Systems Biology: unravelling complex networks?

Peter Grindrod
(University of Bath and University of Oxford)

Desmond J. Higham
(University of Strathclyde)

J. Keith Vass
(The Beatson Institute, Glasgow)

Alastair Spence
(University of Bath)

September 2006

ABSTRACT

In this paper we discuss the background and challenges inherent in analysing biological systems from a data driven perspective. We focus on networks arising in cell biology, which, from a mathematical viewpoint, can be studied with the tools of applied graph theory. In particular we discuss how taking a systems biology approach necessitates

- the development of alternative, high level (top down), concepts, hypotheses, and models that can be imposed on the available data: how can evolutionarily successful (robust or stable) system of interactions, and mechanisms be encapsulated and transmitted; why are things the way they are?
- the development of practical methods which can both validate existing systems and suggest new ones: how can methods from network analysis and spectral theory for large interaction matrices be employed to suggest distinct types of protein-protein (or gene-gene) group associations (mapping to possible functions) in an efficient manner; including measures of stability for such results?

We present some background, ideas and illustrative responses to each of these questions, filtered, of course, through our own viewpoints and experience. Though early days, we suggest that such consequent activity could be embedded usefully within some of the various systems biology projects taking place over the next few years.
1 Introduction

Post-genome research is focussed on understanding how the variations encoded within genome sequence relate to phenotypical functions - from the cellular level upwards. From the start almost all of this work was driven by the data challenges in moving from sequence to gene to protein to metabolic process to function to organism. In that sense it was “bottom up”. Relationships between individual variations and the increased tendency towards (likelihood of) development of phenotypes have been identified directly in some cases. Such early successes act as exemplars or touchstones in guiding our impressions of the larger picture. Recently sequencing of many individuals has revealed many more divergences and even more sources of diversity, including amplified parts of chromosomes – without apparent disease consequences.

Genetic diversity, in turn, leads to complex differences in phenotype - visible at the single gene level by the variance of transcription being negatively correlated with the closeness of relationships and is smallest for identical twins [Cheung, 2003]. In genetics one of the most useful example of mathematical biology, until the last few years, has been the study of genetic linkage (see, for example, [Lander 1987]). At the genomic scale the real power of mathematics has yet to gain full acceptance among biologists.

Taking the bottom up approach represents a series of huge challenges. How do we deal with the amount of data available for individuals (full genotyping) versus the relatively small number of case/control units available for specific phenotypes? Whilst a function controlled by a single mutation can be identified, by a search through the genome, how do we deal with the combinatorial explosion of possible cooperative/interactive relations between genes (epistasis)?

The larger picture is of the “many to many” relationships between the combinations of possible sequence and thence protein variations, and the myriad of possible functions, mediated by highly variable cellular environments. “Systems biology” is the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions. Rather than focusing down on the individual components or aspects of the organism, systems biology focuses on all of the components as one system. A complication is that some system properties are not identifiable within a single cell; for example, a cardiac pacemaker which depends on the construction of membrane polarization across many cells, leading to a counter flow of ions as the neurons fire.

Before diving down into any particular system, there is a higher level challenge to think about: how can evolutionarily successful (stable) system of interactions, commands and controls be encapsulated and transmitted: why are things the way they are? And how might the self-organisation of the system itself be achieved or structured? By starting top down and imposing alternative organisational conceptual models onto the experimentally obtained data we might gain some insights - or at least develop some insightful questions. This approach is almost always a productive one within science – and usually proceeds long before the data collection is complete. For example, the imposition of the structure of the periodic table onto the sequence the elements.

In such a “top-down” approach, we may start out from a theoretical perspective, and suggest alternative ways, “models”, to describe the relationships between proteins and function; and then consider these alternatives in the light of current knowledge and
planned data exploitation. This may allow us to test, or discriminate between, the alternative models.

We have an emerging partial picture of the whole proteome, with protein pairs connected via an “interactions” network. In a recent review, questions and challenges were presented (Grindrod 2005). Networks based on observations must contain errors and this fact itself should guide the analytical approach – since over reliance on the veracity of the data will doom any approach. Fortunately since the earliest attempts to characterize such networks it turned out that global parameters are relatively robust. Perhaps more surprisingly the methods of associating genes or proteins by their connected “role” within the network (the so called inverse problem – seeking to infer co-function via total associations) may also be robust to levels of error [Grindrod 2002].

2 Background

The ongoing development of high through-put methods of data collection means that a variety of data types will be available for analysis. Genes code for proteins which when present control the chemical pathways providing the signaling and functional responses to enable cells of distinct types to respond to their environment dynamically. The systems biology paradigm challenges (bio)analysts to consider functions at the cellular or organism level as being derived from the integrated and interacting (sub) network of genes, proteins and biochemical reactions: it inverts the usual bottom up or reductionist processes of analysis in favour of a systems definition and modelling paradigm. What type of rules or features might such systems exhibit? What should this mean in practice? What are the implications of the insights to be distilled directly from the data?

There were three catalysts for the ideas and suggestions presented in this paper.

1. In [Grindrod 2005] the authors made an exhaustive review of many aspects of the use of networks and graph theory within proteomics presented in around 40 recent (for then) publications. Such a review is invaluable for new entrants in getting up to speed with the subject. The key issues and suggestions identified there were as follows – and the present paper is in part meant to respond to these.

| o | It should become common practice to present and publish network (graph theoretical) properties of protein interaction networks, such as the degree distribution, the clustering coefficient, and so on, calculated according to their standard definitions. |
| o | There are a number of alternative model candidates for characterizing large interaction networks and, since this is an ongoing area of theoretical research, it would be useful for datasets to made easily available to the mathematical and physics communities. |
| o | Experience of network analysis shows that global and emergent properties of networks are critically dependent on small scale structure. |
| o | Biologists should attempt to analyse large scale networks themselves and avoid premature reduction to subsets of proteins: mathematicians and physicists should be encouraged further to find opportunities to collaborate. |
The many to many nature of the proteins to functions relationship requires a network approach not a clustering type of methodology. The availability of protein-interaction network data is likely to challenge and force the pace of research within network theory in the next few years from [Grindrod 2005].

2. We are part of a collaboration between the universities of Strathclyde and Bath and the Beatson Institute in Glasgow, aimed at introducing methods of numerical linear algebra and spectral theory to the analysis of large networks defined from co-expression data. The methods have been used successfully to identify clusters of related samples and genes in a study of oral cancer [Hunter 2006]. The task of recognizing distinct classes of structure within such networks and analyzing the stability of the results to errors propagating through these systems is key, both in providing bioscientists with actionable analytical options and, crucially, in providing associated measures of confidence. This project, which is EPSRC funded, has also enabled individuals within mathematical sciences to develop their knowledge and relevance to current data-driven analytical problems. Some of the ideas here have resulted directly form discussions within this collaboration.

3. There has in the last few years been a one-time increase in funding, for the subject of systems biology: a paradigm that focuses on all of the active components within a cell or an organism as single system; most likely complex and nonlinear. From an analytical (mathematical) point of view, the analysis of large high-throughput data sets appears to be still in its infancy, and so, in responding to the systems biology challenge, there is a need to cast the net widely in developing and testing models and methods. Therefore we suggest that such programmes should contain some element of both conceptual modelling as well as the development of consequent data analytical methods. This is an immediate opportunity. In a few years time the appointments will have been made and the money spent, so this suggestion is timely for both the funders and the principal investigators. In 2001 Grindrod reviewed the UK spending of the research councils on bioinformatics by surveying all those in receipt of grants at that time [Grindrod 2001]. Then the focus of bioinformatics was found to be biased towards the data collection, data normalization and curation, the management of online access to data bases, and perhaps some “dice and slice” or “drill down” tools for user access (remote or otherwise). Very few smart methods were available to take a data driven view of the “system” under observation. Most of the more sophisticated work was on gene sequences, rather than on analysis of microarray data, and so on. Much of the analysis was being developed within silo (within genome or within programme). At that time, the limit of the data-analyst's ambition was perhaps to use the unsupervised clustering methods popularized in an early paper by Eisen et al, with the code available from the Stamford website [Eisen 1998], or something similar. It is essential, in our view, to use the systems biology opportunity as a vector to move on to more modern, realistic and practical concepts and methods.
An evolutionary successful genome

Let us think ahead of the data supply: what do we expect? Any functional genome, active within any individual living organism, has to satisfy some fairly straightforward requirements:

- **Completeness.** It should encode the information to build all cellular functions, and tissue types, and control the developmental process.
- **Efficiency.** It must be relatively succinct – there are far more functions than genes it seems: so the genome cannot be a list of instructions.
- **Robustness.** It must be robust, both to replication during cell division, and to being successfully spliced with another working genome during the processes of reproduction.
- **Evolutionary memory.** Logically the information at work today must be descended from similar information that defined not just individuals from directly preceding generations, but the genomes of all evolutionary forerunner organisms, backwards in time. Much of this is likely to be redundant and therefore becoming imperfect and perturbed at no cost.
- **Managed failure.** It must have the ability to fail gracefully and completely when some element or other cannot be achieved due to damage, noise, or external intervention.

Though perhaps obvious, these together suggest a number of corollaries. Firstly, if the genome information is something like an executable application to be installed and run within an “environment”, then both efficiency and robustness demand that there must be substantial re-use of certain sub-functions, like subroutines or objects, playing a similar role within context. This de-risks the chance that one or more instances of the same mechanism will fail due to noise or perturbation: they either all function or they all fail. In biological terms this is “cross talk” between functions where distinct processes may share (re-use) components within their subsystems. We may see distinct signaling pathways sharing a number of proteins, together providing a specific mechanism that is similar in each case. It is interesting to speculate whether the more basic (sub) functions that become re-used within many higher order processes are those that are older within an evolutionary context (shared across genomes).

Secondly, it seems reasonable (from the perspective of our thought experiment) that there might exist some proteins which are “more essential” than others either
- because they are part of a mechanism that is reused often within a number of functions, or
- because they have an evolutionarily novel role that is highly organism specific, and therefore are “mission critical” to the particular organism.

Can we differentiate between these distinct two types of essential role? The term “lethality” is often used in the literature to describe proteins whose presence appears necessary. Are these proteins more highly connected (part of some mission critical command and control function), or are they embedded within more fundamental and hence highly reused subsystems? Hence there would appear to be two distinct possible types of “lethality”: do we see one or the other or both?

We see such phenomena very often in theory – indeed this is common in graph theory where microscopic structural changes made gradually (controlled by a graph’s defining parameters) result in phase changes (rapid switches) in the macroscopic properties –
this is often termed as emergent behaviour or self-organization. For example, scale-free networks, where the degree distribution (the probability \( P(k) \) that any given vertex has degree \( k \)) behaves like an inverse power law, have a high degree of robustness against random errors. Upon removal of randomly chosen nodes from the network, the diameter (which characterizes the ability of two nodes to communicate with each other) is affected less than when perturbing an exponential network (a network whose degree distribution decays exponentially). However scale-free networks are highly vulnerable to perturbations of their few highly connected nodes; the diameter increases rapidly and the network breaks into many isolated fragments, [Albert 2000]. This leads the authors in [Joeng 2001] to check whether a correlation between lethality and connectivity exists in their protein interaction network. They correlated the number of links with the phenotypic effect of removing an individual from the yeast proteome. The findings imply that highly connected proteins with a central role in the network's architecture are three times more likely to be essential than proteins with only a small number of links to other proteins. Hence the authors argue that despite the importance of individual biochemical function and genetic redundancy, the robustness against mutation is also derived from the organization of interactions and the topological positions of individual proteins.

More recently there have been some exciting further attempts to relate proteins’ properties such as lethality and viability (according to whether an organism can survive their removal or not) to graph/network properties [Demetrius 2006]. This is a highly promising field. Should systems biology projects contain some elements of this type of work at a generic level, prior to focusing on specific systems?

An evolutionarily successful design for a genome ought to exhibit other features too. For example having some proteins that can, on occasions, substitute for one another would be a highly useful property (indeed it is a response to lethality). Then if a particular mutation occurs all is not lost. It may also be possible that whole sub-functions might substitute for one another providing an alternate means of chemical signaling (back-ups) within specific cellular functions.

4 Data driven characterization and systems definition

To be definite let us consider the simplest type of protein interaction data that may be available. Much of what we say here also applies to gene expression data and co-expression networks that can be suitably defined, based on microarray data. Inevitably at a conceptual level and a practical (modelling) level we will have to make some assumptions and many of these will be (convenient) approximations to the truth: therefore such assumptions need to be made explicit as caveats.

Suppose we have a known set of \( N \) proteins, which will be visualized as \( N \) vertices of a graph. Typically there will be thousands or a few tens of thousands of proteins known to be encoded by the genes. Let us suppose that pair-wise protein interaction data is available for all possible pairs. A positive interaction of some sort will be represented by an edge in our graph. Typically this will be some measure of when the proteins can physically interact together in some suitable chemical environment. Here there is an opportunity for ambiguity: the interactions themselves are most likely highly tissue dependent. Nevertheless for the purposes of our thought experiment let us suppose that a well defined set of conditions has been employed. The result is a large undirected
A graph that likely contains some false positives (edges that should not be present) and false negatives (missing edges). Let us assume that there a relatively few of these but keep in mind that any analytical methods too dependent on individual paths are likely to be a risk. Typically some of the proteins will have known functional roles attributed to them, validated by direct discovery and observations; whereas many proteins will have no such attribution.

How does our present graph compare with other similar graphs, for other tissue types or organisms, or from other data sets? Are all proteomes structurally similar or do they contain distinct idiosyncratic characteristics? To answer such questions one needs to characterize the global structure of the whole.

We believe this is best done by representing (modelling) the graphs in question within one or more suitable class of graphs and deriving the macroscopic parameters that define or calibrate the specific instance. To be definite, suppose that we have a class of graphs which is dependent upon a number of parameters. (For Erdos-Renyi random graphs there is but one such parameter – the constant probability that any particular edge is present). Then we must estimate values for these parameters from the data. Often this is best achieved by fitting some parameters to estimates derived directly from the data, or by maximizing the likelihood of the data as a function of the parameters. What is the probability of observing the data given any possible parameter values? This is a classical likelihood approach, and alternative classes of models (hypotheses) can be compared using likelihood ratios or multiple hypothesis testing [Jaynes 2003].

For example in [Grindrod 2002] the author introduced a class of “range dependent” graphs based upon a one dimensional lattice ordering: all vertices are indexed by an integer, and then the probability that any vertex \( i \) is connected to any other vertex \( j \) is some monotonically decreasing function, \( f \), of the range, \( |i-j| \). For many choices of \( f \), including Grindrod’s two parameter original, \( f(|i-j|)=\frac{?}{?}-|i-j| \), for some real macro parameters \( ? \) and \( ? \) in \((0,1))\), the average degree (and other degree distribution moments) and the Watts-Strogatz clustering coefficient (the tendency for association to be transitive [Watts 1998]) are calculable. Hence the free macro parameters can be determined directly by fitting them (or hybrid parameters) to estimates from the data.

When datasets (graphs…) are published it should become routine for researchers to publish some of the properties of the network: the degree distribution, the Watts-Strogatz clustering coefficient, and so on. There is much interest in identifying scale free networks (or at least networks showing scale free over some range of vertex degrees), where \( P \) decays like a negative power of \( k \) over a wide range of \( k \) values.

The visualization of networks, as trees or pin balls, is of little use to researchers – though such figures emphasize the size of the task at hand, they are more often used to justify closing down the focus onto a reduced system.
Models, models everywhere...

Many random graph models could prove useful for characterizing networks. Such models, that is, formulas for generating edges, may be calibrated and compared by matching their topological properties against real target protein-protein interaction networks. Besides general classes of scale free, small world and other random graphs, some models have been developed specifically for analyzing such interaction networks. These include:

- **Range dependent networks [Grindrod 2002]**: the probability that any vertex is connected to any other vertex is some monotonically decreasing function of the range, controlled by a few global parameters.

- **Geometric models**, where proteins are assigned random locations in Euclidean space [Przulj 2004], and connections are defined between proteins that are sufficiently close together.

- **A lock and key model**, defined by bipartite graphs [Morrison 2006], where proteins are randomly assigned binding domains and connections arise between matching domains. The ideas behind this model can be traced back to [Thomas 2003]. See also related models in [Caldarelli 2002], [Deeds 2006] and [Przulj 2006a].

In addition to measuring global properties such as average path-lengths, clustering coefficients and degrees, more sensitive local network properties can also be used to fit parameters and help decide between competing models. The powerful concept of graphlet frequencies—the relative abundance of possible subgraph building blocks—is developed and tested [Przulj 2006b].

Search engine concepts, where graph theory and algorithmics has had a huge impact in helping to tame massive data sets, may also be relevant. In the case of gene-gene networks defined from microarray data, [Morrison 2005] showed that the same ideas prove worthwhile.

A useful overview of networks in biology is given in [de Silva 2005] and Alon's book provides a recent general reference [Alon 2006].
Defining and validating (sub) systems

Representing given data (a list of edges) as a particular realization of a graph generated within such a class, with the fixed estimates of the macro parameters, constitutes a tricky "inverse problem". Consider Grindrod's range dependent graphs, introduced above, where each vertex must be represented by an integer: then the resolution of the inverse problems looks like a rather nasty search over all possible N! orderings of the N vertices, as we seek to maximize the likelihood of the observed data. However, it can be shown that spectral methods (relaxation methods) applied to the Laplacian matrix (associated with the adjacency matrix) can produce an approximate solution to this inverse problem [Higham2003a]; this method also connects us directly to a large body of work on unsupervised clustering using the Fiedler vector [Feidler1975, Higham2003b].

Immediately, using these ideas, all of the proteins are embedded into some ordering that typically contains both small and large clusters within which the pair-wise interactions are (relatively) transitive. Many proteins of unknown function are therefore amenable to a global game of guilt by association and sub systems of proteins, corresponding to some function or other can be discovered. The work of recognizing special sub graphs within large networks is important for both undirected and directed sub-graphs.

In [Grindrod 2003] the inverse problem was applied to the yeast proteome data containing around 3500 distinct proteins. Very many proteins “of unknown function” were embedded within (near) cliques.

Recently [Vass 2006] used an unsupervised clustering method to validate a system for leukemia and lung cancer datasets, and in both identified a highly associated sub graph of about 100 genes known to be responsible for controlling DNA copying and cell-division. While information on a few of these would be unremarkable, the automatic rediscovery of relationships gathered over several decades gains credibility for the approach. From this sub graph identification of (mostly) negatively correlated genes was trivial. Further support for the biological relevance of the methods was provided by the fact that the second group of genes in turn contained genes with identified roles in suppressing the DNA copying genes. Two extremely complex biological patient datasets for different diseases revealed essentially the same information using spectral methods to successively reorder genes. The spectral method performed better than an expert in identifying the DNA related genes, it is unlikely that any biologist would have started out by identifying all of the targets to analyze this graph using her prior knowledge – there are simply too many genes for anyone to cast their nets so widely. The second group, though it has many genes that “make biological sense”, also contains several that evoke a speculative response, are they new targets for the control of cell division?

Searching for particular structures within an adjacency matrix is equivalent to some unsupervised discrimination methods [Higham (2003b)]. [Grindrod 2002] showed directly by recovering graphs originally generated within the class and subsequently “shuffled” to hide the underlying structure, that the solution and the methodologies are robust to some false negatives and false positives. Hence the discovery of highly associated sub graphs is itself robust to some data error.

Moreover in recent work [Spence 2006] has shown how the sensitivity of gathering subsets of associated proteins form a larger network can be tested and quantified. It is not enough to have some kind of highly associated cluster – one should consider the
significance of such discrimination in the light of errors in the original data which may have been propagated. Fortunately some messy linear algebra can provide alerts to poor practice. So not only can bioscientists employ such state of the art methods to identify sub sets of the proteome that contain proteins of interest, they could also be given health warnings. The important topic of cross-talk adds extra interest to the probabilistic membership of a cluster where the role of two or more groups of proteins may depend on shared components.

Many bioscientists will have a very good idea of some of the proteins that are at work within the functional (sub)-system of their own interest: proteins already known to control or influence some element of a particular function. However we feel strongly that it is a cardinal mistake to close down the system definition too quickly and focus on one’s desired set of proteins. Instead some method or other such as those discussed above should be applied to a much larger dataset (the whole proteome if possible) to demonstrate that the particular subset of proteins to be considered is suggested (and therefore validated) directly by the data, independent of the subjective knowledge of the analysts. This would give the users the confidence that there are no other proteins that could have been included (on the basis of their ability to interact with those proteins within the system). In short the practice of (sub)system definition (in which analysts become focused upon some subset of the proteome) requires some hard evidence that justifies the a priori exclusion of all other proteins.

6 Summary

In this short paper we have discussed the need for some speculative “top down” concept development, and offered some thought experiments, designed to suggest how genomes have developed successfully, and what features and properties they might display. The formulation and testing of such alternative hypotheses should, in part, drive future data collection and experiment.

We have suggested that some of the features observed within interaction data, such as cross talk and substitution, are natural consequences of a successful and robust encoding of the information necessary to define an organism. Furthermore we have also suggested that current notions about lethality may need to be better defined or classified; and that substitution (back-up sub-systems) is a natural evolutionary response to earlier lethality. What further corollaries of distinct possible mechanisms (by which such information might be held) could be suggested and tested?

Next we have considered the problem of analyzing gene or protein interaction data and discussed some of the alternative (mathematical) models for interactomes available. These and their defining parameters should be standardized and used to distinguish between what is common and what is specific across genomes. Moreover we have discussed various approaches to data-driven systems definition (unsupervised discrimination) and illustrated how these have been used to suggest (sub)-systems corresponding to co-functional roles and also validate the subjective and experimental based judgments of bioscientists. There is a role for both approaches: they are complementary not exclusive, and should be employed to justify sub-systems considered within systems biology programmes and projects.
We have also considered the need for the mathematics of such analyses to become robust and contain considerations of stability (sensitive dependence on specific elements within the data). By presenting just some of the alternative methods of attack to distinct data driven problems our selection has been subjective and hence not exhaustive. Rather than acting as a general review, this paper is intended to stimulate those working at the interface between systems biology and mathematical methods, and to inspire new contributions.

Acknowledgement
This work was supported by EPSRC grant GR/S62383/01.

References


