Systems Biology
Doctoral Training Centre

D.Phil. Research Proposal: Evolution, dynamics and modularity in biological networks

Sumeet Agarwal*
Supervisors: Charlotte Deane*, Nick Jones*, Mason Porter*

January 14, 2009

Background

Recent years have seen the collection of vast quantities of data on expression levels and interactions of genes and gene products in various organisms. Such data has been analysed and modelled in many ways. One of the most popular paradigms for modelling interaction data has been to look at it as a network of linked components [30]. This approach has led to the construction of several types of biological networks, such as those for gene regulation, protein interactions and metabolic reactions. In this project, our broad aim will be to look at ways of integrating different kinds of experimental data in order to get insights into the functional organization and dynamics of biological networks. Another goal will be studying the evolutionary history of biological networks: an understanding of the mechanisms of their evolution may help us to explain some of their observed properties and how they relate to network robustness.

Our current focus is on protein-protein interaction networks, which represent experimentally observed physical binding interactions between proteins in a cell (these are collectively referred to as the *interactome*). The development of high-throughput screening techniques has led to the compilation of large interaction datasets, in particular for yeast. The two major experimental methods used are yeast two-hybrid (Y2H) screening [12,13,24,36], and tandem affinity purification followed by mass spectrometry (TAP/MS) [15,23]. The quality and reliability of available datasets is a major issue; recent studies [37] indicate that the properties of Y2H and TAP/MS data differ substantially, with the latter mostly picking up interactions that are part of protein complexes, whilst Y2H is better at capturing more transient binary interactions. Consequently, interaction networks constructed from the two kinds of data also tend to have different characteristics, and one of the key problems in this area is how to combine insights from the different experimental sources in order to obtain a comprehensive picture of the interactome's organisation.

One of the major theoretical approaches to modelling sets of interacting elements in various domains has been to look at them as graphs or networks [30]. Here the elements are represented by nodes, and interactions are represented by links between nodes. In particular, for protein interaction networks, the usual method is to have a node for each protein, and place a link between all pairs of interacting proteins. Typically, no direction

*sumeet.agarwal@dtc.ox.ac.uk, deane@stats.ox.ac.uk, Nick.Jones@physics.ox.ac.uk, porterm@maths.ox.ac.uk

or weight is attached to these links; thus the interactome is generally modelled as an unweighted, undirected network. Various properties can be studied for such networks - a particularly relevant one in this context is that of community structure [11, 17]. Roughly speaking, a *community* (also sometimes referred to as a cluster or module) is a set of nodes with a higher than expected number of links amongst them, as compared to links to nodes outside the community. Many real-world networks have been shown to possess significant community structure [30], as compared to random networks with the same distribution of node degrees. In particular, several studies have pointed towards modular organisation of the proteome [16, 20, 26], with densely interacting groups of proteins being responsible for specific functions and processes. Community structure has also been seen in metabolic networks; a study by Guimerà and Amaral [19] showed that metabolites could be assigned distinct roles in the network by means of a topological analysis based on communities.

The dynamics of these biological networks are also important: for example, it is clear that the interactome is not a static network, but in fact the actual set of proteins and interactions active at a given time and place is highly dependent on the physiological conditions. Looking at gene expression levels using microarray data [14, 25] can tell us which proteins are being expressed in a given situation. In a pioneering study, Han *et al.* [20] combined protein interaction and expression data to show the existence of two kinds of 'hubs' in the protein interaction network of *Saccharomyces cerevisiae*: party hubs, which interact with a number of partners simultaneously, and date hubs, which interact with several partners at different points in time. It was posited that these two kinds of hubs may play a key role in the modular organization of the interactome, with party hubs coordinating a given function and date hubs serving to link different functions. However, there has been considerable debate over the validity of the date/party hub distinction [4–6], with some studies suggesting that it may at least in part have been an artifact of the particular datasets and statistical techniques used, and that the actual range of 'roles' taken on by hubs is more diverse than a simple dichotomy. For instance, Komurov and White [26] showed that another kind of distinction, that between statically and dynamically expressed proteins, may also be of significance. Since the notion of node roles in a network is essentially a structural one, it may better to assess these by way of structural metrics, as in [19], rather than by using statistics based on expression data.

There are several other approaches to looking at network dynamics and integrating multiple kinds of information. Recent work by Hegde *et al.* [21] has sought to use microarray expression data to reduce the interaction network to an 'active' subnetwork under different sets of conditions (i.e., the subnetwork contains only interactions between proteins being expressed in that condition), and then study the differences between these subnetworks to learn how the roles of different genes are affected by a changing cellular environment. Maraziotis *et al.* [28] take the alternative route of clustering genes by expression data and using these clusters to assign weights to known protein-protein interactions, followed by community detection on the weighted network. Another type of data which has recently been made available relates to phenotypic effects of gene deletion. Hillenmeyer *et al.* [22] look at the growth response of single-gene deletion strains to a wide range of chemical and environmental stress conditions, such as the presence of drugs. They find that 97% of gene deletions exhibit a measurable growth phenotype relative to the wildtype, and also show that meaningful gene groups can be extracted by clustering the phenotypic profiles across different conditions. Such data can serve as another source of information for the analysis of proteome organisation.

Finally, some recent work by Bonneau *et al.* [10] has attempted to take these approaches further by constructing a predictive model of dynamic interactome response to environmental changes for the archæon *Halobacterium salinarum NRC-1*. This is done by data-driven discovery of functional and regulatory relationships amongst genes and abiotic environmental factors. The model is able to accurately predict dynamic transcriptional responses of genes for a number of experiments representing completely new genetic backgrounds and environments. The success of this kind of modelling suggests a very high degree of completeness in the constructed networks, and provides a possible way out of the perennial problems of data quality and completeness. There have also been other promising efforts to integrate multiple types of data and use them to construct network-based models for predicting expression and phenotypic responses to genetic perturbations and environmental changes [27, 38]. Approaches of this sort may be a pointer to the future direction we need to take in order to achieve a systems-level understanding of the functioning of biological networks.

Research Plan and Initial Results

Our focus in the initial part of this project has been on looking at node roles in the protein interaction network, in particular in the context of the date/party hub hypothesis. We have been working largely with the protein interaction datasets which were used to perform the experiments leading to the hypothesis [6, 20]. Our attempt has been to examine structural properties of the network and of individual nodes within it, and see whether these are in consonance with a date/party type dichotomy. Towards this end, we initially partitioned the network into communities, via maximising one of the metrics which has been devised for quantifying community structure, called *modularity* [32]. This metric formalises the notion of within-community links being more frequent than one would expect in a corresponding random network. The general problem of finding the optimal partition into communities as per this metric is computationally intractable [7]; but there exist many approximate optimisation algorithms. The one used by us was the spectral partitioning method proposed by Newman [31]. We found that most of the communities discovered in this network are meaningful in the sense of having a high degree of functional homogeneity, as measured by shared protein annotations from the Gene Ontology (GO) database [2]. We then looked topological measures for each node, such as the number of links within its community and the distribution of links across communities, and based on these categorised nodes into roles as per the method of Guimerà and Amaral [19]. We found that the proposed date and party hubs did not fall cleanly into any particular role or subset of roles, and also that there was no apparent bimodality in the distribution of hub roles but rather a more evenly spread continuum of varying roles [1].

Based on these results, obtained for multiple datasets and also verified by looking at other measures of node importance such as centrality measures [33], there seems to be good evidence for refuting the date/party hub hypothesis for protein interaction networks. More recently, we have been looking at possible alternative ways of understanding the different structural functions of proteins in such networks. One such possibility is to look at centrality measures on links (i.e., interactions) rather than nodes. Using a standard measure known as betweenness centrality [17, 30], which measures how important a link is to global network connectivity, we have found a fairly strong inverse correlation between a link's betweenness and the functional similarity of the two proteins it links. There is also some indication of a centrality 'threshold', a point beyond which there is a sudden drop in mean functional similarity (again assessed based on GO annotations); and this threshold appears to be a function of the network size. This leads to the idea that there may broadly be two types of interactions, those occurring within communities of functionally linked proteins and those serving as bridges between such communities. This notion also corresponds fairly well to the long-established theory of 'strong' and 'weak' ties in the social networks literature [18, 34].

In the short-term, our primary aim will be to see if similar results are replicated across multiple datasets for different species (our current results are for relatively small yeast datasets), as well as examining whether this sort of link categorisation might be a general property of modular networks from various domains, or whether it is in some way specific to our protein/biological networks. Also, there is evidence to suggest that one of the reasons for observing a date/party hub distinction in some cases may be the combination into a single dataset of interactions from both Y2H and TAP/MS experiments, which tend to differ markedly in their properties [37]. We will look at the relation between data source and link centrality properties as well, along with the question of reliability of data from different sources, which we will try to assess based on network structure [8].

A key goal for the short to medium-term will be to look at approaches to data integration, as indicated earlier [10, 27, 38]. We now have access to a number of datasets relating to proteins and the genes that code for them, giving us information about their interactions, expression levels, localisation, response to environmental changes, knockout phenotypes and so forth. A major challenge is to devise appropriate frameworks and algorithms to be able to build models that draw on all of this diverse knowledge. Bonneau et al. have taken some steps in this direction: they stress the importance of appropriately integrating data obtained from various high-throughput technologies into a comprehensive model that can quantitatively predict how an organism's cellular networks will interact with the environment and what responses this will lead to [10]. To start, we will attempt to look at this issue largely in the context of the organisation of the proteome and the roles of different proteins, and how these things are affected by changing physiological conditions. Related to this, another medium-term task will be to formulate a more appropriate notion of 'roles' for proteins, which might serve as an alternative to a date/party hub conception. Such a notion may take into account multiple levels of or-

ganisation in the interactome, such as the possibility of communities themselves being grouped. We will attempt to use the recently proposed framework of hierarchical random graph models [9] to study multi-level structure in interaction networks. It may turn out that in terms of network structure, it is more appropriate to talk of roles for protein interactions rather than for individual proteins; this is perhaps suggested by our current results.

In the longer term, one of our aims is to look at the evolutionary processes that may have shaped the structure and dynamics of biological networks. Comparison of interactome organisation across different species can give us some insight into how it evolves; for instance, a study on available human data shows enrichment for interactions between proteins from the same evolutionary lineage, which may suggest a mechanism of preferential attachment between such proteins [35]. If we observe properties like date and party 'hubness' for certain proteins, we might also ask if such properties are biologically conserved. In general, can observed variation and conservation in these networks be explained by some of the proposed models of network evolution [3]? It has been shown that a machine learning approach based on frequencies of small motifs can be used to confidently predict growth mechanisms for the protein interaction network of *Drosophila melanogaster* [29]. We will look to use similar approaches to predict such mechanisms for other species, and for change across species. It would be of interest to study appropriate evolutionary models, or formulate them if required, to see what they imply about network functionality and robustness.

Another long-term goal will be to work towards a unified framework for modelling and understanding the various kinds of biological networks that are currently studied, such as gene regulatory networks, protein interaction networks and metabolic networks. This can be seen as a natural extension of the data integration approach mentioned above; it is clear that the entities in each of these networks are linked, and there is a continual flow of information between them. Ultimately, we might like to be able to think of an entire biological system (which may be a cell, an organism or even an entire population or society) as being modelled by one single network (or perhaps by some more complicated formal structure derived from it), which would have many different levels of organisation. It would be interesting to consider what the basic elements of such a structure might actually represent: would they be things like genes or proteins, would they be of multiple types, or would they perhaps need to capture some more abstract notion of biological building blocks? It is unlikely that definitive answers to such questions will be available any time soon, but we hope in the course of this thesis to take a few steps towards this ultimate end.

Proposed Timeline

- Next 3-4 months (till April 2009): Study edge centrality properties for bigger yeast protein interaction networks and networks in other species; look at how these compare to observations on real-world networks from other domains, such as social networks. Look at variations by experimental source and data reliability. Collect, write up and submit for publication results on evidence against date/party hub view of proteins, and alternative notions such as weak/strong type interactions in protein networks.

- Next 8-9 months (till September 2009): Start to study data integration approaches with a particular view to improving the understanding of protein interaction networks. Combine at least some of expression, genetic interaction and knockout phenotype data with interaction data to create more comprehensive networks; study community structure and other structural features in these networks to better categorise proteins and protein modules. Also look at hierarchical structure models for these networks and their possible use in refining the notions of protein roles and functional groupings.

- 3rd and 4th years (till September 2011): Extend further the work on data-driven integrated models, including looking at predictive models of system response to environmental changes. Enlarge the 'system' by attempting to incorporate multiple interaction types like genetic regulation, protein-protein and metabolic reactions into the same network-based model. An intermediate stage towards this may be a sort of 'network of networks', with entities of the different types represented in distinct networks with some linkages between them. Also, look at models for evolution of biological networks, based on comparative studies of network properties across species. In the final year, collect results and observations from the different strands and present them as a coherent thesis.

References

[1] Sumeet Agarwal, Charlotte Deane, Nick Jones, and Mason Porter. Dynamically organised modularity in protein interaction networks. Project report, Systems Biology DTC, University of Oxford, 2008. Available at http://www.physics.ox.ac.uk/cm/cmt/agarwal.

[2] M. Ashburner et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genet.*, 25:25–29, 2000.

[3] Albert-László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004.

[4] Nizar N. Batada et al. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biology*, 4(10):1720–1731, 2006.

[5] Nizar N. Batada et al. Still stratus not altocumulus: Further evidence against the date/party hub distinction. *PLoS Biology*, 5(6):1205–1210, 2007.

[6] Nicolas Bertin et al. Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, 5(6):1202–1205, 2007.

[7] Ulrik Brandes et al. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.

[8] Pao-Yang Chen, Charlotte M. Deane, and Gesine Reinert. Predicting and validating protein interactions using network structure. *PLoS Comput Biol*, 4(7):e1000118, Jul 2008.

[9] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[10] Richard Bonneau *et al.* A predictive model for transcriptional control of physiology in a free living cell. *Cell*, 131:1354–1365, 2007.

[11] Santo Fortunato and Claudio Castellano. *Encyclopedia of Complexity and System Science*, chapter Community structure in graphs. Springer, 2008.

[12] Micheline Fromont-Racine, Jean-Christophe Rain, and Pierre Legrain. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet.*, 16(3):277–282, 1997.

[13] Micheline Fromont-Racine et al. Genome-wide protein interaction screens reveal functional networks involving sm-like proteins. *Yeast*, 17(2):95–110, 2000.

[14] Audrey P. Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

[15] Anne-Claude Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[16] Anne-Claude Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, 2006.

[17] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, 99(12):7821–7826, 2002.

[18] M. S. Granovetter. The strength of weak ties. *Amer. J. of Sociology*, 78(6):1360–80, 1973.

[19] Roger Guimerà and Lúis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[20] Jing-Dong J. Han et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

[21] Shubhada R. Hegde, Palanisamy Manimaran, and Shekhar C. Mande. Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Comput Biol*, 4(11):e1000237, Nov 2008.

[22] Maureen E. Hillenmeyer et al. The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science*, 320(5874):362–365, 2008.

[23] Yuen Ho et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[24] Takashi Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, 98(8):4569–4574, 2001.

[25] Patrick Kemmeren et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, 9:1133–1143, 2002.

[26] Kakajan Komurov and Michael White. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol. Sys. Bio.*, 3:110, 2007.

[27] Insuk Lee et al. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature Genet.*, 40(2):181–188, February 2008.

[28] Ioannis Maraziotis, Konstantina Dimitrakopoulou, and Anastasios Bezerianos. An in silico method for detecting overlapping functional modules from composite biological networks. *BMC Systems Biology*, 2(1):93, 2008.

[29] Manuel Middendorf, Etay Ziv, and Chris H. Wiggins. Inferring network mechanisms: The Drosophila melanogaster protein interaction network. *Proc. Natl Acad. Sci. USA*, 102(9):3192–3197, 2005.

[30] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[31] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

[32] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[33] Nicola Perra and Santo Fortunato. Spectral centrality measures in complex networks. *Phys. Rev. E*, 78(3):036107, 2008.

[34] Anatol Rapoport. Contributions to the theory of random and biased nets. *Bulletin of Mathematical Biophysics*, 19:257–277, 1957.

[35] Jean-François Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

[36] Peter Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[37] Haiyuan Yu et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, 322(5898):104–110, 2008.

[38] Jun Zhu et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genet.*, 40(7):854–861, July 2008.