Experimental Data for Systems Biology

based on a review article by Bensimon et al. Ann. Rev. Biochem. (2012)

Limitations of the linear directional pathway paradigm

One of the central goals of biology is to understand how an organism's genetic information acts together with environmental influences to shape that organism's phenotype. Until recently, the connection between genotype and phenotype was thought of as a linear process (Figure 1), in which one gene, encodes one protein, which carries out one function.

This paradigm is based on two assumptions that have dominated the thinking of generations of biologists and guided the development of the techniques of molecular biology. Firstly, it postulates a direct link between gene and protein function, implying that it is sufficient to know all genes and their translation products in order to explain the functioning of an organism. Secondly, it places individual proteins and their associated functions in linear pathways, implying that every function “downstream” is affected by an upstream block, while every function “upstream” is unaffected by its downstream blocks.

Before powerful "omics" techniques made it possible to determine the full complement of genes, proteins and RNAs in a biological system, it was widely assumed that the impossibility to fully explain phenotypes through genotypes was due to the existence of additional genes and pathways that had not yet been discovered.

Nowadays, thanks to whole genome sequencing, there are many organisms of which all genes are known and whose complete inventory of RNA, proteins and metabolites has been determined. Yet, apart from very few gene defects, for which the linear directional pathway paradigm appears to be applicable, it remains very difficult to understand the molecular sources of most disease phenotypes.

The reasons for these difficulties are more likely conceptual rather than merely technical. It appears that the paradigm of the single, directional pathway often fails, because it neglects the context of the molecules in the cell as well as cross talk and feedback mechanisms between pathways.

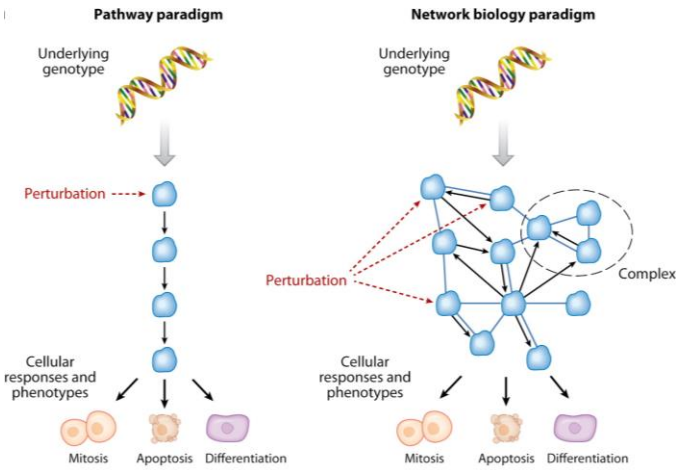

Figure 1. *illustrates the difference between the traditional "one gene, one protein, one function" pathway paradigm (left) and the modern network paradigm (right). In the pathway paradigm information flows from the genotype to the phenotype in a linear fashion and one gene ultimately influences one phenotype. Under this paradigm, a perturbation is thought to influence the phenotype by interacting with a single member of the pathway. Under the network paradigm, a network of interactions replaces the linear pathway. Information flow involves branching and feedback loops. A perturbation's influence on the phenotype results from that perturbation's effect on multiple members of the network and the networks response to that perturbation. (From Bensimon et al. Annual Reviews of Biochemistry (2012))*

From linear pathways to networks: the network paradigm

The new network paradigm replaces the linear directional pathway linking genotype and phenotype by a network (Figure 1).

Each node represents a molecule of interest, such as a gene, any of its products (e.g. RNA or proteins) or smaller molecules such as co-factors, messenger molecules and metabolites. The edge between two nodes represents a relationship, such as a physical interaction, an enzymatic reaction or a functional connection. While the molecule-centric linear pathway paradigm primarily considers a network's nodes, the new paradigm is also concerned with network edges.

The network paradigm assumes that biological networks are dynamically modulated at different time scales by external (e.g. environmental) or internal (e.g. genomic alterations) perturbations and that the properties of the

entire network determine the phenotype. In this sense, an organism's genomic information expresses itself as much in the structure and topology of the network as in the identity of the individual network components.

Implications of the new network paradigm

The network biology paradigm has several important implications. First, it requires a different, more integrative view of biological processes, as the contextual relationships between molecules move to the forefront. Second, network biology provides new opportunities for and critically depends on new experimental and computational approaches, including methods to visualize networks, methods to infer network topology and structure and methods to simulate and model the dynamic behavior of networks and their phenotypic consequences. Thus, network biology has been stirring novel technologies that focus on the measurement of contextual relationships of molecules, rather than on simply enumerating molecules in a catalogue format.

A general roadmap for studying biological networks

Much like the biological network themselves the process of constructing models for these networks is not a linear series of consecutive steps but involves iterative loops and feedback mechanism. Ideker et al. (Science, 2001) have proposed the following generic roadmap for the construction and refinement of network models.

Define all genes in the genome and the subset of genes, proteins, and other small molecules constituting the subnetwork of interest. If possible, define an initial model of the molecular interactions governing pathway function, drawn from previous genetic and biochemical research.

Perturb each pathway component through a series of genetic (e.g., gene deletions or overexpression) or environmental (e.g., changes in growth conditions or temperature) manipulations. Detect and quantify the corresponding global cellular response to each perturbation with technologies for large-scale mRNA- and protein-expression measurement.

Integrate the observed mRNA and protein responses with the current, pathway- specific model and with the global network of protein-protein, protein-DNA, and other known physical interactions.

Formulate new hypotheses to explain observations not predicted by the model. Design additional perturbation experiments to test these, and iteratively repeat steps (ii), (iii), and (iv) until a satisfactory model can be obtained.

The shift from the pathway to the network paradigm also shifts the requirements for experimental data

To support the roadmap outlined above, the experimental techniques used need to be able to generate datasets that ideally fulfill all of the following criteria:

Firstly, the data must be complete, i.e. all the nodes and edges of the network under investigation should be measurable;

Secondly, the data need to be reproducible, i.e. identical results should be obtained in each repeated measurement of a network;

Thirdly, the data need to be quantitative for us to detect dynamic changes of network components and further the data need to be measurable at a reasonable pace so that data can be collected from a large number of samples representing multiple, differently perturbed states of a system.

At first these criteria may appear to be valid for any type of scientific data, but this is not the case. The example from protein mass spectrometry below illustrates that the shift from the linear-pathway to the network paradigm also requires a shift in data collection methods.

The pathway paradigm posits that the inability to explain a certain biological phenomenon is likely due to the existence of a, yet uncharacterized, molecular species or of a new interaction between known species. The new species or interaction may then provide a hint for a new pathway that leads to the phenotype under investigation. This calls for an experimental strategy that focuses on the discovery of new molecular species and new protein-protein interactions, needs that are best met by so-called shotgun mass spectrometry techniques that aim to detect and identify as many different molecular species as possible at the lowest possible concentrations. Meanwhile, the breadth and sensitivity afforded by the shotgun approach comes at the price of reduced quantitative accuracy, experiment-to-experiment reproducibility and speed. Under the linear pathway paradigm, however, these disadvantages have little relevance, because new scientific insights are expected to come primarily from the discovery of new molecules and interactions.

By contrast, the preferred protein mass spectrometry approach for the network paradigm is the SRM (selective reaction monitoring) method. This method allows the measurement of only a rather small subset of predetermined peptides that were specifically chosen to reflect the concentration of those proteins that belong to the subnetwork under investigation. Reducing the number

of peptides being measured allows them to be measured with much greater accuracy and results in less data-collection time per sample. This makes it possible to collect SRM data for a large number of samples (each sample representing a different perturbation of the network) and to make meaningful quantitative comparisons between these samples. Both are essential for studies under the network paradigm, because here we do not expect to understand the organism through the discovery of new network components, but by comprehending subtle changes in the interactions between different network components. A limitation of the SRM technique is the relatively small number of proteins (network nodes) (50-100) that can be analyzed from a sample, which limits the size and number of subnetworks that can be analyzed by the method. However, recent developments in mass spectrometry, a technique called SWATH-MS, for example, have significantly increased the number of proteins that can be quantified from a sample to 3000-6000, while maintaining the performance characteristics of the SRM method that are important for its use in network biology.

Distinguishing different types of networks: physical interaction networks vs. signaling networks and functional networks.

Networks consist of nodes and edges. In the study of molecular systems, the nature of the nodes is clearly defined, as each node corresponds to a type of molecule. There are, however, many different types of interactions (edges) connecting the nodes. Here are three examples. The first type of edge is that of a physical contact between the molecules that the connected nodes represent. This type of edge is undirected ("A binds to B" and "B binds to A" are equivalent statements). The network that connects nodes through physical binding interactions can be called a physical interaction network.

Alternatively, an edge could indicate that one of the nodes sends a signal to the other node. This type of edge is directional - a kinase phosphorylates its target, but not vice versa. Networks formed by these and similar signaling interactions will be referred to as signaling networks. In some cases, signaling interactions take place between network nodes that are also in physical contact with one another, but this is not necessarily the case. Rather, the signaling process could take place through a small-molecule messenger or, as in the case of some ion channels, via through-space electrostatic interactions.

The third type of network edge represents functional interactions, such as those that are observed in synthetic lethality screens. These non-directional edges indicate that two nodes are involved in functionally related processes.

On their one, each of these types of networks will show only a partial view of the system. To be able to comprehend the actual biological system, we must combine all three types of networks.

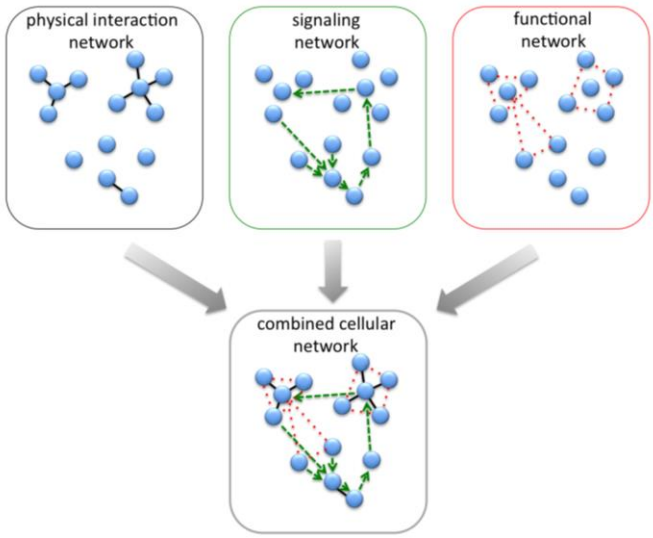

Figure 2. *Network nodes can be connected by multiple different types of interactions (edges), which are revealed by different types of experimental data. By integrating more and more different types of networks into a combined network view we obtain a more complete view of the biological system.*

Correlations as evidence for functional network interactions.

The expression "correlation is not causation" is one of the mantras of data analysis and often bears the connotation that correlations represent an inferior type of scientific evidence that can - or even should - be disregarded. While it is certainly true that correlation between events alone is no proof for one event causing the other, evidence of correlation should nevertheless be considered as valuable information.

Possible reasons for the correlation of two events (A and B) are for instance 1) pure chance and A and B have nothing to do with one another or 2) A and B are linked to each other indirectly, for example by an event C that causes both A and B or 3) A causes B or B causes A - in a probabilistic sense at least.

Case 1) clearly provides no biologically relevant information and would reduce the quality of a network model if it were included. The probability of such "chance"

associations occurring can, however, be calculated and with the proper statistical tools for significance testing and multiple-testing correction, such chance associations can be filtered out efficiently.

This leaves two possible reasons for correlation. Both offer valuable information, particularly in research scenarios where the edges and nodes of the network are still to be defined. Here, even correlations that are due to indirect linkage (case 2) can provide new hypotheses for potential functional interactions which can be the basis of subsequent experiments.