by many genes (redundancy) (Fig. 1a). However, even with this concept, the full complexity of the genome–phenome relationship remains elusive. The problem is deeper. For instance, many core regulatory genes, such as *Ras*, *Myc*, *NF-κB* are not only pleiotropic, but exhibit disparate cell regulatory functions that depend – in a complex manner – on the cellular context [4]. As an increasing number of advocates for a complex system approach in biology point out [5,6], many phenotypic traits are properties that emerge from the collective action of individual genes as a result of non-linear interactions between them (Fig. 1b). These 'emergent properties', including cellular differentiation, cell activation states, homeostasis and mental functions, cannot be understood by the study of the parts in isolation, nor by 'adding up' those effects characterized in isolation [7]. Once more, Aristotle's notion that the 'whole is different from the sum of its parts' holds true. To be of practical use, however, we have to translate this ancient wisdom into molecular terms, with a formalism that can serve as a tool to study the complexity of the genome–phenome relationship.

So, what makes the whole different from the sum? The 'non-additive' ingredient in the genome is the regulatory interactions between genes and proteins that include non-linear cross-regulatory and auto-regulatory feedback loops. Together, they form a genomic regulatory network – Goethe's 'spirit's band', as it were, which orchestrates the parts and makes them act in symphony, not solo. In fact, as post-genomic biology takes on an integrative view, the notion of network has taken a center place in bioinformatics [3,8,9]. The concept of regulatory networks is a higher-level, coarse-grained abstraction that should not be confounded with metabolic networks, which essentially represent the physicality of the biochemical reactions, although the latter is connected to the former (e.g. metabolic enzymes that are directly regulated by growth factor signal transduction pathways). The term 'regulatory network' includes both the protein–gene interactions in the nucleus, and the cytoplasmic and extracellular protein–protein interactions. Thus, it encompasses the traditional concept of 'signal transduction pathways', which largely neglect feedback loops and cross-talk.

The advent of network concepts in bioinformatics raises the following question: How will network-related information be incorporated into current database structures, and how can the concepts of genome-wide regulatory networks contribute to our understanding of the complex genome–phenome relationship?

Networks and biological database structure

Gene and protein databases have evolved around the idea of simple genotype–phenotype relationships. As their original function was to collect, catalogue and classify individual genes and proteins, they are not suited to represent emergent properties that arise from networks of interactions and distributed information processing and, thus, these emerging properties cannot be assigned to individual molecules.

The simplest extension of existing databases to embrace the idea of a genomic regulatory network is to add information

about the interaction partner to the functional annotation (Fig. 2a). This has been done for the yeast protein databases (MIPS [10] and YPD [11])* where interaction partners, derived from individual or high-throughput experiments, are part of the annotation. Although such extension contains only crude binary information (i.e. presence or absence of interactions) without immediate significance to emergent biological functions, the simple list of protein interactions can be converted to reconstruct a crude map of the global network topology (wiring diagram; Fig. 2b). Despite the limited coverage and accuracy of current interaction data [12], the availability of such preliminary topology maps has already elicited many studies that have provided insights into fundamental organizational properties of regulatory networks at the genome scale [13] (see later).

The growing importance of network information leads to the question of how network topology should be represented in databases. Future annotations will have to contain the category of network-related properties of individual proteins, such as identity of interaction partners, connectivity (valency of a node), in-degree, out-degree, modality of interaction, centrality, and so on [13], as well as quantitative parameters such as binding affinities. Beyond the extension of current databases, a standardized formal language to represent network-related features of genes and proteins will be crucial to assimilate the growing volume of functional data on relationships between genes or proteins.

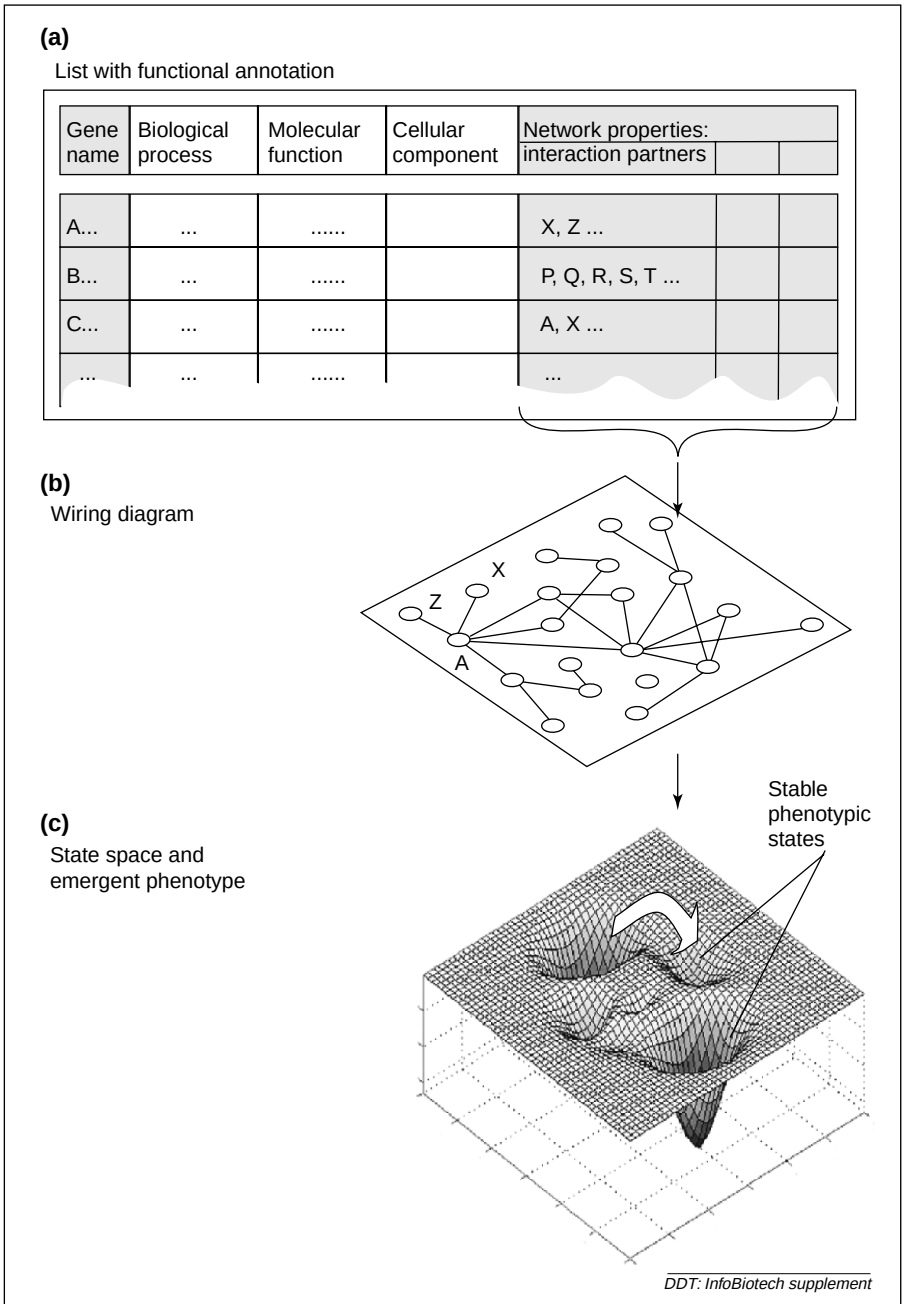

Figure 2. The integrative genomics approach to predicting complex phenotype. The figure depicts increasing levels of abstraction. (a) Database annotations containing interaction information can be translated to (b) a network architecture. The state space structure [(c) schematic projection of *N* dimensions into 3D space] represents the dynamic constraints of the network, and is essentially reflected in the phenotypic behavior. In this arbitrary 3D representation, every position in the X–Y plane would correspond to one state of the network of the cell (defined by a distinct gene-activation profile) at a given time. The 'altitude' (Z-axis) represents the 'potential' (in terms of a gradient) that pushes that cell towards the attractor state.

Reverse engineering the network architecture

The availability of massively parallel methods for monitoring gene activities (e.g. following gene-specific perturbation to the network) has opened the door to systematic elucidation of the network architecture; that is, reverse engineering of the actual wiring diagram of the molecular interactions of a genome, based on the temporal behavior of a large set of genes and proteins. An increasing number of algorithms for solving this network inference problem, typically relying on some simplifying assumptions, are being developed [14,15]. Beyond the

*MIPS, Münich Information Center for Protein Sequences; YPD, Yeast Protein Database.

primary goal of inferring the topology map (i.e. which genes regulate other genes), reverse engineering should ideally also lead to the elucidation of the modalities and quantitative parameters of interactions (i.e. how a gene regulates another). This information is crucial for simulating the dynamics of the network (see later).

However, despite the formidable academic challenge of this inverse problem, and the elegance of some approaches, the conceptual novelty of the reverse engineering of wiring diagrams remains limited. In a sense, it is what biology has been all about in the past two decades: finding out 'who' interacts with 'whom'. The difference between current network reverse engineering methods and traditional methods is procedural rather than conceptual. The use of algorithmic approaches to infer the wiring diagram serves merely to circumvent tedious experimental identification and characterization of individual interactions. Moreover, even if assuming that all difficulties of genomics-based reverse engineering (i.e. biochemical noise, averaging over asynchronous cell populations, dynamic links owing to context-dependent affinities and competition) could be overcome, one might ask: Will a complete blueprint of the wiring architecture (i.e. network topology and interaction modalities) bring us closer to understanding and predicting the macroscopic behavior of the cell or the organism? The answer is no. We would need to know how the network behaves as a whole; for that, we would also have to study its global dynamics (Fig 2; see later).

Nevertheless, network topology discovered to date has scientific value in itself and its study can provide new biological insight. Similar to DNA sequence analysis, network architectures can be subjected to cross-species comparison and can thereby shed light on evolution, taxonomy, organizational principle, and so on [16,17]. Moreover, graph-theoretical features of the generic topology of complex networks, such as connectivity distribution, modularity and hierarchies, can be studied [13,17–20]. For instance, the yeast proteome appears to represent a 'scale-free' network in terms of the connectivity distribution [17] (i.e. it is biased towards containing few 'hubs' or 'master proteins' that interact with a large number of partners) – a fact often under-appreciated by many experimentalists and bioinformaticians. Databases of interactions that are more accurate and complete than the existing ones will be necessary to answer unambiguously the question of whether the yeast and other interactomes are in fact truly scale-free, or not (Thomas Willhelm, pers. communication). From such studies of the generic topology, we will not only acquire a deeper understanding of fundamental design principles of complex of living systems and their evolution, but also derive practical functional properties of a given regulatory network, such as robustness and efficiency of information processing, as well as the role of particular genes [13,17,21].

Network architecture and dynamics: reverse engineering and simulations

Most reverse engineering approaches, based either on traditional, pathway-centered or novel, systematic approaches, provide a qualitative network topology map that contains only static information. By contrast, the phenotypic behavior, such as the switch of cell fates, and developmental processes that are governed by the regulatory network, directly reflect the dynamics of the underlying regulatory network. Two ingredients add 'dynamics', and hence, 'life,' to the topology map of the network: (1) the nodes of the network, that is, the proteins and genes, which can take values (expression level and activation state) that change over time; and (2) the interaction modalities, which dictate how the inputs to a given node affect its state (e.g. unconditional inhibition, collective activation, etc.). With all the information about the quantitative parameters, describing how the inputs ('upstream partners') affect the target nodes ('downstream partner'), one can model and simulate the network dynamics (and thus the system behavior) in a similar way to simulation of aeroplanes, nuclear plants and computer chips. This enables prediction of the epigenetic dynamics of the genome and, hence, the phenotype it governs. However, given the inherent technical challenges in reverse engineering of the topology of genomic networks, and the almost complete lack of information about the quantitative biochemical details, such explicit, exhaustive modeling is beyond our current capacity.

Moreover, the idea of simulating the dynamics of a system, based on the complete information of its architecture, is an ideal scenario that is rooted in engineering sciences; that is, man-made systems, where all the parts and the network of their functional relationships are known previously. With living systems, the process is reversed; this is reflected by the term 'reverse engineering', which must be performed before one can simulate the dynamics of that system. Thus, ironically, one must begin by monitoring the real dynamics intensively to extract the information required for inferring the network architecture, which will then be used to simulate those (very same) dynamics. Therefore, observing the system, reverse engineering the underlying network architecture, and simulating its dynamics, are all interwoven parts of one circular process.

There is another entry point to this cycle of inference: studying the dynamic features of the network and directly relating them to the phenotypic behaviour, rather than reconstructing all the details of the underlying network architecture. In fact, the technological challenge associated with the latter will mean resorting to more modest, but realistic goals as we attempt to link topology and dynamics. The operational questions could then be: How can we, armed with massively parallel and high-throughput technologies, exploit the notion of networks and emergent phenotypes to increase our understanding of living

systems, without subjugating our observation of gene activities solely to the goal of reverse engineering? Instead of first reverse engineering and then simulating the system, is there a short-cut for understanding the dynamics of a system? What alternatives do we have, in using these genome-wide measurement technologies, to reach beyond clustering genes for expression similarities, or collecting data to reconstruct the network architecture?

Network dynamics: focusing on constraints and emerging properties

An intrinsic property of networks is that, because of the regulatory interactions between genes and proteins, the majority of combinations (genome-wide profiles) of gene activities cannot be realized by the cell [3,22]. Instead, the huge number of theoretically possible gene activity combinations is massively reduced to a relatively small subset of characteristic gene activity profiles that satisfy the regulatory interaction rules. Our chance to understand the phenome as the collective dynamics of genes, without having to know the entire wiring program in detail [23] rests on this reduction of a vast space of (theoretical) gene activity combinations.

In technical terms, the constraining of the dynamics of a system by molecular interactions means that the high-dimensional state space of gene and protein activation is highly structured. (The state space is the *N*-dimensional space, in which every point represents a distinct network state as defined by a distinct gene activation profile, where N = number of genes). For example, if gene A (unconditionally) inhibits gene B, then all network states in which both A and B are active will be unstable, which forces the network to 'move' in state space until it hits a stable state (or cycles between a few states). Thus, the network can change its activity profile in only a few directions, following gradients ('trajectories') until it reaches a stable state, the so-called 'attractor state'. The existence of unstable regions and stable attractors impose a substructure to the state space, generating the 'attractor landscape' (Fig. 2c). Accordingly, the network state and, hence, the cell state, can be viewed as a marble on the landscape: it is forced to roll along valleys (trajectories) into the pits (attractors). This attractor landscape captures the phenotypic dynamics: it has long been proposed by great thinkers of modern biology, including Waddington, Delbrück, Jacob and Monod, and Kauffman, that the discrete, robust phenotypic states that we observe correspond to attractors in the state space as defined by the molecular activities of the underlying network (see Refs in [3]). Thus, attractors in the state space map into stable phenotypic states (i.e. cell fates, including differentiation to cell types, proliferation and apoptosis), and trajectories map into directed developmental processes. Although the attractor landscape picture appears to be an intuitive metaphor (as it was when first proposed by Waddington in the 1940s), it can now be reduced to formal and molecular terms, thus providing a conceptual framework for integrating gene action into global phenotypic behavior.

Current massively parallel technologies that measure genome-wide gene activation, such as DNA microarray-based expression profiling, and soon, protein activition, can now be used to characterize the dynamic constraints imposed by the underlying network. Experimental perturbations of tissue cultures or model organisms that elicit defined phenotypic switches, followed by analysis of the high-dimensional trajectories of the corresponding gene expression profile in the gene expression state space, will give a first feel for the attractor landscape and how it maps into the constrained phenotype dynamics. The knowledge of its specific shape and structure will pave the way for determining the possibilities and restrictions of cell and tissue behavior, which can be harnessed for diagnostic and therapeutic purposes. Although this application of DNA microarrays is a more modest approach than attempting to elucidate all the details of the network architecture, it is technically more realistic and practically more relevant.

Implications for drug discovery

The concept of high-dimensional state space and attractors formalizes an approach to understanding emergent properties arising from the complex collective action of genes that is responsible for the inherent difficulty in finding drug targets within a molecular network. An important consequence of the attractor landscape is that to move the network state from one site in the state space (e.g. a stable attractor state representing proliferation) to another state (e.g. terminal differentiation), one often has to operate multiple lever points, that is, affect a multitude of genes or proteins. Only occasionally do we observe a straightforward, conceivable mapping from a single protein to its phenotypic (patho)physiological effect. Specific examples include erythropoeitin (EPO), gastric mucosa H2-receptors and angiotensin converting enzyme (ACE) [24–26]. In these cases, the benefit of inhibiting or mimicking the biomolecule is immediately obvious and, fortunately, the mode-of-action of their major physiological effect is insulated from the rest of the network (Fig. 1). They appear to represent 'islands of linearity' in a sea of non-linearity and complexity, and are not subjected to the process of emergence, which blurs any direct micro-macro relationship. Such molecules are obvious targets for pharmacological intervention. It is probable that such cases of direct genotype–phenotype mapping represent exceptions rather than the rule.

By contrast, effective drugs with robust phenotypic effects in complex pathophysiological situations have turned out to affect many molecular targets, as expected from the model of an attractor landscape with stable states. A classical example are the non-steroidal anti-inflammatory drugs, such as salicylate (Fig. 3), which not only targets the cyclooxygenase (Cox)

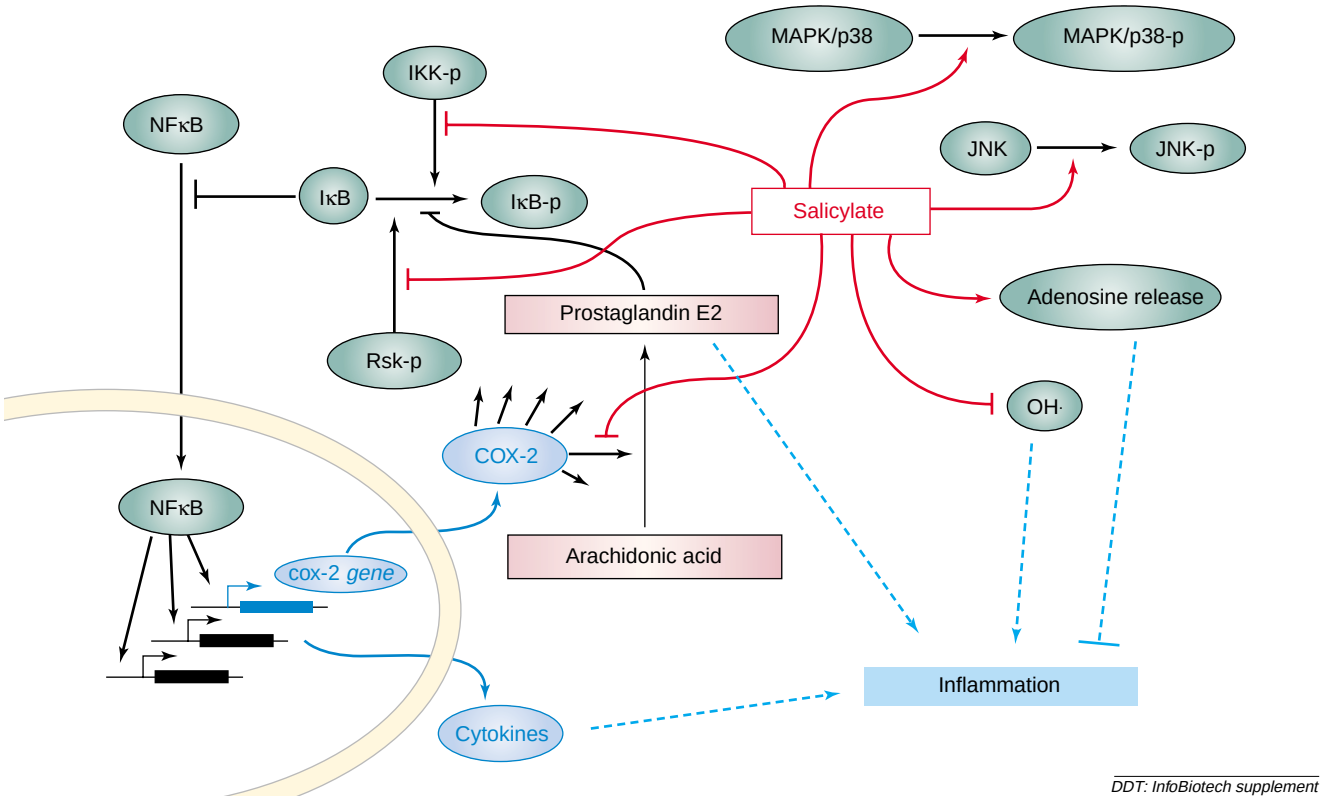

Figure 3. Modulating the behavior of an entire network. The figure depicts the multiple targets of salicylate. Note that many molecules that 'happen' to be targets of salicylate are normally involved in inflammation. Solid lines represent pathway interactions; dotted lines denote indirect physiological effect on inflammation.

enzymes to reduce prostaglandin synthesis, but also affects the *NF-κB* pathway, as well as many other connected cellular targets that all normally contribute to perpetuate the inflammatory state [27–29]. (The actual contribution of these additional targets to the mechanism-of-action remains to be determined.) A recent example of a multi-target drug is the anti-depressant, duloxetine, which inhibits the uptake of both serotonin and noradrenaline [30]. In general terms, the integrative biology view translates into a new principle for rational drug discovery: target network states, not individual proteins. This essentially means that we should explore multi-target drugs or (non-additive) combination therapies both explicitly and systematically.

Concluding remarks

The fundamental, long-term strategic question for drug development, which still relies on a mixture of rational theory and systematic trial and error, is how to split the rational part into the following two strategic alternatives: (1) continue to sieve through protein-function databases to find the remaining low-hanging fruits, that is, proteins that exhibit a one-to-one mapping between genotype and (pathological) phenotype and are thus obvious drug target candidates; and (2) seek to understand the basic features of dynamic molecular and physiological networks and their emergent properties. The second strategy might not yield immediate returns, but might in the future provide the knowledge necessary to predict the precise set of molecular lever points within the machinery of the genomic network. This would then enable researchers to 'steer' a complex phenotype within the virtual epigenetic landscape (which might have been altered by a disease) in the desired direction.

Acknowledgement

This work was supported by an Air Force Office of Scientific Research (AFOSR) grant (F49620-01-1-0564).

References

1 Claverie, J.M. (2001) Gene number. What if there are only 30,000 human genes? *Science* 291, 1255–1257

2 Editorial (2001) The usual suspects. Molecular drug targets 2001. *Nat. Biotechnol.* 19 (6) (Poster)

3 Huang, S. (2001) Genomics, complexity and drug discovery: insights from Boolean network models of cellular regulation. *Pharmacogenomics* 2, 203–222

4 Huang, S. and Ingber, D.E. (2000) Shape-dependent control of cell growth, differentiation and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell. Res.* 261, 91–103

5 Strohman, R.C. (1997) The coming Kuhnian revolution in biology. *Nat. Biotechnol.* 15, 194–200

6 Coffey, D.S. (1998) Self-organization, complexity and chaos: the new biology for medicine. *Nat. Med.* 4,882–885

7 Bar-Yam, Y. (1997) *Dynamics in Complex Systems (Studies in Nonlinearity)*, Perseus Publishing

8 Marcotte, E.M. (2001) The path not taken. *Nat. Biotechnol.* 19, 626–627

9 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285

10 Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34

11 Costanzo M.C. *et al.* (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, anintegrated resource for protein information. *Nucleic Acids Res* . 29, 75–79

12 von Mering, C. *et al.*(2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403

13 Barabasi, A-L. (2002) *Linked:The New Science of Networks*, Perseus Publishing

14 D'haeseleer, P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726

15 Yeung, M.K. *et al.* (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6163–6168

16 Fraser, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752

17 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

18 Strogatz, S.H. (2001) Exploring complex networks. *Nature* 410, 268–276

19 Wagner, A. (2002) Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Res.* 12, 309–315

20 Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science* 296, 910–913

21 Albert, R. *et al.* (2000) Error and Attack tolerance of complex networks. *Nature* 406, 378–382

22 Kauffman, S.A. (1992) *The origins of order*, Oxford University Press

23 Bailey, J.E. (2001) Complex biology with no parameters. *Nat Biotechnol.* 19, 503–504

24 Jelkmann, W. and Hellwig-Burgel, T. (2001) Biology of erythropoietin. *Adv. Exp. Med. Biol.* 502, 169–187

25 Del Valle, J. and Gantz, I. (1997) Novel insights into histamine H2 receptor biology. *Am. J. Physiol.* 273, G987–G996

26 Cody, R.J. (1997) The integrated effects of angiotensin II. *Am. J. Cardiol.* 79, 9–11

27 Yin, M.J. *et al.* (1998) The anti-inflammatory agents aspirin and salicylate inhibit the activity of IκB kinase-β. *Nature* 396, 77–80

28 Alpert, D. and Vilcek, J. (2000) Inhibition of IκB kinase activity by sodium salicylate *in vitro* does not reflect its inhibitory mechanism in intact cells. *J. Biol. Chem.* 275, 10925–10929

29 Vane, J.R. and Botting, R.M. (1998) Anti-inflammatory drugs and their mechanism of action. *Inflamm. Res.* 47 (Suppl. 2), S78-S87

30 Wong, D.T. and Bymaster, F.P. (2002) Dual serotonin and noradrenaline uptake inhibitor class of antidepressant potential for greater efficacy or just hype? *Prog. Drug. Res.* 58, 169–222