Algorithms for protein interaction networks

M. Lappe*[*] and L. Holm†[1]

[*]Max-Planck Institute of Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany, and †Institute of Biotechnology and Department of Biosciences, P.O. Box 56, FI-00014, University of Helsinki, Finland

Abstract

The functional characterization of all genes and their gene products is the main challenge of the postgenomic era. Recent experimental and computational techniques have enabled the study of interactions among all proteins on a large scale. In this paper, approaches will be presented to exploit interaction information for the inference of protein structure, function, signalling pathways and ultimately entire interactomes. Interaction networks can be modelled as graphs, showing the operation of gene function in terms of protein interactions. Since the architecture of biological networks differs distinctly from random networks, these functional maps contain a signal that can be used for predictive purposes. Protein function and structure can be predicted by matching interaction patterns, without the requirement of sequence similarity. Moving on to a higher level definition of protein function, the question arises how to decompose complex networks into meaningful subsets. An algorithm will be demonstrated, which extracts whole signal-transduction pathways from noisy graphs derived from text-mining the biological literature. Finally, an algorithmic strategy is formulated that enables the proteomics community to build a reliable scaffold of the interactome in a fraction of the time compared with uncoordinated efforts.

Introduction

One of the oldest paradigms in molecular biology is the concept of ‘one gene – one protein – one function’. This notion has been captured in hierarchical classification schemes such as the EC-number or the Gene Ontology. It has become clear that the above paradigm is far from the whole truth. Many genes have alternative splicing variants and various post-translational modifications, some proteins have several binding sites and catalyse a variety of different reactions, and the cellular processes or developmental stages in which the proteins are involved are not accounted for at all. The complexity of protein function is captured to a great extent in the biological literature. However, this kind of knowledge is inaccessible to computer algorithms in the sense that it is unstructured information.

The post-genomic view defines protein function in the context of complex networks of specific interactions. This view of a ‘society of proteins’ has been proposed previously [1–3]. Indeed, molecular networks share important architectural features with social networks and the world-wide web [4]. In the present study, we take the view that protein ‘function equals interaction’. Providing interaction information for every gene product is a clean way to assemble the jigsaw puzzle of proteins into a functional map.

The complexity of interaction networks is captured in mathematical terms as graphs $G = (V, E)$. Generally, graphs consist of a set of nodes (or vertices) V linked by either directed or undirected edges E. In interaction networks, nodes represent biological entities such as domains, proteins, complexes or protein families. The edges between these nodes are interactions or functional associations. Each edge can be assigned a weight. For example, the weight could represent the strength of an interaction, such as the dissociation constant K_D or the amount of independent experimental evidence for this interaction.

Like the proteome, the interactome is a dynamic structure. In the first approximation, we will not attempt to model the dynamics of such complex systems, albeit there are attempts based on Boolean networks or differential equations for smaller subsystems [5]. In the present study, we restrict ourselves to model protein–protein interaction networks as static graphs. Although this model glosses over many hairy problems concerning the description of biological function, it provides an abstract overview of cellular networks and the resulting graphs can be subjected to algorithmic and graph-theoretic analysis [6].

Protein classification by interaction networks

A vast number of different experimental and computational methods have been devised to detect protein–protein interactions. The available methods capture different aspects of a whole spectrum ranging from ‘hard’ physical interactions through transient binding (e.g. in signal transduction) to indirect genetic and functional (e.g. metabolic) associations. The common denominator of the most successful experimental techniques is that they measure binding of a ‘bait’ protein to a single or a whole library of ‘prey’ proteins. In contrast, the computational methods measure a degree of association within different ‘functional contexts’, like genes within an

Key words: algorithm, interactome, proteomics, protein interaction network, structure and function, signal-transduction pathway.

Abbreviation used: GBA, guilt-by-association.

[1] **To whom correspondence should be addressed (email liisa.holm@helsinki.fi).**

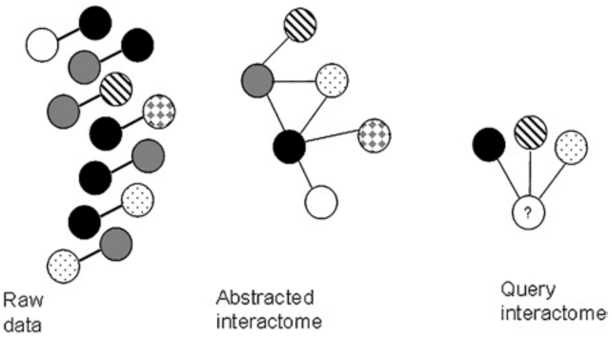
operon, domain fusion within open reading frames or co-occurrences of gene or protein names in Medline abstracts. Efforts are underway to integrate interaction data into publicly available resources [7] (e.g. BIND, DIP, MINT, IntAct and MIPS). Despite the huge differences between all the computational and experimental techniques available, there are nevertheless a number of common emerging properties of interaction networks, namely complexity, incompleteness, noise, a scale-free degree distribution and small-world behaviour.

Experimental results represent typically binary interactions. In other words, each node $v \in V$ is connected to just one other node $w \in V$, representing experimental evidence that the protein represented by v interacts with the protein represented by w. For interaction data derived from the same species, it is possible to assign a finite set of protein names L to the interacting partners represented as the set of nodes V. The labelling function l: $V \rightarrow L$ represents our knowledge about the 'identity' of proteins within the proteome. Given this knowledge, it is straightforward to join the experimental interactions through nodes with identical labels. This method does not work across species, unless we have a way for identifying the 'same' proteins in different species. This identification is usually done by homology inferred from sequence similarity, but any classification from the biological domain can be used. Building a genome-wide interaction is formally the same as any other contraction of a graph (i.e. clustering of the nodes) and differs only in the way the identity function l is defined.

Genome sequencing projects are producing huge numbers of hypothetical proteins of unknown function and structure. The usual way of assigning a putative function or structure to these hypothetical proteins is based on homology between the query protein and another, experimentally characterized protein. Unfortunately, a large proportion of hypothetical proteins cannot be linked by sequence similarity to any known protein family. Non-homology methods have therefore been proposed, based on comparison between the interactions of a query protein with those of previously classified proteins.

GBA (guilt-by-association) implies that a query protein affiliates with the consensus among its neighbours [8]. Imagine colouring the interaction graph according to attributes; GBA works on the assumption that large islands of uniform colour will emerge. This makes sense for predicting attributes such as subcellular localization, since proteins in the same compartment have a good chance of interacting with each other. However, there are biologically relevant entities where interactions involve proteins with very different attributes. For example, the successive steps of a metabolic pathway may involve different enzyme activities. To address prediction problems of this latter type, the actual topology of the overall network is an important piece of information. In contrast with the GBA principle, our EMBED method determines the identity of the query protein by matching the interaction patterns in terms of the 'spectrum' of types of neighbours (Figure 1). The principle of EMBED has been used successfully for structure assignment [9] and function assignment [10].

Figure 1 | Generating interaction networks
Proteins are represented by circles (nodes) and interactions by lines (edges). The raw data gives binary interactions (left). The 'same' nodes are merged. 'Sameness' may be defined by protein names or a classification from the biological domain, leading to abstraction (middle). Unknown query proteins can be classified by matching their interaction patterns to the background information represented by a large interaction network. The gray node is the only one that has the same set of neighbours as the query on the right.



Automatic reconstruction of signalling pathways

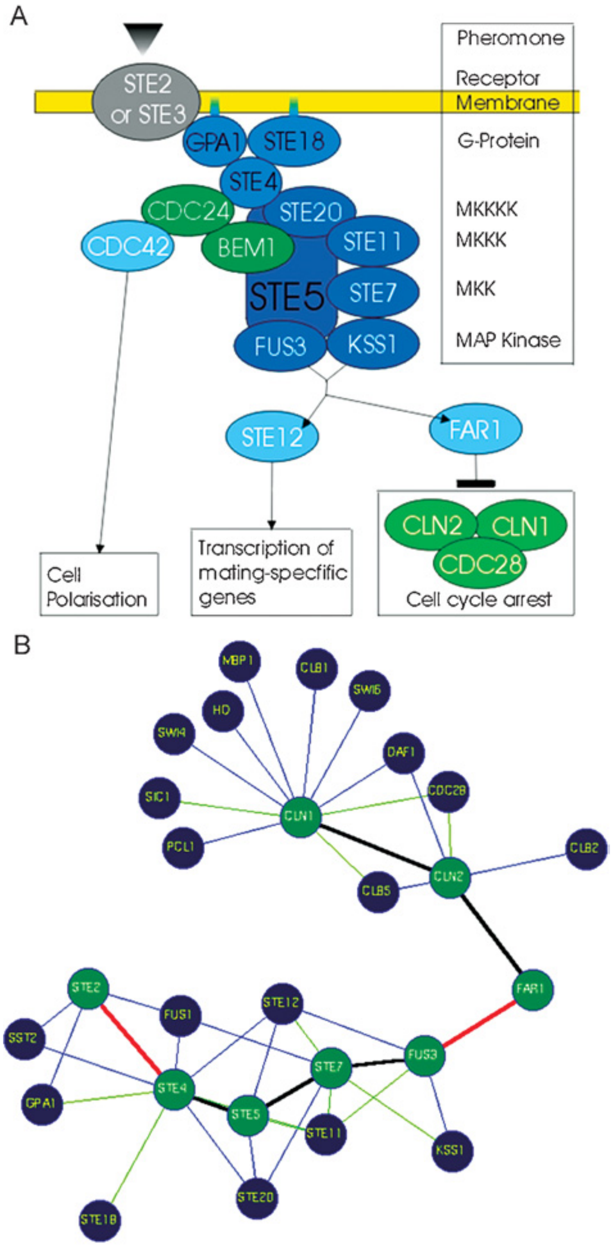
In the following application, an interaction network is generated by tapping into the vast amount of information available from Medline abstracts. We use a statistical approach [11] to compute protein–protein associations from Medline abstracts. Unlike most other experimental data sets, the statistical associations have the advantageous property that they form a weighted graph. In other words, every association comes with a measure representing the significance of the association. This, in turn, is a crude measure reflecting the reproducibility and hence the strength of the underlying physical or genetic interactions.

Densely interlinked clusters of proteins with similar function can be detected in interaction graphs and are called 'functional modules'. Signalling cascades, on the other hand, have a more 'linear' architecture. Signalling cascades are constructed from a variety of different proteins with the purpose of carrying the information conveyed by external stimuli to the nucleus and triggering the appropriate cellular responses. We search our interaction graph for paths that connect the end points of such signalling pathways [6]. The path starts at a given receptor and ends at any of a given set of transcription factors. Owing to ubiquitous weak associations, there is always a short-cut between any two proteins in a small-world network. However, biological pathways rarely coincide with the shortest path. We have demonstrated that a simple algorithm, where the information is routed along the most reliable edges in the network, selects paths that are remarkably similar to known signalling pathways. This algorithm copes with noise and the small-world characteristics of the network without any preprocessing of the data, such as removal of edges below a certain threshold or elimination of highly connected hubs (e.g. [12]). Furthermore, the algorithm is free of any assumption on the length of a pathway, the absolute number of nodes to be incorporated in the pathway or about any intermediates (though known intermediates can be easily incorporated in the search). Since the algorithm implements a greedy strategy, it is robust, fast and delivers reproducible results (unless new trends upset the Medline data set).

Empirically, the algorithm seems to get the topological order of the signalling cascades roughly right (Figure 2). The method presented here is based on precomputed statistical associations from Medline abstracts, so it does not really find new associations or interactions. The scope is hence limited, by definition, to the published results (abstracts) – on well-studied pathways – and does not cover hypothetical proteins. In principle, however, the algorithm would work on any weighted graph, if appropriate weights can be generated [13,14]. We note that path searching algorithms have applications in various problem domains, e.g. sequence alignment [15].

Figure 2 | Reconstruction of signalling pathways

(**A**) The known pheromone signalling pathway [17]. (**B**) Thick lines indicate the 'backbone' linking a cell-surface receptor (Ste2) to a transcription factor (Cln1). The backbone follows the most reliable edges in a yeast interaction network based on statistical associations in Medline abstracts. The thin lines link 'associated factors' to the backbone. These nodes are generally connected to the backbone proteins.



Unravelling unknown interaction networks

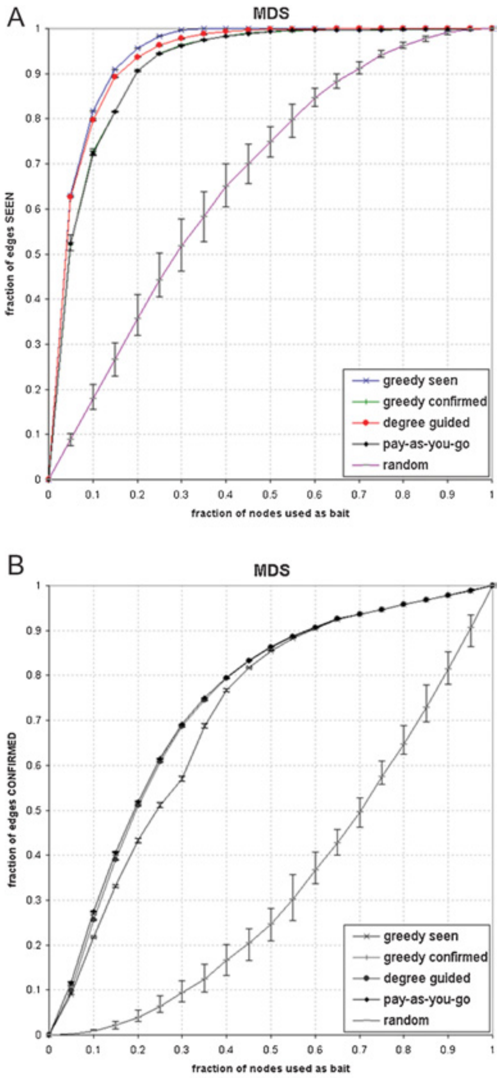
Experimental high throughput techniques (such as yeast-two-hybrid and tandem-affinity purification MS) allow us, with some degree of error, to determine the neighbourhood of a given protein within the interaction network. Currently, we have a complete list of proteins from sequencing of many organisms, but only very limited information is available on the interactome. The vast majority of protein–protein interactions either remain to be experimentally determined or have not been made available in a public database yet. How can we complete the coverage of interaction space with minimal effort in terms of the required number of experiments? To address this question, we modelled the required resources by assuming cost and time to be in a constant proportional relationship to the number of performed pull-down experiments, which is equivalent to the number of proteins used as bait [16]. We are well aware that this simplification leaves out a lot of experimental details, but it leads to a concise model of the overall process of information gain in proteomics.

To simulate the discovery of an unknown interaction network, we use real interaction data sets (for yeast) that are explored from scratch by virtual pull-down experiments. Although no complete data set of interactions is available for a single organism yet, all observations indicate that protein interaction networks are scale-free. Since any randomly selected subset of edges from a scale-free network again follows a power-law distribution, and all interaction data sets available represent different subsets of the overall interactome, we conclude that interaction space as a whole has the same distribution shape as any major subset. That interaction networks of higher organisms, like human, are scale-free as well, seems to be the most reasonable assumption at this point. Thus the simulation results for incomplete interactomes should hold also for the real-life exploration of unknown interaction networks.

Coverage of interaction space translates to edge coverage in graphs. In the present study, edge coverage means to select a subset of nodes such that every edge is connected to, or covered by, at least one node in this subset. There are many different solutions for finding a set of nodes covering all edges (interactions) within the same graph (interaction network).

Figure 3 | Performance of the 'pay-as-you-go' strategy compared with random ordering of baits or a theoretically optimal, degree-guided strategy

The coverage of interaction space is cast as a problem of edge covering in graphs. A virtual pull-down experiment reveals the edges going out from a selected 'bait' node. Edges 'seen' have been covered at least in one direction (**A**). Edges 'confirmed' have been covered in both directions (**B**). An efficient strategy selects 'baits' in such an order that as much of the graph is covered as early as possible.



A subset of highly connected nodes covers a larger portion of the network compared with another subset of the same size consisting of nodes that are less connected. In biological terms, a pull-down experiment reveals the adjacent edges (interacting proteins) of one node (the bait). Consequently, a minimum edge-covering set would allow us to map the interactome with the minimal experimental effort (minimal number of baits). The disadvantage here is that the problem of finding the minimum edge-covering set has been shown to be NP-complete and hence cannot be computed efficiently even on a graph of known topology. In the biological setting, matters are complicated further because the topology of the interactome graph is initially unknown.

The information gain from pull-down experiments is determined by the strategy applied to order the baits. Owing to a scale-free distribution, fast coverage is obtained by initially focusing on the hubs in the network. Unfortunately, locating hubs requires prior global information about the network one is trying to unravel. We have shown that a novel 'pay-as-you-go' strategy finds its way to highly connected nodes near-optimally using only local information that is collected on-the-fly in successive pull-down experiments. Using the 'pay-as-you-go' strategy, 90% of the human interactome can be seen in 10000 pull-down experiments with 50% of the interactions confirmed in both directions. The small-world property ensures a short path between any two nodes and accounts for the quick convergence towards the hubs in the network, independent of the starting point. A scale-free distribution allows our strategy to estimate the number of interactions based on partial information and select the next bait. Remarkably, the pay-as-you-go strategy already achieves near optimal coverage in confirming interactions even in the absence of a reliable measure of interaction degree. Apart from being able to cover the interaction space efficiently without any prior knowledge, the real strength of the method lies in its ability to generate confirmed interaction information close to the greedy confirmed strategy (Figure 3). Given the limitations of present experimental techniques, an interaction has to be repeatedly detected (at least twice) before it can be regarded as safe knowledge.

This work exploits the properties of scale-free networks to tackle otherwise computationally hard problems effectively and by computationally relatively simple means. The 'pay-as-you-go' strategy gives no performance advantage in 'edges seen' in fully randomized networks, but using the right-hand side of the degree distribution in picking baits confers an advantage already in random networks for 'edges confirmed' [6]. An interesting corollary is that if the fitness value of an interaction network depended on 'edges confirmed', it would spontaneously evolve from a random network to a scale-free network, which has the highest computational capacity.

Conclusions

The publication of several large-scale interaction data sets in recent years has been accompanied by the emergence and proliferation of new algorithms for interaction networks. Algorithms exploiting interaction data have to deal with the problems of noise, incompleteness and complexity. Interaction patterns can be used in protein classification to narrow down the scope of hypotheses before experimental verification. Modules or pathways abstracted from interaction networks may help researchers to look beyond the boundaries of a single protein and consider the surrounding functional context. But perhaps the main impact of the work reviewed in this paper lies not so much in any precise prediction of novel biological facts, as in guiding experiments and drawing attention to interesting parallel relationships between mathematical abstractions and the architecture of biological systems.

We thank group members and colleagues for discussion, and especially S. Dietmann, K. Paskiewicz and T. Schlitt for their contributions to data analysis.

References

1 Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998) J. Mol. Biol. **283**, 707–725

2 Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. (2000) Nature (London) **405**, 823–826

3 Chen, Z. and Han, M. (2000) Bioessays **22**, 503–506

4 Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) Nature (London) **407**, 651–654

5 de Jong, H. (2002) J. Comput. Biol. **9**, 67–103

6 Lappe, M. (2003) Ph.D. Thesis, Cambridge University, U.K.

7 Galperin, M.Y. (2005) Nucleic Acids Res. **33**, D5–D24

8 Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) Yeast **18**, 523–531

9 Lappe, M., Park, J., Niggemann, O. and Holm, L. (2001) Bioinformatics **17** (Suppl. 1), S149–S156

10 Vasquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003) Nat. Biotechnol. **21**, 697–700

11 Jenssen, T.K., Laegreid, A., Komorowski, J. and Hovig, E. (2001) Nat. Genet. **28**, 21–28

12 Steffen, M., Petti, A., Aach, J., D'Haeseleer, P. and Church, G. (2002) Bioinformatics **3**, 34

13 von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) Nucleic Acids Res. **33**, D433–D437

14 Lee, I., Date, S.V., Adai, A.T. and Marcotte, E.M. (2004) Science **206**, 1555–1558

15 Heger, A., Lappe, M. and Holm, L. (2004) J. Comp. Biol. **11**, 843–857

16 Lappe, M. and Holm, L. (2004) Nat. Biotechnol. **22**, 98–103

17 Schrick, K., Garvik, B. and Hartwell, L.H. (1997) Genetics **147**, 19–32

Received 19 January 2005