Novel Algorithms for Protein Interaction Networks

Michael Lappe
Wolfson College

A dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.

Email: lappe@ebi.ac.uk

November 26, 2003
To My Family
This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using \LaTeX\ according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.
Novel Algorithms for Protein Interaction Networks

Summary

November 26, 2003

Michael Lappe
Wolfson College

The functional characterisation of all the genes and their gene products is the main challenge of the post-genomic era. Recent experimental and computational techniques have enabled the study of interactions among all proteins on a large-scale. In this thesis I present approaches to exploit interaction information for the inference of protein structure, function, signalling pathways and ultimately entire interactomes.

Interaction information for every gene-product is a clean way to assemble the jigsaw puzzle of proteins into a functional map. In my work interaction networks are modelled as Graphs. The idea of up-casting incomplete interaction information is formulated. The results are abstracted networks which are generic and organism independent while consistently exhibiting scale-free and small-world properties.

These maps prompted me to investigate and develop the concept that protein function can be operationalised via protein interaction patterns. Since the architecture of biological networks differs distinctly from random networks, the functional maps contain a signal that can be used for predictive purposes. The knowledge base for further inference is built on existing biological classifications of protein structure (i.e. SCOP) and function (i.e. GO). I developed a novel algorithmic principle (EMBed) for protein function and structure prediction. This non-homology method does not require any sequence or structure similarity. It enables accurate predictions beyond the limits of current bioinformatics methods which are based on the concept of homology.

Moving on to higher level definitions of protein function, the question of how complex networks can be decomposed into meaningful subsets arises. In order to formulate experimentally verifiable hypotheses the functional maps must be made accessible to human interpretation. I demonstrate how whole signal-transduction pathways can be extracted reliably. Here complex sets of protein associations derived from text-mining the biological literature are used as an example of noisy information.

Finally, I present a model of the overall information gain in the field of proteomics. I formulate an algorithmic strategy that enables the proteomics community to build a reliable scaffold of the interactome in a fraction of the time compared to un-coordinated efforts.
This work opens the door to a new breed of methods and strategies which actively use the scale-free properties of biological networks to their advantage. As a result, these methods are able to approximate $NP$-hard problems in bioinformatics effectively by relatively simple computational means.
Acknowledgements

"He called me in and asked me what I did, exactly.
Have you ever heard of such a thing?
What sort of question is that?
This is a university!"
Terry Pratchett, 'The Last Continent'

The work contained in this thesis was conducted at the European Bioinformatics Institute (EBI) in Cambridge, UK, between September 1999 and August 2003. I was the recipient of an EMBL (European Molecular Biology Laboratory) predoctoral fellowship to work with Dr. Liisa Holm in the Structural and Functional Genomics Group at EBI, an EMBL outstation.

I am indebted to countless people I had the pleasure to meet at EMBL Heidelberg, the EBI, cellzone, the MRC-LMB and throughout the University of Cambridge who helped me in this exciting time with their encouragement, moral and technical support, critical feedback and scientific discussions.

Foremost I have to thank my supervisor Dr. Liisa Holm for her patience, in-depth knowledge, wisdom and open-mindedness. She gave me plenty of opportunity to develop and pursue novel ideas and the critical feedback to put these ideas to effective use. Together with the other members of the Structural & Functional Genomics Group at EBI, namely Sabine Dietmann, Andreas Heger, Matthieu Louis and Jong Park, she provided an environment in which I could share and develop my passion for science.

Also I am grateful to all the members of my Thesis Advisory Committee. Dr. Michael Ashburner, Dr. Dennis Bray and Dr. Christos Ouzonis provided invaluable feedback and advice.

Together with Caleb Webber, who also gave me lots of food for thought and critical feedback on this thesis, I served as pre-doc representative. Among the lively crowd of Ph.D. students who accompanied me at the EBI, I have to thank mostly Anton Enright, Katia Kivinen, Victor Kunin and Thomas Schlitt for the interesting time.
Furthermore, Henning Hermjacob, Rolf Appweiler and the whole IntAct Team provided me with resources in the context of the IntAct project and opportunities to participate and present my work at the Proteome Standards Initiative (PSI) meetings. The students Markus Brosch, Gregory Gimenez and Konrad Pasquiewitz helped me to further and implement the concepts developed in this thesis. Part of this work will be continued at the Cambridge University Biochemistry department and the MRC-LMB, thanks to Florian Hollfelder, Sir Tom Blundell, Andrew Griffiths and the BBSRC.

Also Dan Bolser, Sarah Teichmann, Oliver Niggemann, Clemens Grabher and Yiorgios Skiniotis helped in numerous aspects of this work. I also have to thank M. Bolotin-Fukuhara, A. Goffeau, L. Schmitt and O. Stasik for comments on the pathway extraction system. Furthermore, I am indebted to G. Cesareni, F.C.P. Holstege, P. Kemmeren, I. Lappalainen, L. Montecchi-Palazzi, S. Moeller, G. Neubauer, J. Rick, J. Schulz, A. Valencia, H. Varmark, J. Vilo, M. Vingron and I. Xenarios for helpful comments.

But most of all, I am grateful to my wife Namhee Kwon-Lappe, my parents Gabriele and Norbert and my sister Monika for their love and enduring support over all these interesting years.
List of Tables
List of Figures
Chapter 1

Introduction

“It is embarrassing to know that one is a god of a world that only exists because every improbability curve must have its’ far end;”
Terry Pratchett, ’The Colour of Magic’

1.1 Summary

In the post-genomic era protein function is defined in the context of complex networks of specific interactions. Here I give an overview of computational definitions of protein function and link them to interaction networks. The central guiding hypothesis that protein function equals interaction is formulated.

The main experimental and computational sources of interaction information are briefly introduced and discussed. Finally, the emergent properties of the resulting interaction networks are highlighted: a varying degree of noise and incompleteness combined with small-world behavior and a non-random, scale-free degree distribution.
1.2 Motivation

The main challenge in the post-genomic-era is the accurate annotation of the gene products discovered in the many on-going sequencing projects. A major and almost constant fraction of about 30%-50% of all newly discovered genes have no sequence similarity to any other gene of known function. This limits the scope of widely used homology-based methods such as BLAST (Altschul et al., 1997).

In this context structural genomics initiatives set out to provide experimentally determined structures such that for most sequences a reliable homology model can be obtained. Despite all advances in the field, the experimental determination of protein structures remain time-consuming and expensive.

It is obvious that novel computational and experimental methods for function prediction and annotation are needed. Fully automated computational methods for function prediction suffer from the fact that the broad term protein function remains relatively ill-defined. Clearly the field of molecular biology in general and bioinformatics in particular will benefit from a clean and computationally graspable definition of function.

This is the root of the wide-spread interest in protein interaction information. As novel experimental and computational methods are developed to search for interacting partners of a protein of interest, our view of protein function is transformed. Now the cell is seen as a complex society of proteins rather than an un-coordinated bag of proteins where each member works pretty much in isolation.

Because protein-protein interactions play a major role in all biological processes the post-genomic view of protein-function defines every protein as a part of a complex network of specific interactions. The main challenges and questions I hope to address in this thesis are:

- How are interaction networks constructed from the available data and what are the architectural features of these networks?
- How can protein function be reliably computed via interactions?
- In what ways can the complexity of cellular networks be tackled?
- Is it feasible to provide interaction information efficiently for an entire proteome?
1.3 Concepts

Here I will discuss a few different approaches to protein function and relate these concepts to protein interactions.

One of the oldest paradigms in molecular biology is the concept of one gene - one protein - one function. This encapsulates the assumption that every gene codes for one protein where its’ unique sequence codes for a specific native structure. As an enzyme, this protein structure is responsible for catalysis of specific reactions in the cell. In this context the definition of function is the biochemical function of a protein. This notion has been captured in a hierarchical classification by the Enzyme Commission, the EC-number.

It has become clear that the above paradigm is far from the truth. Many genes have different splice-variants coding for different proteins, notwithstanding different post-translational modifications. Some proteins have several binding sites and catalyze a variety of different reactions. Furthermore, the genetic or developmental view of protein function is not accounted for in this model at all. Of course 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.

Usually protein sequences are composed of domains. Domains can be defined as independent folding units or independent evolutionary units of a protein. Then explicit functional annotation can be attached to defined domains. Such domain definitions can be found, for example, in the member databases of InterPro, such as PfAM (Bateman et al., 2004) or others. The structural definitions of domains are found in databases such as SCOP (Murzin et al., 1995) or CATH.

When the associated text containing the biological annotation is just copied from proteins of known function to new sequences based on spurious sequence similarity, this can result in functionally different proteins getting similar annotations. As proteins may contain several domains, the 'domain-chaining-effect' can lead to false annotations in subsequent rounds of automated function prediction. Clearly a more structured approach is needed.
1.3.1 Operationalizing the Notion of Function

Protein structures already have been organized into hierarchical classifications such as SCOP and CATH, and the FSSP (http://www.ebi.ac.uk/dau/fssp/) which agree well for most of the protein folds. These classifications also deliver functional information, as certain structural classes are involved in biological processes like, for example, DNA-binding or ATP-hydrolysis.

A similar and more structured approach to functional annotation is taken by the Gene Ontology Consortium (GO). Here functional terms are described within an Ontology. As the information is structured, it makes the annotation of protein function described accessible to algorithms. The entries are organized as a directed acyclic graph (DAG), where the direction is from the most general (GO root) to the more specific terms. In GO the distinction between different views of function is accounted for. There are 3 main categories in GO, describing biochemical function, cellular role and subcellular localization. A protein is annotated with at least one label from each of the three categories.

1.3.2 Function equals Interaction

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 by Indeed molecular networks share important architectural features with social networks and the world-wide-web.

Here I will take the view that protein function equals interaction. More specifically, in the context of the broad GO categories of functions, biochemical function can be defined as protein - small molecule interactions. Similarly, subcellular localization can be seen as protein - organelle interactions. Finally, biological processes can be captured via protein - protein interactions. In the following, I will focus on protein - protein interaction networks and refer to biological process as protein function, unless stated otherwise.
1.3.3 Modeling Interaction Networks as Graphs

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 arcs 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 given a weight. For example, the weight can represent the strength of an interaction, such as the dissociation constant $K_D$ or the number of independent experimental evidence for this interaction.

1.3.4 The Complexity of the Interactome

The interactome can be defined as the set of all specific protein-protein interactions occurring in an organism. Like the proteome, the interactome is a dynamic structure.

The size of the potential interaction space of an organism is the number of all possible pairwise combinations of proteins. This can be visualized as an adjacency matrix (see Figure). In such an adjacency matrix $M$ an entry $m_{ij}$ is marked if the proteins $i$ and $j$ are known to interact. The size of this matrix, including self-interactions, is $\frac{n(n+1)}{2}$, where $n$ is the number of proteins. Obviously the potential complexity of a network grows at a quadratic rate to the number of proteins involved. This suggests that the complexity of organisms may not merely be related to the number of genes in a genome, but to the number of ways the resulting gene-products can interact.

In a first approximation I 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. Here I will restrict myself to model interaction networks as static graphs to gain an overview of cellular networks. The resulting graphs will then be subjected to algorithmic and graph-theoretic analysis.
1.4 Methods

The main experimental and computational methods of interaction information will be introduced briefly. I will focus on the kind of interaction information that can be retrieved from the different methods and how the information is modeled as networks. Besides the techniques discussed here, many other methods are available. Because it seems impossible to give an exhaustive overview, I restrict myself here to discuss the methods and datasets used within the scope of this thesis.

1.4.1 Yeast-Two-Hybrid (Y2H) Data

The first experimental method to yield high-throughput, large-scale datasets was introduced by S.Fields and co-workers (1996) based on the modularity of the GAL4 transcription factor. In essence, the GAL4 protein is split into its' DNA binding domain (GAL4-db) and the activation domain (GAL4-ad) (Figure 14.1). In order to test for an interaction between proteins X and Y, a hybrid is formed between X and the GAL4-db on the one hand and Y and the GAL4-ad on the other hand. If X and Y do interact, the resulting complex reconstitutes transcription of a reporter gene. If no interaction between X and Y occurs, the activation domain can not recruit the transcription machinery to the reporter gene (Figure 14.2). This experiment is usually conducted in *S. cerevisiae* (yeast) and because of the two hybrid constructs involved was termed *yeast-two-hybrid* experiments, or *Y2H* for short.

The technique has been employed numerous times to determine interactions for various proteins. The most prominent screens were performed by Ito and co-workers and Uetz and co-workers (2000, 2001), resulting in large interaction datasets for yeast. Another dataset for *Helicobacter pylori* was produced by Rain and co-workers (2005).

The yeast-two-hybrid experiments return pairwise interaction information and hence can be modeled as an edge joining the two interacting proteins. The detection of an interaction in this method is carried out in the nucleus of the cell, which is not the 'natural' environment of all the tested proteins. Hence the method may not be able to detect some interactions due to misfolding of the hybrid constructs. Also, the reporter-gene may be transcribed even in the absence of an interaction due to auto-activation (Aloy and Russell,
1.4. METHODS

Because of these and other complications, an estimated 50% of all reported interactions are assumed to be false positives with a false-negative rate of about 90%.

1.4.2 Tandem Affinity Purification (TAP-MS)

In this experimental setup the protein of interest, called bait, is fused with a tag. This tag immobilizes the bait on an affinity column or micro-beads. Then proteins are purified from solution as they bind to the bait - a process that has been termed pull-down-experiment. The resulting complexes are cleaved and the members separated on a gel. In the end, the members of the purified complex are identified using micro-sequencing on mass-spectrometers. Because this technique relies on the tandem-affinity purification with subsequent mass-spectrometry it has been termed TAP-MS.

The results of two large-scale screens have been published by Gavin and co-workers and Ho and co-workers. The TAP-MS data is more reliable and reproducible, false positive and false negative rates are estimated below 30%. As any interaction has to 'survive' the purification step the resulting data represents relatively stable protein complexes, while more transient interactions may not be detected.
1.4. METHODS

Figure 1.2: Illustration of Tandem-Affinity Purification (TAP) with subsequent mass-spectrometric analysis (MS), (adapted from Kumar and Snyder, 2005)

Ho and co-workers over-expressed the bait proteins. This procedure may have contributed to a higher rate of false positive interactions in the data. This study was carried out at MDS proteomics, and will be referred to as the MDS dataset. Gavin and co-workers mainly targeted all yeast genes which have a human ortholog. The resulting network is, albeit more reliable, less dense and less interconnected than the MDS dataset. As the work was mainly carried out at cellzome AG the resulting network will be denoted as CZ.

The results from TAP-MS experiments can not directly be modeled as binary interactions. The exact pairwise relationships between all the members remains unclear. The spoke and clique model (see Figure 1.2) represent the two extreme assumptions, and the truth is certainly in between the two. Bader and co-workers have shown that the spoke model is about 3 times more accurate than the clique model.

1.4.3 Structural Data (PDB)

The resolved protein structures deposited in the Protein Databank (PDB) can be harnessed to detect domain-domain interactions. An interaction between two domains is defined here as a minimum of 5 pairs of residues which are in contact. A contact is defined as a pair of residues from the two respective domains which are located within a distance of less
Figure 1.3: Modeling Tandem-Affinity Purification (TAP-MS) data as binary interactions. The interaction partners of a bait-protein using Tandem-Affinity Purification and subsequent Mass-Spectrometry (TAP-MS) are called prey. In order to break the resulting set down into pairwise interactions, there are two approaches, called the spoke and clique model. The spoke model (left) models only edges from the bait to all prey proteins. The clique model (right) covers all potential interactions by inferring pairwise interactions among all proteins identified in the experiment. In this sense no distinction between bait and prey is made. Obviously, the clique model represents an over-prediction, as it is a superset of all possible pairwise interactions. The spoke model is about 3 times more accurate in reflecting the true binary relationships.
1.4. METHODS

Figure 1.4: Domain Adjacency in PDB-Structures. The two domains of the PDB structure 1URA are shown as $C_\alpha$ traces in green (left) and red (right). All the atoms in contact within a distance of 5Å from the domain interface are drawn as spheres. The picture was taken from http://interaction.mrc-dunn.cam.ac.uk/local/Httpd/PDBHighlight/, courtesy of D. Bolser.

The 5Å threshold is quite conservative, resulting in accurate data that can readily be modeled as binary interactions. However, false positives may arise from crystal packing artefacts. The resulting dataset is referred to as PDB. The dataset can be sub-divided into intra- and intermolecular interactions, depending on whether the domains are located on the same or different polypeptide chains.
1.4. METHODS

Figure 1.5: Domain Adjacency in Genomic Sequences. In this illustration, 3 domains (A, B, C) could be assigned to the open-reading frame (ORF), shown in blue. The length (number of residues) of the linker-regions in between is shown on top. Given a threshold of 30 residues, in this example an interaction between domain A and B is inferred.

1.4.4 Domain Adjacency

Other approaches to predict domain interactions are based on domain fusion events and genomic context. Here SCOP domain assignments within a yeast protein sequence that were less than 30 residues apart were interpreted and modelled as binary intra-molecular domain interactions (see Figure 1.5).

Obviously, other domain definitions could be used in the same way, i.e. PFAM or SMART. Similar to the clique model in TAP-MS, interactions among all domains within the same open-reading frame (ORF) could be inferred. Using only links between directly adjacent domains connected by a linker region of less than 30 residues is a relatively conservative approach. The resulting data are denoted as ORF.

1.4.5 Text Analysis

In order to tap into the vast amount of biological literature and structure the information contained in scientific articles, several approaches have been developed. One of the most prominent may be natural language processing (NLP) where the algorithm attempt to ‘read’ and interpret texts based on a set of grammatical rules.
A computationally less demanding statistical approach is to count co-occurrences (similar to Table ). Here associations between proteins are derived from their co-occurrence in Medline-abstracts (see Section ). This approach is based on the fact that negative results are published less often than positive results. Hence proteins mentioned together frequently can be assumed to be functionally associated. From the raw co-occurrence counts further statistical measures of association, like mutual-information content, can be calculated (see Section ). Although this data does not constitute a physical interaction in itself, many of the associations found this way correspond to experimentally determined protein-protein interactions. In summary, this represents an inexpensive approach that combines high coverage with reasonable accuracy. Estimates from the overlap with TAP-MS data show that the false-positive and false negative rates are comparable to the results of yeast-two-hybrid experiments.

1.5 Results

Recently, high-throughput screens of protein-protein interactions have taken center stage in the functional characterization of novel proteins. A wealth of interaction data is available from manual annotation and high-throughput screens (for example, Deane et al., 2002) which means that their distribution of interactions often follows a power law (see Figure ). In simple words, most proteins (nodes) have only a few interactions while there are a few proteins which have many interactions (are highly connected). These highly connected proteins keep the cellular network together, and the vast majority of the proteome is contained in one giant connected component. This distribution is very different from what would be expected of a random graph (see Figure ).
Figure 1.6: A histogram of the degree-distribution of four experimental networks (CZ, CORE, DIP, MDS).
Here the probability $P(k)$ of a node (protein) having $k$ edges (interactions) is plotted over the degree ($k$, number of interactions). While most proteins are only involved in few interactions, a few proteins (hubs) are involved in many interactions. The same data is plotted on an absolute scale above and on a logarithmic scale below. In all data-sets, despite their differences in experimental techniques and composition, this distribution follows a power law and the data points approximate a line in a log/log plot (below). The 'broad tail' of this distribution stretches further than shown here (up to $k=100$), the data-sets contain some 'extreme' cases of hubs with over 200 interactions reported for some proteins.
1.6. DISCUSSION AND OVERVIEW

All available networks also exhibit small-world behavior such that the average path length between any two nodes is very small. The average path length is proportional to the logarithm of the number of nodes in the network which means that there is on average a path of less than 4 interactions between any two proteins in yeast. Indeed I observed that about 90% of all proteins can be reached within a path of length 4 or less in the MDS dataset alone (Fig. 5).

All of the available experimental and computational datasets on protein-protein interaction have different degrees of accuracy and coverage, due to their respective systematic advantages and limitations. This renders the integration of the data into a reliable set a non-trivial task (Deane et al., 2002). The investigation of the noise levels and the estimate of accuracies for different datasets or even individual interactions remain important questions. Here, I note that robust algorithms are needed to cope with the problem of noise that persists through all available data to a varying degree.

A puzzling result when analyzing the overlap of the large-scale datasets is that there is little overlap between them (Bader and Hogue, 2003). This is partly due to the inherent noise, but also indicates that our current picture of the interactome is far from complete.

1.6 Discussion and Overview

A vast number of different experimental and computational methods to detect protein-protein interactions now at our disposal. The available methods capture different aspects of a whole spectrum ranging from “hard” physical interactions via transient binding (i.e. in signal transduction) to indirect genetic and functional (i.e 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 operon, domain fusion within open reading frames or co-citations within medline abstracts. Currently there are efforts underway to integrate interaction data, similar to sequence and structure data, into publicly available resources.
Figure 1.7: A histogram of the degree-distribution for the MDS dataset under increasing levels of randomization. Again, the probability $P(k)$ of a node (protein) having $k$ edges (interactions) is plotted over the degree ($k$, number of interactions) (see Figure 1.6). Each curve results from a copy of the MDS data where an ever-increasing fraction ranging from 10%-100% of the interactions are randomized. The resulting curves are shown on a logarithmic scale below, where the original distribution approximates a line.

Compared to the original distribution, the randomized sets have increasingly less nodes with an extremely low or high number of interactions. In a completely randomized set, the distribution is bell-shaped and peaks around the average degree or density (=number of edges/number of nodes) of the network. Obviously, scale-free networks are far from random.
Figure 1.8: Interaction networks are small-world networks. Here the average percentage of nodes that are reached within a distance of \( d \) steps from a randomly chosen start-node is plotted. In all networks over 50% of the network can be reached within less than six steps from any arbitrarily chosen node.
Examples of this are BIND (Bader et al., 2000) and the DIP (Xenarios et al., 2000) and the MIPS collection of protein complexes.

The first task at hand is the integration of different sources of experimental and computational data in order to generate comprehensive datasets for further studies. In Chapter 6, we model interaction networks as graphs in the context of a structural classification, resulting in an organism-independent overview of cellular networks. The concepts of up-casting, down-casting and level of abstraction are introduced. As interaction pattern seem to be evolutionarily conserved, this corroborates the hypothesis that protein function can be captured, at least to some degree, via protein interactions.

Having comprehensive interaction networks available, biological questions can be restated and solved as graph-theoretic problems (Boulton et al., 2004). The main question is how interaction data can be used in a similar way to sequence data in order to predict protein structure or function. With the EMBed-principle I introduce a general framework for accurate inference in interaction networks in Chapter 8.

Higher-level biological function is described in terms of pathways. By modeling the transmission of information through a noisy network environment I attempt an algorithmic definition of signaling pathways. Chapter 9 demonstrates how pathways can be reconstructed from complex datasets.

Finally, given the scope and limitations of current experimental techniques, the question arises how a complete picture of the interactome is obtained in the most efficient way. I present an algorithm that suggests bait proteins with a near-optimal information return on experimental effort in Chapter 10. The pay-as-you-go strategy allows to cover over 80% of unknown interaction networks with less than 20% of the proteome used as bait in successive pull-down experiments.

Hopefully, this work will, in the context the availability of more comprehensive interaction information, contribute to a more accurate characterization of the human genome in terms of protein structure and function in the near future.

1.7 Acknowledgments

I am grateful to D. Bolser for providing Figure 1. G. Gimenez assisted with Figure 2.
Chapter 2

Generating Protein-Protein Interaction Networks

“‘I was confused and uncertain about all the little details of life. But now,’
he brightened up, ‘while I’m still confused and uncertain it’s on a much
higher plane, d’you see, and at least I know I’m bewildered about the really
fundamental and important facts of the universe.’”
Terry Pratchett, 'EQUAL RITES'

2.1 Summary

I present a framework to generate comprehensive overviews of
protein-protein interactions. In the post-genomic view of cellular
function, each biological entity is seen in the context of a complex
network of interactions. Accordingly, I model functional space by
representing protein-protein-interaction data as undirected graphs.
I suggest a general approach called ’up-casting’ to generate interac-
tion maps of cellular networks in the presence of huge amounts
of fragmented and incomplete data, and to derive representations
of large networks which hide clutter while keeping the essential ar-chitecture of the interaction space. This is achieved by contracting
the graphs according to domain-specific hierarchical classifications.
The key concept here is the notion of induced interaction, which
allows the integration, comparison and analysis of interaction data
from different sources and different organisms at a given level of
abstraction. I apply this approach to compute the intersection and
union of different datasets of protein and domain interactions.

The architecture of the resulting networks remains scale-free, as frequently seen in biological networks, and this property persists through many levels of abstraction. Connections in the network can be projected downwards (down-casting) from higher levels of abstraction down to the level of individual proteins to allow prediction of potential interaction partners.
2.2 Introduction

As more experimental data on protein interactions become available, it will become critically important to integrate and compare the data derived from different sources. The analysis of interaction data aims to reveal the organizational principles of cellular networks and to describe the architecture of biochemical and genetic networks. A key difficulty on the way is the incomplete experimental characterization of most biological systems. In order to fill the information gaps, methods to generalize from individual experimental evidence (e.g. Protein-Protein Interactions) to higher level biological entities (e.g. protein families) are needed to generate structural and functional annotations.

Even for such a simple model organism as yeast, with 6000 genes, the current interaction network is far too complex and densely connected to be perceived as a whole. Therefore, it is important to have means to condense the interaction data into "functional modules" in a comprehensive way, especially for the upcoming amounts of experimental data from yeast-two-hybrid, gene expression, co-immunoprecipitation, tandem-affinity purification (TAP) and protein array experiments etc. It is intuitively clear that, as in aerial archaeology, where the structure of an ancient settlement is invisible from the ground and only becomes apparent from aerial photographs, it is necessary to take a step back to get an idea of the bigger picture. Here I present a general framework that is able to integrate data coming from different sources, at different levels of abstraction, from different species, and is able to condense the amount of data in a meaningful way. The result is a glimpse at the overall architecture and topology of cellular networks.

2.2.1 Upcasting: Clustering Interaction Information

Here I apply a graph-theoretic framework to generate protein interaction maps from a given set of binary interactions obtained from experimental data. Such a set can be seen as a graph $G = (V, E)$. Initially, each node $v \in V$ is connected to just one other node $w$, representing experimental evidence that the protein represented by $v$ interacts with the protein represented by $w$. So how do we go about joining these binary interactions in order to get an overview?
2.2. INTRODUCTION

It is fairly obvious that 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 $V$. This defines a labeling (or identity) function $l : V \rightarrow L$, where $l$ represents our knowledge about the "identity" of proteins within the proteome. Then it is straightforward to join the experimental interactions via nodes with identical labels. This method does not work across species, unless there is a way for identifying the "same" proteins in different species by homology. Biologists distinguish two kinds of homology: orthologues, which arise from speciation, and paralogues, which arise from gene duplication. For practical purposes, orthologues are defined as having exactly the same biological function (including interactions) while paralogues may have evolved new properties or lost some ancestral properties. One method would be to link the interaction networks of different species by orthologous genes (COGs), which would be an extension of the above mentioned identity function by a lookup-table defining the "same" node across different species. A more general way is via sequence similarity, using the assumption that proteins above a certain sequence similarity undergo similar interactions. In this case, the labeling function would be based on a threshold $t$ for sequence similarity. Unfortunately such a general threshold has not yet been determined.

I am going about these obvious notions in this rather awkward way to demonstrate that building a genome-wide interaction map is formally the same as any other contraction of a graph (clustering of the nodes) and differs only in the way the identity function $l$ is defined.

There are two general classes of clustering approaches, which can be used to condense the interaction graph. The first approach is un-supervised clustering, which means treating the network as an arbitrary graph and clustering it according to its distribution of edges, without any domain-specific knowledge. Decomposition of the graph into highly connected subcomponents reveals communication hubs and functional modules (sets of genes that are involved in common cellular processes). The second approach is supervised clustering, whereby one assigns a set of known biological properties to all the nodes and clusters the nodes based on these attributes. This is exemplified by the clustering according to subcellular localization by Schwikowski and co-workers to see how much crosstalk there is between different cellular compartments, i.e. edges that link the given clusters.
Figure 2.1: Upcasting in SCOP and Level of Abstraction. The SCOP hierarchy is depicted from the root (top) via class, fold, superfamily family down to domain level. An interaction set between different domains is shown below. A mapping of the interaction into the hierarchy induces higher level interactions. In this example, the interaction highlighted on the upper left induces an interaction between superfamilies 1.1.1 and 1.1.3.
Figure 2.2: Contraction of a Network at Different Levels of Abstraction. The set of binary interactions is shown below and the same as in Figure 2.1. The interactions are linked to SCOP by mapping the nodes of the interactions to the domains hierarchy. The resulting up-casted networks on superfamily, family and domain level (top-down) of the interaction set are shown.
2.2. INTRODUCTION

2.2.2 Contraction of a Graph based on a given Classification

Here I use a given clustering from the biological domain in order to analyze the underlying interaction network in the context of the knowledge encoded in this clustering. Since the set of binary interactions is now seen as interactions within and between higher-level entities (like subcellular compartment or protein-superfamily) this leads us to the definition of induced interactions at a higher level of abstraction.

In more biological terms, an interaction $e = (v \leftrightarrow w) \in E$ between the two proteins $v$ and $w$ induces an interaction between two higher level biological entities (clusters, or inner nodes in a hierarchical classification) $A$ and $B$, if and only if $v$ and $w$ are descendants (or members of the clusters) of $A$ and $B$, respectively. Then $i = (A \leftrightarrow B)$ is called an induced interaction of $e$ given the classification (or clustering) $C$. More verbally, an induced interaction at a given level of abstraction is the interaction between higher-level biological entities (e.g. subcellular compartments, protein families) induced by a clustering of lower-level entities (e.g proteins or domains) given their interactions. The induced network is equivalent to the connectivity structure resulting from the contraction of the original graph under the given classification (see Section 1.4 for details).

As discussed above the clustering of the interaction network is not limited to a single step. Since different clustering methods can be applied consecutively, this is logically equivalent to a hierarchical clustering. Considering a hierarchical clustering leads to the concept of "level of abstraction": The more the interaction network is contracted, the more abstract from the underlying experimental data it becomes. Higher abstraction trades a gain in generality for a loss of specificity (see Figure 2.3). In analogy to moving upwards to more general classes in object-oriented hierarchies, I will call moving towards the root of a hierarchical classification to a higher level of abstraction upcasting. Examples of upcasting are the abstraction of protein-protein interaction networks to interactions between cellular compartments or structural classes (depending just on the hierarchy used for classification).

In this experiment, I used the SCOP classification for the contraction of the interaction data. SCOP classifies protein domains into a hierarchy with different levels of similarity, which are (from the most general to the most specific) : Class, Fold, Superfamily, Family and Domain (see Figure 2.4).
2.2. INTRODUCTION

Figure 2.3: Downcasting higher-level interactions and the resulting ambiguity. An interaction between two-higher-level entities (i.e. superfamilies) is downcasted to interactions among all their respective members.

2.2.3 Downcasting: Transferring Interaction Information

The complementary procedure of projecting a given network between higher level entities back down to lower levels of abstraction is what I call downcasting. This could be, for example, downcasting a network of protein family interactions back to the protein or domain level. Generalizing from individual protein interactions by upcasting them to protein family level and then downcasting them back to the protein level is a way to predict possible interactions. At the same time it is not reasonable to predict that all members of protein-families A and B interact based on the instance of a single observed interaction, additional filters are needed in the downcasting procedure (see Figure 2.3). The development of specific filters and descriptors for protein interactions is required for accurate genome-wide predictions of protein interaction networks.

Here I demonstrate the application of the framework described above. Upcasting of different sources of interaction data to SCOP superfamily level resulted in a comprehensive overviews of the interactome within the context of a structural classification (Figure 2.4). Down-casted and induced interaction networks are used as background information for a novel algorithm for fold prediction (see Chapter 3).
2.3 Methods

2.3.1 Datasets

For interaction data, the following datasets were used (see Section for details):

- Y2H contained 2238 interactions as the result of comprehensive yeast-two-hybrid experiments.
- DIP (Database of Interacting Proteins) at this point contained 5002 interactions from different sources and organisms, including yeast.
- PDB The PDB was parsed to check for domain contacts. All domains pairs with more than 5 contacts within 5Å were used.
- ORF SCOP domain assignments within a yeast protein sequence that were less than 30 residues apart were interpreted as intra-molecular domain interactions.

As a hierarchical clustering these interactions were mapped to the Structural Classification Of Proteins, SCOP.

2.3.2 Clustering and Induced Interactions

Let $G = (V, E)$ be the graph abstraction of the biological system under consideration (the interaction network). Then $C(G) = (C_1..C_n)$ (a clustering, here the SCOP hierarchy) is a decomposition of $G$ into $n$ subgraphs induced on the $C_i$, if $\cup_{C_i \in C} = V$ and $C_i \cap C_{j \neq i} = \emptyset$. The induced subgraphs $G(C_i)$ are called clusters. The set of edges $E_C \subset E$ consists of the set of edges between the clusters. Then the contraction $H =< C(G), E_C >$ is called the connectivity structure. For a given (interaction) graph $G = (V, E)$ and a given (hierarchical) clustering $C$, let’s consider an edge $e = (v \leftrightarrow w) \in E$, where $v$ and $w$ are nodes from the cluster $C_i$ and $C_{j \neq i}$, respectively. Then $e$ induces an edge in the contraction $H$ between the two nodes representing the clusters $C_i$ and $C_j$, respectively.
2.3. METHODS

![Diagram showing protein sequence with SCOP assignments](image)

Figure 2.4: Removing the Overlap from SCOP Assignments. The protein query sequence is depicted by the black line on top. All SCOP assignments (here SCOP1-SCOP4) produced by BLAST with an e-value better than \( t = 10^{-5} \) were considered in ascending order (top-down). If an assignment was overlapping with a previous domain by more than 40 residues it is discarded (i.e. SCOP2). In this example, SCOP1, SCOP3 and SCOP4 represent the final set of structural domains assigned to the query sequence.

### 2.3.3 Mapping Interaction Data to SCOP

All protein sequences were used as queries in a BLAST search against the SCOP database. First I validated that \( 10^{-5} \) is a reasonable cut-off for the assignment by comparing the entries in SCOP against themselves (data not shown). Assignments below this threshold were disregarded. The yeast interaction data from DIP and Y2H was linked to SCOP by assigning them to the domains in SCOP which resulted in the best e-values. Furthermore, the overlapping SCOP assignments were removed using the following greedy strategy: First, SCOP assignments considered in ascending order from lowest to highest e-value. If the current assignment at hand overlapped with a previous, better assignment by more than 40 residues it was discarded (see Figure).

### 2.3.4 Inducing SCOP Interactions

The step to reduce the protein-protein interaction data to interactions among SCOP domains is fraught with a certain degree of ambiguity. A protein-protein interaction between two mono-domain proteins is unproblematic. In this case when both proteins have just one structural domain assigned with a significant e-value each, and both sequences are nearly covered by their
Figure 2.5: The Ambiguity in inducing SCOP domain-interactions from an experimentally determined interaction between two multi-domain proteins.

respective SCOP assignments, there is no doubt which domain-domain interaction does account for the protein interaction observed.

However, considering a protein-protein interaction between two multi-domain proteins, it remains unclear which domains physically interact in vivo. Figure illustrates this point. Let A and B be two interacting multi-domain proteins with \( a_1, a_2 \) and \( b_1, b_2 \) the domains in A and B respectively. Assume each of the four induced interaction in \( I = \{ (a_1 \leftrightarrow b_1), (a_1 \leftrightarrow b_2), (a_2 \leftrightarrow b_1), (a_2 \leftrightarrow b_2) \} \) is weighted with a standard weight of 1. Now the sum of the weight all induced interactions resulting from one protein-protein interaction equals 4. Hence multi-domain protein interactions would contribute more 'noise' to the induced interaction network than mono-domain protein interactions due to the inherent ambiguity.

This could in principle be accounted for by a reduction of the weight of the induced interactions. Let A and B be interacting multi-domain proteins with \( n \) and \( m \) domains each. Then the induced interactions would be weighted with \( 1/(m \times n) \) such that the sum off all induced interactions is back to 1.

The fact remains that not all sequences of interacting proteins are fully covered with SCOP assignments. This results in 'unknown' regions of proteins where in principle an interaction could take place. In other words, the
exact number of domains within any given proteins is unknown. As \( m \) and \( n \) cannot be determined reliably the approach described above fails.

To resolve this issue, I used the relative sequence length to weight the induced interactions among domains. Let \( l(A) \), \( l(B) \) be the length of the interacting proteins \( A \) and \( B \), while \( l(a_i) \) and \( l(a_j) \) denotes the length of the assigned domains \( i \) and \( j \) in protein \( A \) and \( B \). The relative sequence length of a domain is the fraction of the protein covered by this domain. Now the weight of the induced interaction \( i = (a_i \leftrightarrow b_j) \) can be calculated as the product of the relative sequence length of the two assigned domains:

\[
weight(a_i \leftrightarrow b_j) = \frac{l(a_i)}{l(A)} \cdot \frac{l(b_j)}{l(B)}
\]  

(2.1)

In biological terms, a longer domain has a bigger volume and surface area and hence a higher chance to be actually involved in the protein-protein interaction. Although this approximation remains rather crude, it accounts for unassigned domains within both interacting proteins. At the same time, multi-domain proteins do not introduce additional noise into the induced interaction network.

It is noteworthy that the sum of the weights of all domain-domain interactions resulting from a single protein-protein interaction is always \( \leq 1 \) when calculated this way. This equation was employed to calculate the background information network used in Chapter

### 2.3.5 Weighting Up-casted Interactions

In order to calculate the weight of induced interactions among higher level entities, the sum of all induced interactions among lower level entities has to be calculated. Let \( A \) and \( B \) be two nodes in the hierarchical classification \( C \), and \( a_1 \ldots a_n, b_1 \ldots b_m \) be members of the classes \( A \) and \( B \), respectively. The weight of the induced interaction \((A \leftrightarrow B)\) is calculated as

\[
weight(A \leftrightarrow B) = \sum_{i=1}^{n} \sum_{j=1}^{m} weight(a_i \leftrightarrow b_j)
\]  

(2.2)

For example, the weight of the induced interaction between two superfamilies is the sum of induced interactions among all their member families. This equation is applied recursively from the 'leaf' level (the level of domains in the context of the SCOP hierarchy) through all higher levels of abstraction.
2.4 Results

The SCOP superfamily (SF) level represents an unambiguous evolutionary relationship of its members. At the SF level proteins often show clear functional homology and this was chosen as the level of abstraction in this analysis. I used the upcasted interaction networks at superfamily level and generated the union and intersection between different sets.

2.4.1 The Union of PDB, ORF and Y2H Datasets

The results of the first analysis of protein-protein interaction data in the context of a structural classification are shown in Figure 2.3 in Park et al., 2001). In short, the Y2H, PDB, and ORF datasets were upcasted to SCOP superfamily level. This interaction map provides a comprehensive overview of the interaction data available at the time. The main advantage is that it displays about 8151 interaction between SCOP domains in terms of 664 induced interaction between superfamilies. Without up-casting such a graphical overview of all interactions would have been impossible due to the complexity of the interactome. I produced the final layout of the graph manually, as no layout algorithm was available at the time that could deal with large scale-free networks and produce a legible output.

2.4.2 The Intersection between DIP and Y2H Data

I compared the recent Y2H data published by Uetz et al., 2000; and recent PPI data gathered by Blom et al., 2004; with the PPI database of DIP (Bader et al., 1999). The SCOP class 1B were used in this analysis, excluding multi-domain, membrane or artificial proteins and models. An interaction was considered as fully assigned if both nodes constituting this edge could be mapped reliably to SCOP. This resulted in 667 fully assigned interactions from DIP (19%) and 149 from Y2H (7%). An additional 1252 interactions in DIP and 679 in yeast only one interaction partner could be given a structural assignment. These partially assigned edges were later used to embed (see Chapter 1.) the unassigned interaction partner in the context of the overall network, which yielded interaction-network-context derived structural assignments for those proteins which BLAST could not assign based on sequence similarity.
Figure 2.6: The Giant Component at SCOP-Superfamily Level, resulting from the union of Y2H and PDB derived data. See Figure [ ] for details.
Figure 2.7: The Minor Components at SCOP-Superfamily Level. Each family is represented by a shape according to its class in the SCOP database and identification number in version 1.48 of SCOP. (Ellipses: all-\alpha; rectangles: all-\beta and small proteins; triangles: \alpha/\beta; diamonds: \alpha + \beta; pentagons: multi-domain proteins; hexagons: membrane and cell-surface proteins.) The size of the shape of a family is proportional to the number of family interactions it undergoes. Black edges to a shape means some members of the family interact with each other. The interactions are color-coded as follows: blue, interactions in the PDB only; red, interactions in the PDB and yeast; grey, interactions in yeast only. Inter and intramolecular interactions can be distinguished by the type of line connecting two families: continuous line, intermolecular; thin broken line, intramolecular; thick broken line, both intra and intermolecular. More than one-third of the families are part of the giant component shown in blue, centering around the immunoglobulin superfamily (2.1.1) and the P-loop nucleotide triphosphate hydrolase superfamily (3.30.1). Other highly connected nodes are 1.110.1 (armadillo repeat), 4.117.1 (protein kinases, catalytic core) and 2.41.1 (trypsin-like serine proteases). Figure shows 229 out of at least 572 protein families that interact in the PDB and yeast. The remaining families are part of smaller isolated clusters here.
### 2.4. Results

<table>
<thead>
<tr>
<th>SCOP SF</th>
<th>Description</th>
<th>Cellular function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2.2</td>
<td>Chaperone J-domain</td>
<td>protein folding</td>
</tr>
<tr>
<td>1.41.1</td>
<td>EF-hand, calcium binding</td>
<td>signal transduction</td>
</tr>
<tr>
<td>1.110.1</td>
<td>Ras GEF</td>
<td>signal transduction</td>
</tr>
<tr>
<td>1.111.1</td>
<td>ARM repeat</td>
<td>signal transduction</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Immunoglobulin</td>
<td>signal transduction</td>
</tr>
<tr>
<td>2.32.2</td>
<td>SH3-domain</td>
<td>signal transduction</td>
</tr>
<tr>
<td>2.36.1</td>
<td>Sm motif of SNRNPs</td>
<td>Spliceosome</td>
</tr>
<tr>
<td>2.64.3</td>
<td>Trp-Asp repeat (WD-repeat)</td>
<td>signal transduction</td>
</tr>
<tr>
<td>3.2.1</td>
<td>NAD(P)-binding Rossmann-fold domains</td>
<td>Enzymes</td>
</tr>
<tr>
<td>3.3.1</td>
<td>FAD/NAD(P)-binding domain</td>
<td>Enzymes</td>
</tr>
<tr>
<td>3.32.1</td>
<td>P-loop containing NTP hydrolases</td>
<td>Enzymes</td>
</tr>
<tr>
<td>3.50.1</td>
<td>Actin-like ATPase domain</td>
<td>enzyme domain</td>
</tr>
</tbody>
</table>

Table 2.1: The breakdown of the 12 most populated hubs reveals 6 superfamilies involved in signal transduction or immunity, 4 widespread catalytic domains and 2 components involved in the assembly of multisubunit molecular machines (spliceosome and chaperones). The strong representation of signaling pathways is natural since these are based on specific protein-protein interactions (which may be localized to adaptor modules like the SH3-domain). The enzyme domains may have been co-opted in evolution to power proteins involved in a variety of cellular functions. For example, the actin-like ATPase domain is part of actin (cytoskeleton), several sugar kinases (metabolism) and heat shock protein 70.
2.4. RESULTS

(see Table) There were 1683 interactions in DIP and 1410 in Y2H where both interaction partners remained unassigned, which involved 47% of all the nodes in DIP and 63% of all the nodes in Y2H.

2.4.3 SCOP-Upcasting Conserves the scale-free Network Architecture

Figures and show the essential scale-free structure of all the interaction networks upcast to SCOP superfamily levels (Cappe et al., 2004). Scale-free networks are commonly observed in biology and the protein-protein interaction networks are no exception. Obviously the scale free property is retained in the contraction using SCOP, as there are few hubs with high connectivity keeping the network together while most nodes have a relatively low connectivity. Figure shows that the degree-distribution of the interaction network stays virtually identical throughout several levels of abstraction, from the SCOP domain up to superfamily level.

The most highly connected nodes remain more or less the same in both the union and the intersection (Figures and ) despite the differences in the databases used. The hubs for Figure are listed (top down from the graph) and discussed in Table.

2.4.4 Fast Comparison of Interaction Data

As mentioned before, the procedure described above using a given hierarchy to contract an interaction network is not the only possible way to compare interaction data from different sources. The comparison of the Y2H and DIP resulted in virtually identical graphs using two routes of upcasting: (1) map proteins in Y2H to DIP on the sequence level, then upcast equivalent interactions using SCOP; (2) upcast proteins in either Y2H or DIP individually using SCOP, then identifying equivalent interactions from their SCOP labels. Although it seems obvious given that Blast does fairly well on detecting homology and homology is a transitive function, it is a cross-validation of the resulting overlap. That both methods generate fairly identical results means that using SCOP as an identity-function is equivalent to using sequence homology (in this case BLAST) for integrating interaction networks.

The important practical implication is that this allows the strategy to
Figure 2.8: Union of Y2H and DIP data. An edge \( e = (A, B) \) in the graph was created if there are at least two edges, one \( d = (e, f) \) from DIP and another \( y = (x, z) \) from Y2H given that the \( e \) and \( x \) belong to SF \( A \) while \( f \) and \( z \) belong to SF \( B \). Then this edge was weighted with the number of underlying interactions in both sources. So each edge in the resulting graph has at least 1 underlying interaction from each source, and therefore a weight from 2 up to 19 (denoting the number of experimental interactions that induced this edge in the graph). The weight is represented by the width of each edge. Note that for all nodes in the graph there are self-interactions linking the node with itself which have been omitted in the graphic representation for the sake of clarity.
Figure 2.9: A histogram of the degree-distribution for the interaction data upcasted to different levels of abstraction. Here the probability $P(k)$ of a node (domain, family, superfamily) having $k$ edges (induced interactions) is plotted over the degree ($k$, number of interactions). The distribution stays virtually identical on all levels of abstraction.

On all levels of abstraction, this distribution follows a power law and the data points approximate a line in a log/log plot (lower plot). The distribution is stretched further on family and superfamily level compared to domain level, as some nodes become higher connected. This demonstrates that the network does indeed become more inter-connected on higher levels of abstraction as more 'context information' is accumulated.
assign each new genome and its interactome to SCOP and then subsequently compare the resulting interaction graphs within SCOP rather than comparing all genomes crosswise. Obviously, since the resulting upcasted graphs are smaller, a comparison within SCOP is much faster and easier to compute.

2.5 Discussion

In summary, I have introduced and formalized the general concepts of induced interactions, level of abstraction, upcasting and downcasting.

The concept of upcasting is similar to 'information transfer by homology' which is routinely used in functional annotation and structure assignment to genomes and in metabolic reconstruction. Here I have phrased this principle in a formal framework which applies to any classification (not limited to homology) and any source of interaction information. In particular, I note that SCOP is one of several structural classifications available that could be applied here. However, CATH and the FSSP (http://www.ebi.ac.uk/dali/fssp/), agree well for most of the protein folds and the differences in these classifications should not affect the resulting overview qualitatively.

The concept of downcasting could be used to predict novel interactions between individual proteins, but this is crucially dependent on employing refined classifications, or efficient filters, that preserve the property of interaction. This means the development of descriptors for protein-protein interfaces would complement methods for predicting interactions from gene fusion events. These descriptors of protein-protein interactions should be preferably designed to work on a sequence (like motifs) and not only on a structural level (like surface patches). Having such descriptors in place will enable genome-wide prediction of protein-interaction networks and clustering of proteins by interface similarity in "functional space".

Currently available interaction data and structural assignments are still far from complete. Hence it remains unclear whether catalytic domains (e.g. 3.32.1 assigned to SCOP) are interacting with other proteins or whether the experimentally observed interaction involves multi-domain proteins which have a specific 'interaction domain' in addition to the domains assigned to SCOP. I hope to clarify the domain issue in the future work.
2.5.1 Structure-Function Relationship

The up-casted induced networks give an overview of complex protein interaction networks and visualize interaction patterns in a structural context. The resulting graphs remain scale-free as opposed to random networks. The question why biological networks exhibit scale-free degree distributions has to remain open at this point (see Section 2.5.2 for further discussion).

The main observation here is that structural families tend to interact with one another in a specific, non-random way. This comes as no surprise, as a specific structure-function relationship is a central working hypothesis in molecular biology.

In addition, this observation supports the hypothesis that protein function can, at least to some degree, be operationalized via interaction patterns. As interaction patterns are conserved and non-random when up-casting up to superfamily level, the resulting induced networks contain a clear signal. This signal can be used for predictive purposes. Embedding of interaction patterns into the context of up-casted networks as background information results in a novel method for fold assignment (see Chapter 2.7).

2.6 Acknowledgments

Sarah Teichmann and Jong Park provided the PDB, ORF and Y2H datasets. Oliver Niggemann assisted in the graph theoretic analysis of the up-casted networks.
Chapter 3

Protein Fold Assignment via Interaction Patterns: *EMBed*

“[…] it is well known that a vital ingredient of success is not knowing that what you’re attempting can’t be done.”
Terry Pratchett, 'EQUAL RITES'

3.1 Summary

With the emerging large scale data-sets on protein-protein interactions, the question arises how accurate predictions on protein structure and function can be achieved in the resulting networks. Here I present a novel framework that successfully takes the networks’ topology into account. This *EMBed* principle is based on embedding a query pattern of interactions into the context of an upcasted interaction network called the background information network. It does not require any sequence similarity of the query protein to a protein of known structure or function and thus is a non-homology method.

As an example, I describe an application of the *EMBed* principle to fold assignment. The method predicts structural classifications from the SCOP hierarchy at over 80% accuracy up to rank 5. A query pattern is re-identified as the highest ranking prediction correctly in 75% of the cases while the signal-to noise ratio is about 100:1. The prediction accuracies represents a ten-fold improvement over the ‘guilt-by-association’ principle, the only non-homology in-
ference method for interaction data available so far.

The very fact that it is at all possible to predict structural labels successfully based solely on interaction patterns is probably the most noteworthy result. EMBed exploits the specificity of the structure-function relationship in the form of the non-random distribution of interactions between structural domains. The EMBed algorithm is general and can be applied to any hierarchical classification of protein structure or function.
3.2 Introduction

Similar to the exponential growth of available sequence and structure data observed over the last decade, an enormous growth in available interaction data can be expected for the near future. The integration of all the upcoming interaction data from various methods and organisms will form interaction networks of unprecedented complexity and size. But a central question remains: given all these networks, how can this information be harnessed to arrive at an accurate prediction of the function and possibly the structure of the proteins involved?

Together with the publications of the first large-scale yeast-two-hybrid interaction screens, prediction of protein function was attempted based on the resulting interaction data (Mayer and Hieter, 2000). The logic underlying the so-called guilt by association principle, or GBA for short, may simplistically be summarized as follows: 'If two proteins interact with one another, they usually participate in the same, or related, cellular functions.'

As described in Chapter 7, function can be operationalized via functional attributes associated with every protein. These attributes are labels from, for example, the Enzyme Commission (EC) numbers, the Gene Ontology (GO), or the Structural Classification of Proteins (SCOP).

Following the GBA principle these labels can be transferred along established interactions. The benefit of such a non-homology approach is that it works in the absence of any similarity of the protein under study to a previously characterized protein. Hence a non-homology method can be used to predict function for the large fraction of ORFans within newly sequenced genomes - those sequences which have no detectable homology to any protein of known function.

Given the unknown protein and its interactions, how is a prediction made about its' function? The unknown node q under consideration I will call the query node, and the known interactions of q the query pattern. Using the GBA principle, the functional attributes from all interacting proteins of the query node q are collected. Then, within this set of neighbors, a majority vote is taken. So functional attributes are transferred from the neighborhood onto q - thus the name guilt by association, which was introduced by (Walker et al., 1998). The significance of a functional attribute is estimated using the

41
frequency of their occurrence.

Variations of the GBA principle arise from using different statistics to calculate the significance of a functional attribute within the set of neighboring nodes. The method is not necessarily restricted to the direct neighbors \( (d = 1) \), but can include all nodes within a distance of \( d \) interactions from \( q \) (see Figure 4). Due to the small-world behavior of the interaction networks any distance \( d > 3 \) will always include a major fraction of the overall network (see Figure 5) and hence is of no practical importance.

In how far does the interaction neighborhood really reflect different aspects of protein function? It is easy to see why the GBA principle should work for sub-cellular localization: If, for example, all interacting partners of \( q \) are located in the cytosol, it can be safely assumed that \( q \) is located in the cytosol as well (see Figure 6). Obviously this method may also work to some extent for associating a protein to larger biological processes, albeit this definition of function may be less clearly reflected in neighborhood relationships. Considering the biochemical definition of function, it is even less clear why interacting proteins should bind the same set of small molecules or even catalyze them in the same way. These considerations have been confirmed by Figure 7. They found prediction accuracies of 72.7% for subcellular localization, 63.3% for biological process and 52.7% for biochemical function using guilt-by-association.

Interactions between members of the same structural family or superfamily are frequently observed (see Section 7). This would indeed allow interaction information to be harnessed for structure prediction using the GBA principle. As shown in Figure 8, GBA performs better than random but does not deliver SCOP predictions at a reasonable accuracy. As one would expect, interacting proteins do not necessarily share the same structure.

The guilt-by-association principle only accounts for existence of a protein and its' associated attributes in the interaction neighborhood. The actual topology of the overall network is an important piece of information that is left out. Here I will explore to the idea to use the topology of interaction networks. The resulting non-homology method produces accurate assignment of structural labels from SCOP via interaction patterns alone.
Figure 3.1: Illustration of the guilt-by-association (GBA) principle. In the center is the unknown query node \( q \). The lines depict known interactions, while the nodes symbolize proteins. Functional attributes (like structure or sub-cellular localization) are represented by the different shapes and colors. All proteins with a direct interaction to \( q \) have a distance of \( d = 1 \). Proteins that can only be reached in another step from the first shell have a distance of \( d = 2 \) from \( q \) and so forth.

Assume, for example, node color codes for subcellular localization and red represents mitochondrial proteins. From the knowledge that proteins located in the mitochondria are over-represented in the first shell of \( q \), the query protein would be assigned as a mitochondrial protein according to the guilt-by-association principle.
Figure 3.2: Finding a SCOP label for a given interaction pattern. The graph here is identical to Figure 3.1 and represents the background information of interactions on SCOP superfamily level.

In this example, the interaction partners of a query node $q$ could be mapped to the superfamilies highlighted by the red circles (1.111.1, 2.64.3, 3.32.1). There is only one node, namely 3.50.1 (highlighted in green), that satisfies the constraint to interact with all of the superfamilies identified by the query pattern. This node represents the most likely assignment of $q$ to a structural superfamily.
3.2.1 Embedding Query Patterns

The interaction maps in Chapter see Figures and clearly demonstrate that structural entities (like superfamilies) interact with a specific set of other structural classes. Here specific means the interaction network is non-random on a given level of abstraction. As interaction patterns are conserved and specific, the interaction networks contain a signal that can be used for structure prediction. This network of induced interactions \( I \) among SCOP labels \( S \) (as described in Section ) is referred to as the background information network \( B = (S, I) \).

This is illustrated in Figure for (some of) the known interaction partners of \( q \) the respective SCOP labels can be determined. The nodes corresponding to this set of SCOP labels can be located in the overall network representing the background information. In contrast to the GBA principle, the working hypothesis here is that proteins with similar interaction patterns have similar functions and structures.

Now the task is to find a node in \( s \in S \) such the query pattern is matched in an optimal way with \( I \). In other words the query pattern is embedded in the background information network.

The node \( s \in S \) in the background information network which satisfies the constraint to interact with all (or most) of the nodes identified by the query pattern is the most parsimonious explanation for this pattern. Hence the node \( s \) identified this way represents the most likely structural assignment for \( q \).

Put generally, the structure-function relationship can be exploited in a reverse fashion: by (partial) knowledge of protein function through interaction patterns the number of structural domains consistent with this information is very much reduced. Following this principle it should be possible to predict structural labels reliably based on interaction patterns alone. Such an application of interaction information to fold assignment has not yet been undertaken.

3.2.2 Algorithmic Principles

The main algorithmic idea is to re-identify a query node \( q \) from any network modeled as an undirected graph \( G = (V, E) \) just with the information about the neighbors of \( q \) at hand (see Figure ). Although subgraph-isomorphism is an \( NP \)-complete problem for unlabeled graphs, in this special case the
Figure 3.3: Identifying a query node by network context. The query pattern consists of an query node $q$ (black sphere) representing an unknown protein together with known interaction partners (grey spheres). This query pattern is depicted on the left.

The embedding of the query pattern works as follows: On a first pass, the direct neighbors of the query node $q$ are identified in the network representing the background information (right). These nodes are called activated. Then in a second pass every activated node sends a token to all of its’ neighbors in the network. These tokens are represented by the incoming edges on the right.

Obviously, the node $q$ is re-identified by receiving the highest number of tokens. Furthermore, the node has a high likelihood to be re-identified even if the initial interaction information was incomplete and noisy.
background information network is a labeled graph and \( q \) is the only unlabeled node.

Let \( q \) be a node from the graph \( G = (V,E) \) and \( D \) be the set of direct neighbors of \( q \) such that \( \forall d \in D : (d \leftrightarrow q) \in E \). In a first pass all nodes \( d \in D \) are identified or activated. In a second pass all activated nodes send one token to each of their respective neighbors.

Let \( w(v) \) denote the number of tokens gathered by any node \( v \in V \). It is obvious that no other node in the graph will gather more tokens than \( q \). Hence the condition \( w(v) \leq w(q) \) can be asserted. In fact, the case that another node will gather the same amount of tokens as \( q \) such that \( w(v) = w(q) \) is only encountered if the all neighbors of \( q \) are neighbors of \( v \) as well. In this case the interacting partners of \( q \) form a subset of the interacting partners of \( v \). Accordingly, \( q \) and \( v \) can be expected to have similar structure.

In biological terms, all nodes that gain a token in the second pass are potential structural assignments for \( q \) given the background information network. The set of all nodes that receive a token form the result-set of the query. The nodes in this result-set can be ranked according to the number of tokens they received. If all edges of \( q \) were part of the query then \( q \) has to turn out on rank 1 of the result-set. Although the top-ranking predictions may not be unique, if the interaction information is complete and accurate this method will re-identify \( q \) based only on neighborhood information.

In addition, the method even works if the query pattern is noisy or incomplete. As long as at least a single true edge of the original interaction pattern remains in the query pattern, \( q \) will receive at least one token and hence be a member of the result-set. However, the more of the original pattern is lost, the less likely \( q \) will turn out on the top rank.

In summary, the algorithm outlined above describes a method for structure prediction that can work with noisy and incomplete interaction data. All “knowledge” about protein structures is encoded in the background information network. Obviously the construction of the background interaction network itself is crucial to the accuracy of this approach.


3.3 Methods

3.3.1 Interaction-Datasets and Mapping to SCOP

The protein interaction data available from DIP and yeast-two-hybrid experiments were downloaded and mapped to SCOP version 1.48 using BLAST with an e-value worse than $10^{-5}$ and overlapping assignments were discarded. Finally, interactions between SCOP labels were induced, weighted by the relative sequence length of the assignment.

3.3.2 Generating the Background Information Network

The background information network $B = (S, I)$ consists of the all induced interactions $I$. The labels on all levels of SCOP form the set of nodes $S$, excluding the class level at the top of the hierarchy. This class level was omitted because there are only three classes and it helped to keep the graph at reasonable size. All edges are weighted with the sum of the induced interactions (see details).

3.3.3 Embedding Algorithm

The $EMBed$ algorithm embeds a query pattern $P$ into a given background information network $B = (S, I)$ with $S$ being the set of SCOP labels and $I$ the induced interactions between them. The query pattern consists of the query node $q$ and its directly interacting proteins $D_P$.

$$P = \{(q \rightarrow p) : p \in D_P\} \quad (3.1)$$

A (partial) mapping function $M : O \rightarrow S$ from the set of all proteins $O$ to SCOP labels $S$ is given. In the first pass, the query pattern is linked to the background information network via $M$.

$$D_A = \{r : (p \rightarrow r) \in M, p \in D_P, r \in S\} \quad (3.2)$$

This mapping represents the known SCOP assignments for all proteins directly interacting with $q$. Now $D_A \subset S$ are the nodes $directly$ $activated$ through $D_P$ from $q$. 
Figure 3.4: The **EMBed** Algorithm: constructing a subgraph from \( q \). The task is to assign a structural label to the query protein \( q \) on the left. The information given is the set of direct interacting proteins \( D_P \). The arcs from \( q \rightarrow D_P \) represent the known interactions or the **interaction pattern** of \( q \) (see equation).

The induced interactions among SCOP labels form the background information, illustrated by the grey shaded area. The SCOP assignment for all protein in \( D_P \) provides a mapping from \( D_P \rightarrow D_A \), denoted by the dotted lines. The nodes in \( D_A \) are SCOP labels within the background information network and they are **now activated** if they are SCOP assignments for any node in \( D_P \) (equation).

In the second pass all activated nodes \( D_A \) send tokens to all neighboring SCOP labels along induced interactions. Because they are **indirect** assignments they are denoted as \( I_A \). The weight of these tokens is the same as the weight of the interactions \( D_A \rightarrow I_A \) (equation).

The result-set \( I_A \) is now ranked according to the sum of all the tokens gained in the second pass. This score is equivalent to the sum of the weight of all incoming edges for every node in \( D_I \) (equation). The node with the **maximum score** in \( D_I \) is the most likely structure assignment for \( q \).
Figure 3.5: Benchmarking EMBed Predictions. The SCOP assignments for any given protein produced by BLAST represented the standard of truth, depicted on the left. The predictions produced by EMBed for this protein as a query node $q$ were ranked according to their score in descending order. The EMBed prediction with the lowest rank that matched a BLAST assignment was reported as a true hit.

In the second pass, those SCOP labels within the background information network $B = (S, I)$ adjacent to the activated nodes form the set $I_A$ of indirectly activated nodes, where

$$I_A = \{ s \} : (r \leftrightarrow s) \in I, r \in D_A, s \in S$$ \hspace{1cm} (3.3)

All the nodes in $I_A$ are potential SCOP predictions for the query node $q$. The score for every node in $I_A$ is calculated as the sum of the weights of all incident edges from $D_A$. Let $s$ be a node in $I_A$, $p$ a node in $D_A$, and $w(s \leftrightarrow p)$ the weight of the induced interaction $(s \leftrightarrow p) \in I$. Then

$$score(s) = \sum (w(p \leftrightarrow s)), p \in D_A, s \in I_A.$$ \hspace{1cm} (3.4)

All nodes $s \in I_A$ are ranked according to their respective score. The node in $I_A$ with the highest score represents the best prediction, as it is the maximal match of $P$ to the induced interactions in the background information network.

In general, the EMBed algorithm can also be defined as a solution to the maximum flow problem from the source $q$ to a sink $s \in I_A$ (Laptev et al., 2004). However, given the construction of the weighted subgraph $q \rightarrow D_P \rightarrow D_A \rightarrow I_A$, this is equivalent to the maximum sum of all incident edges for a node $s \in I_A$. 
### 3.3.4 Benchmarking

SCOP domains (SCOP version 1.48) were assigned to all protein sequences in *S. Cerevisiae* using BLAST. All SCOP assignments with an e-value ≤ 10^-5 were regarded as significant and constituted the 'standard of truth' of structural domains for every yeast ORF. The SCOP predictions produced by *EMBed* were ranked according to their score. An *EMBed* prediction was regarded as correct if the predicted SCOP label was also found by BLAST. The correct prediction with the lowest rank was reported as a true hit (see Figure).

The percentage of proteins with a correct assignment at any given rank $x$ and the cumulated percentage of correct predictions with rank $\leq x$ are plotted in Figure. A 'leave-one-out-test' on the predictive accuracy was performed such that the query node $q$ and all adjacent interactions were removed from the background information network prior to the *EMBed* prediction for $q$. In this exclusive case the query pattern had to be matched against the background information induced from other experimental data than the query pattern.

In contrast, in the inclusive case the interactions induced from the query pattern were left in the background information network. This way I measured the accuracy of *EMBed* in re-identifying interaction patterns that were already experimentally observed.

Finally, the probability of arriving at a correct prediction by chance using *EMBed* was measured. For every query protein $q$ the number of interacting partners $k$ (or the degree of $q$) was calculated. Then a set of $k$ randomly chosen proteins was generated. Subsequently, SCOP predictions with *EMBed*, ranking and benchmarking were performed on this randomized set as described above.

### 3.4 Results

The very fact that it is at all possible to predict structural labels successfully based on solely interaction patterns is probably the most noteworthy result. *EMBed* exploits the specificity of the structure-function relationship in the form of the non-random distribution of interactions between structural domains. The method works in the absence of any sequence similarity and produces accurate structure assignments.
3.4. RESULTS

3.4.1 Accuracy

After providing the first ‘proof of principle’ that interaction patterns could be used for fold assignment I was able to improve the predictive accuracy significantly. The main difference between the current version and the results reported initially lies in the way the background information was constructed. Instead of using a standard weight of 1 for every induced edge, the edges were weighted according to the relative sequence length of the assigned domains (see Section 3.2). This way less noise was introduced into the network representing the background information. The current accuracies of SCOP-predictions produced by EMBed are summarized in Figure 3.2. EMBed produces a list of predictions which are ranked according to their score. The prediction with the highest score is the one most likely to be correct. For the evaluation of the performance I distinguish two cases:

- In the exclusive case the interactions induced by the query pattern were removed from the background information. Here the query pattern had to be matched against interactions induced from other proteins. From a systematic viewpoint, this leave-one-out-test is more consistent with the original analysis where about 1% of the interaction data was set aside as a test-set.

- In the inclusive case the data induced by the interactions used in the query were left in the background data. Hence this analysis shows how accurately a pattern is re-identified by EMBed. The re-identification will be of more practical interest as the available interaction information will become more comprehensive over time.

About 50% of the highest ranking assignments produced by EMBed match an assignment from BLAST in the exclusive case. Including all predictions up to rank 5, a correct assignment was found for about 80% of the query proteins. Compared to 25-30% on rank 1 and 43% up to rank 5 found to be correct in the current version is almost twice as accurate.

In the inclusive case the accuracy increases to over 75% at rank 1 and about 95% at rank 5. This means that if the query pattern is available in the database then the correct structural label is very likely re-identified within the top 5 predictions.
Figure 3.6: Accuracies of SCOP-predictions produced by EMBed. The percentage of EMBed predictions matching a structure assignment by BLAST is plotted over the rank of the prediction. On the left the percentage of correct predictions at any given rank is shown. On the right the cumulated percentage of all correct predictions with rank \( \leq x \) is plotted. In the exclusive case the query pattern was removed from the data prior to prediction. The curve marked inclusive shows the accuracy in re-identifying the query pattern in the data. The random curve shows the probability of arriving at a correct prediction by chance using a randomly generated query pattern in EMBed.
3.4.2 Signal-to-Noise Ratio

In order to estimate the signal-to-noise ratio of the *EMBed* SCOP predictions I applied the same principle to random query patterns. These random query patterns were generated by replacing the interaction partners in the original data by randomly chosen proteins. As the number of interaction partners for every node in the network stayed the same, the statistical distribution of the original interaction data was kept.

With these random neighbors as the query pattern for q, a prediction was generated. Subsequently, the predicted SCOP labels were ranked and compared with the known BLAST assignments for q. Figure 3.4 shows the probability of retrieving a correct structural label using a random query pattern stays well below 0.5% on rank 1. Hence a conservative estimate of the signal-to-noise ratio of *EMBed* would be about 100:1.

3.4.3 Comparison with guilt-by-association

How much predictive power is gained by the *EMBed* principle in exploiting the network topology compared to the guilt-by-association principle?

For a comparison of the respective performance of both methods I implemented the $\chi^2$ test (as described in Figure 3.3). Then prediction of subcellular localization for all interacting proteins was performed, based on data provided by My reference-implementation of the GBA principle achieved over 70% accuracy for subcellular localization. This result matches the performance of GBA described by Hishigaki and co-workers and suggests that the reference-implementation worked correctly. However, a predictive accuracy of over 70% was reached when cumulating the correct predictions up to rank 7 (see Figure 3.5). This puts the performance of GBA into perspective: only 22 categories were used in the study by Hishigaki and co-workers, while the dataset used here was even reduced to 9 distinct categories of subcellular localization. At this point it is noteworthy that both the GBA and *EMBed* method have to choose among thousands of SCOP labels when applied to fold assignment. The reference-implementation of GBA was then applied to the prediction of SCOP labels. The results are summarized in Figure 3.6. Looking at the highest ranking predictions, *EMBed* is 10 times more accurate than GBA. Most encouraging, *EMBed* is more accurate in SCOP assignments than GBA is on the prediction of sub-cellular localization, although SCOP assignment
3.5 Sample Predictions

A sample of assignments obtained for hypothetical yeast proteins are shown in Table 3.1. YEX6_YEAST matches the sequence conservation pattern of SNRNP9s (small nuclear ribonucleoproteins) and is clearly a remote family member. YHW1_YEAST has weak sequence similarities to cytoskeletal assembly proteins, and structural proteins are often helical so the alpha-alpha superhelix fold assignment is not an unreasonable prediction. Both YES2_YEAST and YHX1_YEAST hit the P-loop superfamily, but there is no identifiable sequence similarity to support a fold prediction. The target SCOP class is also extremely large, so this hit may be spurious. The assignment of PAF1_YEAST is a possible discovery, which is supported by functional similarities. PAF1_YEAST is involved in transcriptional regula-

### Table 3.1: Fold Assignment Predictions for Hypothetical Yeast Proteins

<table>
<thead>
<tr>
<th>Swissprot ID</th>
<th>length</th>
<th>e-value</th>
<th>SCOP label</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEX6_YEAST</td>
<td>93</td>
<td>1e-10</td>
<td>2.36.1.1</td>
<td>SNRNP</td>
</tr>
<tr>
<td>YHW1_YEAST</td>
<td>637</td>
<td>0.5</td>
<td>1.111</td>
<td>alpha-alpha superhelix</td>
</tr>
<tr>
<td>YES2_YEAST</td>
<td>125</td>
<td>100</td>
<td>3.32.1</td>
<td>P-loop containing NTP hydrolases</td>
</tr>
<tr>
<td>PAF1_YEAST</td>
<td>445</td>
<td>-</td>
<td>1.4.5.15.1</td>
<td>DNA-binding domain from rap30</td>
</tr>
<tr>
<td>SED4_YEAST</td>
<td>1065</td>
<td>-</td>
<td>3.17</td>
<td>flavodoxin-like</td>
</tr>
<tr>
<td>YNI6_YEAST</td>
<td>102</td>
<td>-</td>
<td>3.56.1.1.13</td>
<td>phosphoribosyltransferase</td>
</tr>
<tr>
<td>YG33_YEAST</td>
<td>275</td>
<td>-</td>
<td>2.52.1.2</td>
<td>molybdenum cofactor biosynthesis</td>
</tr>
<tr>
<td>YHX1_YEAST</td>
<td>630</td>
<td>-</td>
<td>3.32.1.13.3</td>
<td>P-loop containing NTP hydrolase</td>
</tr>
</tbody>
</table>

is undoubtedly a much harder problem.

From the data shown in Table 3.1, I estimate the signal-to-noise ratio of GBA to be in the range from 3:1 (subcellular localization) up to 12:1 (cellular role). In summary, both accuracy and signal-to-noise ratio increase about ten-fold by using the information from networks topology for predictions.
3.5. SAMPLE PREDICTIONS

Figure 3.7: Comparison of EMBed with guilt-by-association Here the percent of correct predictions made by both methods is plotted over the rank. The performance of EMBed in both the inclusive and exclusive case are shown for reference (see Figure 3.6).

In predicting sub-cellular localization (GBA:SCL), the guilt-by-association reference-implementation reaches over 70% accuracy only on ranks $\geq 6$. The dataset used was reduced to less than nine categories for subcellular localization containing significant numbers of interactions, hence the maximum rank here is seven.

The application of GBA to the prediction of SCOP labels (GBA:SCOP) reaches about 5% correct predictions at rank 1 and 13% at rank 5. Hence EMBed outperforms GBA about ten-fold.
tion, and the SCOP assignment is to a DNA-binding domain in the 'winged helix' superfamily.

3.6 Discussion

The *EMBed* principle represents a novel framework for fold assignment based on embedding interaction patterns into a background information network. I implemented the algorithm using a set of protein-protein interactions derived from DiffDB and yeast-two-hybrid experiments (Ito et al., 2000) and mapped to the Structural Classification of Proteins (SCOP).

The advantage of the method described here is the ability to produce assignments for proteins where BLAST could not detect any clear homology. It extends fold assignment to proteins beyond the current limit of homology-based methods. Since *EMBed* relies on context-information rather than homology, this approach can be characterized as a non-homology method. Homology is only needed to determine where the interaction partners of the query protein are located within the overall network, but not for the query protein itself. The only information required is the interaction pattern of the query protein.

*EMBed* outperforms the existing *guilt-by-association* principle ten-fold. Obviously, accurate fold assignments are produced in the face of noisy and incomplete data. Not only the available interaction data is still far from complete and not perfectly accurate (see Chapter... In addition, despite the ongoing structural genomics projects, not all proteins have reliable structural assignments. As a result, the background information network is consequently fraught with noise, and the mapping of interaction patterns to the network is partial most of the time. However, these problems will be of less significance as more interaction information and more proteins structures will become available.

Despite the shortcomings of the available data, the application to fold assignment seems to outperform sequence-based methods (compare (McGuffin et al., 2004), but is still lacking a rigorous statistical model to evaluate the predictions made. The scoring function is quite simple and straightforward, and there is still room to improve the assignment procedure by improving the scoring function.

So far there is no other prediction method available that uses interaction
information for fold assignment, although gene neighborhoods have been used to assign function (Huynen, Snel et al. 2000). More recently, an approach similar to \textit{EMBed} has been described, using interaction data with the MIPS functional classification scheme. Although a different classification scheme is used by Vazquez and co-workers, their results match and corroborate the accuracy of \textit{EMBed} for structural and functional predictions.

The ideas underlying the \textit{EMBed} principle are general and invariant to the source of interaction information as well as to the type of classification used. Hence the framework described here can be applied beyond structural classifications. For example, to apply \textit{EMBed} to the prediction of functional labels means employing a different hierarchical classification of protein function, just in the same way as SCOP encodes knowledge about protein folds. For example, a collaboration with the IntAct team to apply \textit{EMBed} to the prediction of functional labels from the Gene Ontology (GO) (Ashburner et al., 2000) based on the data curated in the IntAct interaction database is currently underway.

3.7 Acknowledgments

Konrad Pazquiewitz helped with the reference-implementation of the \textit{guilt-by-association} principle. Gregory Gimenez, Markus Brosch and Henning Hermjakob assisted in the ongoing application of the \textit{EMBed} principle to GO label prediction for the IntAct project.
Chapter 4

Automatic Identification of Signaling Pathways

“He was determined to discover the underlying logic behind the universe. Which was going to be hard, because there wasn’t one. The Creator had a lot of remarkably good ideas when she put the world together, but making it understandable hadn’t been one of them.”
Terry Pratchett, ’MORT’.

4.1 Summary

Current protein-protein interaction networks, however incomplete they may still be, already are so complex that they cannot be perceived or analyzed in depth as a whole. Here, I aim at automatic extraction of signaling pathways from complex networks in a way that is readily accessible to human interpretation.

This analysis focuses on a collection of statistical associations among proteins derived from text-mining and validated by experimental high-throughput interaction data. I model the transfer of an external signal from a given receptor to a set of transcription factors in the nucleus based on concepts of information flow. Derived from first principles, the method is fully automatic and does not require any training set. I adapt the maximum spanning tree algorithm for fast and accurate reconstruction of pathways.

This pathway extraction strategy is general, insensitive to noise in biological networks and readily applicable to any set of receptors.
in any organism. Here, I show that the method reliably extracts known signaling cascades in yeast, yielding results that are consistent with available biological knowledge and meet human expectations. Implicitly, as the method reconstructs pathways, this work gives an algorithmic definition of 'signaling pathway'.
4.2 Introduction

Despite the huge differences between all the computational and experimental techniques available, there are nevertheless some common emerging properties and problems of interaction networks, namely complexity, incompleteness, noise, a scale-free degree distribution and small-world behavior (see Chapter). In summary, the dilemma is that the ever-increasing wealth of information arising from experiments and computational analysis, while aiming for completion, gives rise to scale-free small-world networks that are so complex that they cannot be perceived or analyzed as a whole.

The goal is to tackle the complexity of these networks and extract subsets which represent biologically meaningful processes. These subsets should be readily accessible to human interpretation and serve as a basis to formulate experimentally verifiable hypotheses. Classically, the complexity of biological networks has been tackled by decomposing the complex cellular machinery into several key pathways.

Despite large scale efforts to comprehensively screen for protein-protein interactions, the picture of the yeast interactome is far from complete. Here, I tap into the vast amount of information available from Medline abstracts to complement the available experimental data. I use a statistical approach (similar to Marcotte et al., 2002) to compute protein-protein associations from Medline-abstracts. This procedure yields a good coverage of interaction space with reasonable accuracy. Overall, 41129 associations covering 3616 proteins in yeast were detected. The resulting associations significantly extend the available data in a completely automatic way. They link experimental complexes, capture transient and indirect interactions, and close gaps in the databases which are due to technical limitations of the available high-throughput approaches.

Here the network arising from the statistical associations is used as a typical test-case for a noisy, scale-free, small-world network representing the complex cellular environment. Unlike most other experimental datasets, the statistical associations have the advantageous property that they form a weighted graph. In other words, every association comes with a measure of strength 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.
4.2.1 Aiming for higher Levels of Organization

Previous attempts to extract higher levels of organization from networks revolve around the notion of *functional modules*, which have been defined as strongly connected components of interaction networks mainly consisting of proteins involved in the same function. 

This approach can be tackled by graph-based clustering of the interaction networks, based on either topological properties and / or the clustering of functional attributes of the proteins. This definition comes close to what may be regarded as protein complexes, which were shown to be dynamic entities undergoing re-arrangements in response to changing conditions (Gavin et al., 2002). However, traditionally biologists view the complexity of biological networks not in terms of compact modules but in terms of *pathways*, like metabolic pathways or signal transduction pathways.

How to extract such pathways from large protein-protein interaction networks in a fast and reliable manner? This question led to the idea of modeling the information flow in such a network. The signal received by a receptor is a valuable piece of information that is passed on like a packet in a computer network. While the information should not get lost in the maze, it is important to note that no protein has an overview on how to optimally route the information. So, at each stage there can be only *local* decisions about how to pass the information on towards its' destination in a reliable fashion.

As I will focus further on modeling the flow of information through cellular networks, this analysis will concentrate on the reconstruction of signaling pathways. Although the approach described here may not work for metabolic pathways, the apparent analogy is that metabolic pathways essentially model the flux of metabolites.

In contrast to clusters of proteins of similar function resulting from the definition of 'functional modules' above, signaling cascades have a different, more 'linear' architecture. Signaling cascades are constructed from a variety of different proteins such as receptors, enzymes, channels and regulatory proteins with the purpose of carrying the information conveyed by external stimuli into the nucleus and triggering the appropriate cellular responses. This also involves information processing, for example amplification, modulation and integration with other signals. Therefore deriving a static picture of signaling pathways, which are very dynamic in their response to external stimuli and environmental factors, is an abstraction not easily achieved.
Figure 4.1: Illustration of the pathway reconstruction algorithm. The calculation in this example starts from a single receptor on the cell-surface $S$. Initially, the pathway $P$ only consists of $S$. All adjacent nodes to $P$ are in $Q$ (depicted as cone-shaded nodes). Subsequently in each step $P$ (diamond shaped nodes) is extended along the maximum edge leading into $Q$. Note that this maximum edge is not necessarily adjacent to the node added in the previous step, as illustrated in the transition from the stage depicted on the top right to the one on the lower left. All hitherto unknown nodes adjacent to $P$ are added to $Q$ while backward edges are excluded. The process stops when a pre-defined number of transcription-factors ($T$) is hit. Finally backtracking from $T$ to the start-node $S$ results in the backbone pathway, which is subsequently augmented with associated factors.
4.2.2 From first Principles towards an Algorithm

How might an information channel as described above be constructed from bio-molecular components? The basic idea is this: given a start- and end-node, extract a path from a weighted Graph. However, the application of the shortest-path algorithm to this problem is inhibited by the *small-world* behavior of the underlying network. Due to ubiquitous weak associations there is always a short-cut between any two proteins in such a network. In particular, this means that biological pathways rarely coincide with the shortest path.

In contrast, I model the concept of information flow in a network from the cellular exterior into the nucleus. In this context, the (statistical) strength of an association from the text-mining is taken as the capacity of information this edge can transmit. In order to transmit the information reliably, the information is routed along the strongest edges in the network.

Put simply there is no *routing information* available, every protein "sees" only its direct neighbors and does not "know" anything about the global network connectivity. Hence these packets or units of information are transferred to a direct interacting neighbor based on local information and interactions. The requirements are that this forwarding of information has to happen in a reliable manner until a destination point is reached. The forward direction of the information flow can be ensured by a simple memory mechanism which prevents backwards flow. In physical terms, the information is passed on as a series of state changes of the proteins involved, like phosphorylation state or other post-translational modifications, conformational changes or cleavage.

In the case of signaling pathways it is relatively straightforward to define a start and stop-set for the pathway search. Receptors for extra-cellular signals are located in the cell-membrane and can be distinguished by their trans-membrane domains. The stop-set can be defined by the set of all known transcription factors. Through the control of the transcription factors the corresponding cellular response is triggered.

How is the biological intuition translated back into a computer algorithm? I define network connectivity based on the co-citation of gene names in Medline abstracts and derive information about the reliability of an interaction indirectly from the number of co-citations (see Table). Based on the concept of information flow outlined above, the pathway-finding algorithm follows the most reliable associations first, resorting to weaker associations
only if no other, stronger link is available (see Figure 1). If the strength of an association in the network is interpreted as its height in a landscape, the algorithm tries to stay on high, firm ground as long as possible. It then only takes the slowest descent and directly uses stronger links as they become available, until the destination is reached. Backtracking which nodes have been visited en route results in a linear backbone pathway connecting the start node to the destination node. I then combine the resulting pathway-backbone with only those proteins that are strongly (above average) associated to it. This reflects that although the strongest associations may not result in the main pathway but they do give the general direction of information flow. The resulting graphs represent a compilation of a huge body of data in a form that is easily accessible to subsequent analysis by biologists.

4.3 Methods

4.3.1 Experimental Datasets

Experimentally derived data on yeast interactions was retrieved from the following four databases:

1. DIP
   Database of Interacting Proteins
   (March 2002 release; http://dip.doe-mbi.ucla.edu/)

2. MINT
   Molecular INTeraction Database
   (http://cbi.com.uniroma2.it/mint/)

3. CZ
   Cellzome dataset
   (http://www.cellzome.com)

4. MDS
   MDS Proteomics dataset
   (http://www.mds-proteomics.com)

While DIP and MINT contain binary interactions, the TAP-MS data from CZ and MDS contain sets of proteins from a number of pull-downs. Here the data was broken down into a complete set (clique model) of binary interactions and marked as bait-prey and prey-prey interactions respectively. All experimental data were used only to highlight which statistical associations are experimentally verified as physical interactions, they were not used in the calculation of the pathways.
4.3.2 Start and Stop-set

For defining the sets of start-nodes and terminal nodes, I downloaded data from YPD. As start nodes for the calculation all proteins marked as receptors or with seven trans-membrane domains where used. To generate the set of terminal-nodes I compiled a list from YPD containing all proteins described as transcription factors.

4.3.3 Synonym Dictionary

Protein names were mapped to gene / protein names occurring in Medline abstracts using a synonym dictionary. The synonym dictionary containing the complete set of unique yeast ORF identifiers and the corresponding gene names and their synonyms was compiled from publicly available information in the databases SGD, MIPS, and SWISS-PROT.

4.3.4 Text Corpus

A total of 61470 journal abstracts were downloaded from Medline using the SRS server (http://srs.ebi.ac.uk). Each downloaded Medline entry was required to contain at least one ORF/gene name or one of its associated synonyms in the text body of the abstract or in the title. In addition, the MESH term 'Saccharomyces cerevisiae' was required to limit the search to the chosen model organism, yeast. These selection criteria for Medline abstracts mean that completely uncharacterized yeast ORFs, which are not described in the literature yet, are excluded from the current statistics.

4.3.5 Text Mining

Occurrences of yeast gene names were tallied as 1, if a gene name or any of its synonyms was present in a Medline abstract, and 0 otherwise (see Table). In other words, I use a discrete two-state model of presence or absence of a gene in a Medline entry. The data set yielded a total of 41129 co-occurring pairs of genes, among which 10285 were co-occurring in two or more abstracts. A total of 3616 proteins are covered by this network.
Proteomic analysis of nucleoporin interacting proteins.
Allen NP, Huang L, Burlingame A, Rexach M.
Department of Biological Sciences, Stanford University,
Stanford, California 94305-5020, USA.
The Saccharomyces cerevisiae nuclear pore complex is a
supramolecular assembly of 30 nucleoporins that cooperatively
facilitate nucleocytoplasmic transport. Thirteen nucleoporins
that contain FG peptide repeats (FG Nups) are proposed to
function as stepping stones in karyopherin-mediated transport
pathways. Here, protein interactions that occur at individual
FG Nups were sampled using immobilized nucleoporins and yeast
extracts. We find that many proteins bind to FG Nups in highly
reproducible patterns. Among 135 proteins identified by mass
ectrometry, most were karyopherins and nucleoporins.
The PSFG nucleoporin Nup42p and the GLFG nucleoporins Nup49p,
Nup57p, Nup100p, and Nup116p exhibited generic interactions with
karyopherins; each bound 6--10 different karyopherin betas,
including importins as well as exportins.
Unexpectedly, the same Nups also captured the hexameric Nup84p
complex and Nup2p. In contrast, the FXFG nucleoporins Nup1p,
Nup2p, and Nup60p were more selective and captured mostly
the Kap95p.Kap60p heterodimer. When the concentration of
Gsp1p-GTP was elevated in the extracts to mimic the
nucleoplasmic environment, the patterns of interacting proteins
changed; exportins exhibited enhanced binding to FG Nups, and
importins exhibited reduced binding. The results demonstrate
a global role for Gsp1p-GTP on karyopherin-nucleoporin
interactions and provide a rudimentary map of the routes that
karyopherins take as they cross the nuclear pore complex.
PMID: 11387327

Table 4.1: Example of Text-Mining in a Medline-Abstract. The synonyms
recognised are highlighted (i.e. Nup2p). This abstract contributes a count
of 1 to the co-occurrence between all the 12 different highlighted genes and
proteins. The sum of all co-occurrences derived from all medline-abstracts
available yield the network used for the pathway extraction.
4.3. METHODS

The decision of which association is the strongest within a given set during the pathway calculation is based on the raw co-occurrence count. Only when several potential associations have the same number of co-occurrences, is the value of the mutual information used to break the tie. The mutual information statistic ($MI$) from information theory measures the extent of association between two genes in a given document collection beyond that expected resulting from chance. The mutual information $MI(X, Y)$ between two genes $X$ and $Y$ is given by:

$$MI(X, Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$  \hspace{1cm} (4.1)

4.3.6 Pathway Reconstruction Algorithm

The pathway algorithm implements the greedy strategy to add the strongest edge from the neighborhood of the growing pathway in every step.

First of all, the undirected, weighted graph of pair-wise associations is translated into a directed graph by replacing every edge $(u \leftrightarrow v)$ into two directed arcs $(u \rightarrow v)$, and $(v \rightarrow u)$.

First the undirected graph representing all the pair-wise statistical associations derived from the text-mining is translated into a directed graph such that every undirected edge $e = (u \leftrightarrow v) \in E$ is represented by two arcs $e_1 = (u \rightarrow v)$ and $e_2 = (v \rightarrow u)$. The weight on both edges $e_1$ and $e_2$ is identical to the weight of $e$.

Now all the nodes that 'have' information are marked as potential members of the pathway $P$. Initially, this will be only the start-nodes. Here the start-node represents a receptor on the cell surface. In every step, the information can only be passed on to proteins that can be directly reached in one step. These are represented as the set $Q$ of direct neighbors of $P$. In the description of the algorithm in pseudo-code (see Table 1), the nodes of the pathway are denoted as $P_N$ while the pathway edges $P_E$. Similarly, the nodes in $Q$ directly adjacent to the pathway are denoted as $Q_N$ while all edges leading from $P \rightarrow Q$ are denoted as $Q_E$.

From all the arcs $P \rightarrow Q$ that could extend the Pathway now the strongest one is chosen (Table 1, line 7), and the newly reached node is moved from $Q$ to $P$ (Table 1, line 8-10), hence modeling the information transfer along the most trusted connection available at this step. Subsequently, $Q$ is updated by adding all the neighbors of this new node to it (Table 1, line 17-18).
In order to forward the information deeper into the network, all edges that lead backwards from \( Q \) onto a known node in \( P \) are eliminated (Table 1 line 20-21). This simple mechanism ensures the directionality of the information flow. Algorithmically speaking, I grow a maximum spanning tree from the start-node into the network.

The procedure terminates if a node within a set of designated terminal nodes (stop-list) is hit (Table 1 line 11-14). In this work, this stop-list represents all known transcription factors triggering the cellular response in the nucleus. The pathway can now be reconstructed by backtracking along all the used arcs back to the start-node (receptor). In fact the pathway finding can be repeated until a given number of transcription factors have been reached. The result of this procedure is what I call the backbone pathway.

### 4.3.7 Adding associated Factors

Because the backbone gives the rough direction of the information flow, without claiming to be the pathway along which the information necessarily travels in vivo, proteins closely associated with the backbone pathway are added in a second pass. The 'first shell' already contains a significant number of nodes. In order to keep the results concise, not all of the (up to several hundred) proteins that have an association with a member of the backbone are displayed. For all adjacent nodes the sum of co-citations with the backbone, \( S_c \), is computed. This yields an overall average value of co-citation \( A_c \), and only the proteins where \( S_c \geq (A_c/\text{widthfactor}) \) are kept. In order to keep only the essential associations (edges), then for every adjacent node its average strength of association with the backbone is determined. By dropping the associations below average, every associated protein is linked to the backbone only via those edges that have an above average weight among all the associations of this protein to the backbone.

### 4.3.8 Parameters

As described above, the algorithm is deterministic given the weighted interaction graph and essentially parameter-free. The amount of output is influenced by two parameters, which control the search depth and search width. The search depth determines how many entries in the stop-list (transcription factors) have to be reached before the algorithm terminates. The benefit of this approach is that there is no need to make any assumptions about the
findPathway(Graph $G = (V,E)$, startNodes $S$, TerminalNodes $T$, int searchDepth)

01. nrOfHits $= 0$
02. $P_N = S$
03. $P_E = \emptyset$
04. $Q_N = \{t : t \in (V - S), s \in S, (s \rightarrow t) \in E\}$
05. $Q_E = \{e = (s \rightarrow t) : s \in S, e \in E\}$
06. while ($Q_E \neq \emptyset$ and nrOfHits $\leq$ searchDepth)
07. \hspace{1em} $e_{max} = e : e.weight \geq f.weight, e \in Q_E, f \in Q_E$
08. \hspace{1em} $P_E += e_{max} = (u \rightarrow v) : u \in P_N, v \in Q_N$
09. \hspace{1em} $P_N += v$
10. \hspace{1em} $Q_N -= v$
11. \hspace{2em} if ($v \in T$) // new node v is a terminalNode
12. \hspace{3em} nrOfHits += 1
13. \hspace{3em} excludeNeighbours($v$)
14. \hspace{3em} backTrack($v$, $S$)
15. \hspace{2em} else
16. \hspace{3em} // Update Q
17. \hspace{4em} $Q_N += \{w : (v \leftrightarrow w) \in E, w \in V - (P + Q)\}$
18. \hspace{4em} $Q_E += \{e = (v \leftrightarrow w) \in E, w \in V - (P + Q)\}$
19. \hspace{2em} end if
20. \hspace{1em} $B = \{(v \rightarrow r) : r \in P + Q, (v \rightarrow r) \in E\}$
21. \hspace{1em} delete($B$) // remove the backward edges
22. end while

end findPathway

Table 4.2: Pathway-Finding Algorithm in Pseudo-Code
maximal length of a pathway. A different search depth does not alter the trajectory of the backbone. The effect is merely that a branch is added to the backbone pathway for every additional node in the stop-list.

Of course the effect of the actual value of the search depth on the resulting pathway-figure is dependent on the network context of a given start-node, but in biological terms it can be roughly summarized as follows: When stopping on the first or second transcription factor, the result usually is a very basic signaling cascade. When stopping after the third transcription factor has been reached, the graph also contains the most directly reached downstream process. When five or more transcription factors are included, alternative routes and feedback loops via closely associated proteins emerge. Therefore the search depth is an easy way of controlling the amount of information to be extracted for a given receptor.

The search width controls the fraction of associated factors which are kept as 'flesh' around the backbone. A threshold of 1 means that only nodes which are equal or above average connected to the backbone pathway are kept. If this threshold is lowered, then a higher number of nodes are kept. Dividing the average connectivity of all associated factors by the width factor yields the threshold that is used. For example, a search width of 2 will keep more (roughly about twice as many) associated nodes, while a factor of 0.5 will result in a less dense graph. The advantage is that there are no assumptions to be made about the absolute number of proteins to be included. The search width controls the number of proteins being kept with respect to the average connectivity of all associated factors to the backbone. This implements a local adaptive cut-off in the sense that this average is different for every local pathway within the global network. Similarly, the associated factors are connected to the backbone only via their strongest edges, as all edges with a weight below average for every associated factor are dropped.

4.3.9  Visualization

After the main pathway has been calculated (by merging all the traces of the backtracking from all reached transcription-factors and their associated factors) the resulting graph is exported for visualization. In the current implementation, the graph is exported in Graph Description Language (GDL) and displayed using aise (http://www.absint.de). In aise the force-directed layout was used to produce the final figures.
Figure 4.2: Schematic representation of the pheromone response pathway in yeast, adapted from [29]. The order of the signaling cascade is reliably recovered by the pathway-reconstruction algorithm (see Figure 4.1).
Figure 4.3: Graph representation of the Ste2p (pheromone response) pathway reconstruction. On the left the 'backbone' is shown (width-factor of 0.0002), while on the right the same pathway is shown with associated proteins (width-factor 0.2). The search depth is set to 3, resulting in the inclusion of three transcription factors (Ste12p, Swi4p, Swi6p). All nodes are coloured according to available sub-cellular localisation data, according to the scheme in Figure 4.5. The graph corresponds well to the known structure of the Ste2p pathway (shown in Figure 4.6).
4.4 Results

I have developed an algorithm that extracts signal transduction pathways from Medline abstracts. The method is based on modeling information flow along trusted associations. The basic idea here is to generate a backbone along the most trusted edges of the global interaction network that could mediate an external signal and trigger a cellular response. The algorithm does not involve any clustering steps or pre-filtering of the input data. The implementation of the principles of information flow results in a method that works un-supervised, without prior training, and in a completely automated way. The fast and robust greedy strategy applied here makes the method well suited for an interactive environment in which researchers can explore and navigate all possible signaling pathways. I call this superset of signaling pathways the transductome.

Because S. cerevisiae is comparatively well characterized, I focus on this model organism in this analysis. However, since this method is based on pre-computed statistical associations derived from biomedical literature (Medline abstracts), it can be applied to any organism as long as there is a body of literature available. The results of this work are two-fold:

- Firstly, from the biological perspective, the pathway-finding algorithm produces an overview over a large body of biological literature in an easily understandable graphical format. In the absence of any clear-cut definition of pathways or well defined signaling pathways, it remains difficult to quantify accuracy here. Hence I had to rely on textbook knowledge and expert opinion in order to get an idea about the usefulness of the output.

- Secondly, from a computational perspective, the algorithm developed here is of general interest since it can be applied to data derived for any other organism and even networks built using different experimental data or computational techniques. The complete calculation and visualization are done within less than a minute on a standard desktop PC. The algorithm is very tolerant to noise in the input data due to the dynamic filtering that resorts to weak associations only if no stronger links are left.
4.4. RESULTS

4.4.1 Examples of Computed Pathways

Figures illustrate results obtained for three different pathways. The reconstruction of the pheromone response pathway (compare Figure 2 and Figure 3) works very well. Most of the key players of the signaling cascade are identified and the ordering of the proteins on the backbone reflects the established biological model of the cascade (Figure 4).

The pheromone response pathway in the yeast Saccharomyces cerevisiae is a well studied example of a signaling pathway involving a G-protein and a mitogen activated protein kinase (MAP kinase) cascade. It consists of a membrane bound pheromone receptor Ste2p, which links to a hetero-trimeric guanine-nucleotide-binding protein (G-protein) (Gpa1p, Ste4p, Ste18p). The subsequent signal propagation is via a MAP kinase cascade of Ste20p, Ste11p, Ste7p, and Fus3p or Kss1p, which are bound to the scaffolding protein Ste5p and regulate the transcription factor Ste12p. The pheromone signal triggers the onset of the mating process which involves complex changes in cell cycle (Far1p, Chl2p, Chl1p, Clb5p, Dafl1p, Swidp, Swifp, Cdc28p) and cell wall integrity (Figure 5 and Figure 6). All of the proteins named above are recovered as part of the backbone pathway or associated factors, and also includes three additional factors (Fus1p, Mata1p, Cdc42p).

The second example pathway, starting from Mid2p, is required for activation of the cell wall integrity signaling pathway. It shows a considerable crosstalk with the pheromone pathway, merging at Ste11p (Figure 7). The cell wall integrity signalling pathway consists of the small G-protein Rho1, protein kinase C (Pkc1p), and a mitogen-activated protein kinase cascade Bck1p and Slt2p, which appear as associated factors in the computed pathway, are components of this kinase cascade. Rho1p and Pkc1p are found in the computed backbone pathway.

The third pathway starts from GPR1p, another G-protein coupled receptor (Figure 8). Gpr1p is coupled to Gpa2p and involved in the pathway of pseudo-hyphal differentiation via a cAMP-dependent protein A kinase pathway (Cyr1p-Tpk1p/Tpk2p/Tpk3p). Tpk3p cannot be found in the backbone but it is present in the ‘flesh’ linked to Ras2p and Sra1p. Ras2p also regulates the TPK pathway besides others. Ras2p is a quite prominent homologue of the human RAS protein, this might be one of the reasons why Ras2p is not predominantly linked to the TPK-cascade.
Figure 4.4: Graph representation of the MID2 pathway reconstruction. The 'backbone' on the left is calculated with a width-factor of 0.0002, while on the right the same pathway is shown with associated proteins (width-factor 0.15). Here the search depth has been set to 2, resulting in the inclusion of the transcription factors Swi4p and Swi6p. All nodes are coloured according to available sub-cellular localization data, according to the scheme in Figure
Figure 4.5: GPR1 Pathway reconstruction. Again the 'backbone' (left) was calculated using a width-factor of 0.0002, while on the right the same pathway is shown with a width-factor of 0.3. The search depth here is set again to 2, yielding Mig1p and Cyc8p as the first downstream transcription factors found. All nodes are coloured according to available sub-cellular localization data, according to the scheme in Figure 4.4.
4.4. RESULTS

Subcellular Localisation
- Unknown
- Cytoplasm
- Cell neck
- Cell membrane
- E.R.
- Mitochondrial
- Nuclear
- Nuclear Rim

Figure 4.6: Colour Scheme coding for sub-cellular Localization: All the nodes (proteins) in Figure 4.5 were colour-coded according to this scheme.

The edge thickness in all examples is drawn according to the strength of the statistical associations between two proteins (number of co-citations). The backbone edges are drawn in black when there is experimental evidence confirming an association as a physical interaction, and coloured red if it is a statistical association only. Similarly, the edges connecting the associated factors with the backbone are shown in green when they are experimentally confirmed and coloured in blue otherwise. All proteins on the backbone except the transcription factors are shown as diamond shaped nodes, with the receptor drawn at an increased size. The associated proteins are depicted as circles with the transcription factors being enlarged.

Based on looking at many computed pathways, I find that in order to estimate the reliability of the predictions it is helpful to take the thickness of the edges into account, which reflects the number of co-citations which has been found for the respective protein pair. A further indicator of the reliability of a pathway is the amount of ‘flesh’ added to the backbone, ‘bare’ backbones at high search widths usually indicate less reliable connections.

4.4.2 Validating Statistical Associations using Experimental Datasets

The implementation of the algorithm presented here is solely based on statistical associations among proteins in S. cerevisiae computed from Medline abstracts. In the context of this analysis the experimental datasets were not
Figure 4.7: Network reliability at different Levels of Mutual Information. The solid circles denote proteins, while the dotted lines represent experimentally confirmed interactions. The solid line (ski8-spo11) is a potentially novel interaction.

Here the Cellzone dataset was used as a reference standard. The inner dotted circle contains interactions with an MI value greater than 0.0038. At this level the baseline accuracy of the network is about 38% reliability. Including associations with a lower but still positive (MI ≥ 0) mutual information increases coverage while the accuracy drops down to 28%.

used in the computation but were utilised in the visualization stage in order to demonstrate which of the statistical associations overlap with known protein-protein interactions.

A significant fraction of the statistical associations overlap with experimentally determined interactions. For example, the baseline accuracy of the text-derived associations is in a similar range as the yeast-two-hybrid datasets compared to the Cellzone data as a reference set (for comparison, see Figure 4.7). Furthermore, the relative percentage of statistical associations with experimental confirmation increases within subsets of increasingly stronger associations (see Figure 4.7). Although I do not claim that all of the associations represent direct physical interactions, this supports the assumption that a functional relationship between proteins can be inferred from their co-occurrence.
4.5 Discussion

I have described an algorithm based on first principles of information flow that finds signal-transduction pathways along trusted edges in a protein-protein interaction network. The main contribution is the ability to create an overview of hundreds of Medline abstracts by condensing them into a graphical format that can be readily understood and interpreted. Although each individual association or interaction has been described in detail in the corresponding articles, it is the context and overview that the algorithm provides that are novel. This helps researchers to look beyond the boundaries of a single protein and consider the surrounding functional context to generate testable hypotheses. The power of this "network thinking" comes from bridging individual biochemical function of a protein to the higher level of cellular function, which is here operationalized by network context.

4.5.1 Advantages

Since the algorithm implements a greedy strategy, it is robust, fast and delivers reproducible results. Empirically, it seems to get the topological order of the signaling cascades roughly right. A major advantage facilitating further studies of the predicted pathways is that every association in the graph is linked to the underlying articles. This allows for rapid and detailed manual inspection of the underlying text-body as well as for a direct evaluation of the graph in the context of the available literature. The calculation can start from a whole set of start-nodes and consider a large set of terminal nodes (i.e. all known transcription factors). This enables the inclusion of known intermediates for a given pathway as well as the investigation of the cross-talk between different signaling pathways. The algorithm copes with noise and the small-world characteristics of the network without any pre-processing of the data, like removal of edges below a certain threshold or elimination of highly connected hubs (compare, for example, to...)

Furthermore, the algorithm is free of any assumptions on the length of a pathway, the absolute number of nodes to be incorporated in the pathway, or about any intermediates. Nevertheless, known intermediates can be easily incorporated into the search by including them in the start-set.
4.5.2 Accuracy

Experimental data was not used as input to the algorithm, so it can be used for validation. The percentage of associations verified by protein-protein interactions is higher within the pathways than the average in the full text-derived interaction network. However, this results from the combination of two factors: Firstly, stronger statistical associations in general have a higher overlap with the experimental data (see Figure) and secondly the algorithm is designed to prefer strong associations whenever possible. Since I did not have any pre-defined reference set of signaling pathways available, I had to rely on the feedback from biologists, which was astonishingly positive. The pathways were usually directly recognized by experts and regarded as useful, since many of the proteins in a graph are known to be involved in the signaling pathways they represent. Furthermore, the ability to link known pathways with other downstream processes can directly lead to novel hypotheses about cellular function.

The significance of the resulting pathways can be roughly estimated as follows: The backbone pathways presented here contain about 10 nodes. As the dataset contained 41129 associations for 3616 proteins I can assume an average of at least 10 adjacent edges for each node in the network. This means there are $10^{10}$ different possible pathways of length 10 for any given start-node. Given the many known factors of the pathways presented here that are recovered by the algorithm clearly shows that the results are far from random. Although this estimate does not constitute a statistical model of significance, this gives an idea of the complexity of the overall problem. Obviously, enumerating and evaluating all pathways with a given scoring scheme is computationally expensive. This highlights the advantages of a reliable greedy algorithm based on local decisions.

4.5.3 Limitations

Since the algorithm presented here is based on pre-computed statistical associations from Medline abstracts, it does not find really new associations or interactions. The scope is hence limited, by definition, to the published data (abstracts) and does not cover hypothetical genes or proteins. I have shown that the model of intra-cellular flow of information works well for signal transduction. However, the same paradigm may not be appropriate for metabolic pathways which would have to model the flux of metabolites.
4.5.4 Extensions

The low computational demands of the algorithm make this method suitable for an interactive environment which allows for exploration and iterative refinement of a pathway from a given set of receptors. The algorithm can work with any set of association or interaction data as long as there is an objective function available that delivers a numerical index of association strength. Since the crucial step is which edge of the neighborhood of the pathway is explored next, the algorithm can work with an integrated set of experimentally determined interactions, as long as there is a reliable way of comparing the relative strength of the interactions from different sources and experimental methods.

Although some measure of reliability is available within most datasets (i.e. the number of IST hits in the yeast-two-hybrid data), so far these measures are incompatible towards each other. Attempts to overcome this can be found in which would enable the use of experimental data directly within the framework outlined here. Another possibility would be to overlay experimental protein-protein interaction networks with measures of gene co-expression in order to generate a weighted graph. The absence of a general, weighted measure of reliability for the interactions resulting from high-throughput experiments is a limit of the available data but not a shortcoming of this method. In the absence of large scale data on the physico-chemical properties of protein-protein interactions, the measure of co-occurrence was used here as a (rather crude) approximation of binding strength and specificity.

4.5.5 Philosophical Remarks

The observation that putting together the relatively simple principles of information flow yields meaningful pathways is a result in itself. It implies that, even in the absence of a clear-cut definition of what makes a signaling pathway, this pathway-finding algorithm is able to capture to some extent the way information is processed in the cell. Alternatively, it could be argued that it captures the way pathways are described in the biological literature. Then this method is just able to extract pathways in a similar way as we cope with the complexity of biological networks. The question of how far our reasoning about biology in terms of pathways meets the biological reality must remain open here.
There is an interesting relationship of this method to the insights about path finding in scale-free networks. It has been shown that forwarding packets to the neighboring node with the highest connectivity results in good coverage of the network and fast transfer of information. However, in the case of signal transduction pathways one is not interested in fast forwarding of the packets of information, but in reliable information transmission. Hence, here the strongest (or most trusted) connections is used first, since the information is valuable and has to be passed on to the ‘right’ nodes. Obviously, fast and indiscriminate signaling into any cellular process would create havoc in a biological system. The potential to achieve good coverage in the context of scale-free networks based on local rules is explored in the pay-as-you-go strategy (see Chapter).

4.5.6 Application as an Annotation Tool

Finally, it worth noting that this method is ideally suited as an annotation tool in the context of emerging interaction databases. The ability to produce a simple graphical overview of a huge body of text with direct links from each edge to the underlying literature will ease the efforts of curators significantly.

4.6 Acknowledgments

The calculation of the statistical associations and the implementation of the text-mining was done by Dr. Sabine Dietmann, who also kindly provided Figure. Thomas Schlitt assisted in the biological analysis of the reconstructed pathways.
Chapter 5

Unraveling unknown interaction networks with near-optimal efficiency

"He tried to explain that magic had indeed once been wild and lawless, but had been tamed back in the mists of time by the Olden Ones, who had bound it to obey among other things the Law of Conservation of Reality; this demanded that the effort needed to achieve a goal should be the same regardless of the means used."

Terry Pratchett, 'The Colour of Magic'

5.1 Summary

The functional characterization of all genes and their gene products is the main challenge of the post-genomic era. Providing interaction information for every gene-product is a clean way to assemble the jigsaw puzzle of proteins into a functional map. I demonstrate that the information gain is determined by the strategy applied to order the baits of pull-down experiments. Due to the scale-free distribution, fast network coverage is obtained by focusing initially on the hubs in the network. Unfortunately, locating hubs requires prior global information about the network one is trying to unravel. I describe a novel pay-as-you-go strategy that finds its way to highly connected nodes near-optimally using only local information collected on-the-fly in successive pull-down experiments.
Using this strategy, I estimate that 90% of the human interactome can be covered by 10000 pull-down experiments with 50% of the interactions confirmed in both directions.
5.2 Introduction

Here I define protein interaction space as the set of all specific interactions among proteins in the cell. Experimental high-throughput techniques such as yeast-two-hybrid and TAP-MS allow, with some degree of error, to determine the neighbourhood of a given protein within the interaction network. Although over 100 different organisms are now already sequenced and many other sequencing projects are underway, only very limited information on the interactome is available. For example, DIP currently contains only 958 interactions for human and 246 for mouse while there are 15178 interactions recorded for yeast (Saccharomyces cerevisiae). 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 complete coverage of interaction space be achieved with minimal effort in terms of the required number of experiments?

I operationalised the required resources by assuming cost and time to be in a constant proportional relationship to the number of performed experiments (pull-downs), which is equivalent to the number of proteins used as bait. I am well aware that this simplification leaves out a lot of experimental detail, 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, I used real interaction data sets (for yeast) which were explored from scratch using virtual PullDown experiments (see Table). Although no complete dataset of interactions is available for a single organism yet, all observations indicate that protein interaction networks are scale-free (see Figure). Since any randomly selected subset of a scale-free network again follows a power-law distribution, and all interaction datasets available represent different subsets of the overall interactome, I concluded that interaction space as a whole has the same distribution as any major subset. Thus, the simulation results on incomplete interactomes should hold also for the real-life exploration of unknown interaction networks.

5.2.1 Covering Interaction Space

I model interaction networks as graphs where all proteins are nodes linked by edges representing the protein interactions. Thus, coverage of interaction space translates to edge coverage in graphs (see Figure). In contrast,
Figure 5.1: Two edge-covering sets \( (C_1, C_2) \) for the same graph. The graph \( G = (V, E) \) consists of 9 nodes \( (V = \{a, b, c, d, e, f, g, h, i\}) \) linked by 10 undirected edges \( (E) \). Edge-covering sets are indicated by the cone-shaded nodes. Both set \( C_1 \) (left) with 6 nodes \( (\{a, b, d, e, f, h\}) \) and set \( C_2 \) (right) with 3 nodes \( (\{c, f, h\}) \) satisfy the condition that every edge is adjacent to a node in \( C_1 \) or \( C_2 \), respectively. In the adjacency matrix representation of \( G \) an entry \( m_{ij} \) is marked (in red) to denote an interaction between proteins \( i \) and \( j \). The matrix \( M_{C_2} \) shows that \( C_2 = \{\{c, f, h\}\} \) is a covering set, since the respective rows and columns (grey) of this set cover all interactions. The matrix \( M_{ord} \) illustrates how the weight of a graph is calculated from any given ordering of the nodes. In this example the virtualPullDown experiments (Table) are conducted such that 1st:c, 2nd:f, 3rd:h, 4th:a etc. are used as bait. The ordering of the nodes is marked on the diagonal of \( M_{ord} \). The upper right triangle of \( M_{ord} \) denotes at what time-step an interaction is seen while the lower left triangle of the matrix records at what time-step the interaction is confirmed according to the ordering. A performance plot is generated by plotting the number of edges seen and confirmed after time-step \( t \). The earlier edges are seen and confirmed, the lower the sum of the weight of all entries in \( M_{ord} \) becomes. Hence this time-weight (the sum over all entries in \( M_{ord} \)) can be used as an indicator for the performance of an ordering in revealing the topology of the network.
5.2. INTRODUCTION

structural genomics aims to provide enough experimental structures such that for every protein sequence a homologous structure can be found (Vitkup et al., 2001), in other words, to provide a node covering set of experimental structures within a network of homology relationships among sequences. There are many different solutions for finding a set of nodes covering all edges (interactions) within the same graph (interaction network). As illustrated in Figure 5.1, a subset of highly connected nodes covers a larger portion of the network than another subset of the same size consisting of nodes which 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 mapping of the interactome with the minimal experimental effort (minimal number of baits). The bad news here is that the problem of finding the minimum covering set of nodes 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. In this work, I will open the door to a new breed of methods and strategies which use the properties of scale-free networks to their advantage and are able to tackle otherwise computationally hard problems effectively by computationally relatively simple means.
5.2.2 The Model: Virtual Pull-Down Experiments

Each pull-down experiment reveals the interaction partners of one protein used as bait. In order to model the information gain by an individual experiment (tandem affinity or yeast-two-hybrid), I introduced the procedure of a virtualPullDown( p, t) (see Section Table). First, protein p is marked as bait in time-point t. Subsequently, all direct neighbours of p (in an experimentally determined interaction graph) are marked as prey if they have not yet been detected as a prey previously from another bait. Similarly, all the edges adjacent to p are marked as "detected at time-point t" if they have not been detected before and "confirmed at time-point t" otherwise (see Figure). The latter corresponds to what has been termed 'reverse-tagging' in TAP/MS experiments. Since every undirected edge links exactly two nodes, the complete process of covering an interactome is modeled with these two additional attributes, seen and confirmed. Given the inherent error-rates of the available experimental techniques the goal is to confirm as many edges as possible in order to arrive at a reliable set of interactions. By using the time-weight as an objective function I could distinguish the efficient strategies.

Assuming a given ordering of the all the n proteins (nodes) from [1..n], I ran successive virtualPullDown-experiments (Table) of every protein at time-point t = [1..n] and measured how quickly a given portion of the edges had been detected or confirmed. The time-weight is lower the more edges are being covered earlier in the process (Figure). Hence this time-weight indicated how quickly the interactions of the network were being discovered by the given strategy. This enabled me to compare different orderings in terms of the time-weight of a graph by taking the sum of all time-points across all edges. This model allows me to answer different questions, for example:

• Given there is funding (or time) for x number of experiments, how large is the fraction of the interactome one can expect to detect?

• How many experiments are needed to confirm a certain fraction (e.g. 50%) of the interactome?
5.3 Methods

5.3.1 Datasets

For simulation purposes, I focused in this analysis on the model organism *Saccharomyces cerevisiae* (bakers yeast) only, since for this species the most comprehensive data about protein-protein interactions is available. In detail, the analysed datasets are:

- **DIP** (Database of Interacting Proteins) is a resource that has gathered information by manual annotation and also includes the high-throughput yeast-two-hybrid datasets. Excluding self-interactions, this set contains 14844 interactions for 4711 proteins in yeast.

- **CORE** is a subset of DIP which contains validated interactions. Hence the interactions contained in this set are more reliable. Consequently, this dataset is smaller and less dense; it contains only 4357 interactions for 2129 yeast proteins.

- **CZ** is a dataset of 2743 interactions determined for 1297 proteins using tandem affinity purification and subsequent characterisation by mass-spectrometry.

- **MDS** follows a similar approach but over-expressed the bait proteins. This may explain why this dataset is denser than the CZ network; it contains 8040 interactions for 1695 proteins.

For the simulations on the CORE dataset, the actual simulation was done within the DIP network. The number of edges seen and confirmed was then determined only within the CORE subset of validated interactions. This analysis addresses the concern that by targeting highly connected proteins first, the resulting information would be less reliable, which is clearly not the case (see Figure).

For the sake of simplicity, I assumed reflexivity by using an undirected graph for modelling interaction space. This means an interaction $e = (u \leftrightarrow v) \in E$ is equally likely to be detected from both interacting proteins. Hence $(u \rightarrow v)$ and $(v \rightarrow u)$ have the same likelihood to occur in a virtual PullDown experiment (Table), using proteins $u$ or $v$ as bait, respectively.
virtualPullDown( Protein p, TimePoint t)
1. p.bait = t
2. N = getNeighbours( p, f_p, f_n)
3. FOR ALL (n ∈ N)
4. IF ( n.prey == 0) n.prey = t
5. IF ( (p ↔ n).seen == 0) (p ↔ n).seen = t
6. ELSE (p ↔ n).confirmed = t
7. NEXT n

Table 5.1: The Procedure of a virtualPullDown in Pseudo-Code

This has been argued to be not necessarily true for both TAP/MS and Y2H data. The analysis methods could accommodate for this by using a directed, weighted graph instead where the weights on the edges represent the probabilities of detecting a protein along this relationship. However, so far such data is not yet available.

5.3.2 Procedure of a virtualPullDown( p, t)

The virtualPullDown procedure (see Table 5.1) models the information gain from a pull-down experiment with protein p as bait in time-point t. All nodes u ∈ V in the graph G = (V, E) representing the protein-interaction network have the two additional attributes u.bait and u.prey containing the time-points when this protein was used as bait and when it was first detected as prey. All edges e = (u ↔ v) ∈ E of this graph have the additional attributes (u ↔ v).seen and (u ↔ v).confirmed to denote when this interaction has been detected and confirmed. All these attributes are initially set to zero.

The function getNeighbours() delivers the set of all nodes adjacent to the bait protein p. Any rate of false positive (f_p) or false negative (f_n) data can be accounted for by the getNeighbours() function (see Table 5.1). First a fraction f_p of the neighbouring set is replaced by randomly selected nodes, then a fraction of f_n of the nodes in the set are deleted at random. The resulting set N of neighbours is delivered back to the virtualPullDown() function, which then operates on this erroneous set.
getNeighbours( Protein p, FalsePositiveRate $f_p$, FalseNegativeRate $f_n$)
1. $N = \{ n : (n \leftrightarrow p) \in E \}$
2. FOR ALL ($n \in N$)
3. IF (nextRandomNumber() < $f_p$) $n = \text{selectRandomNode}()$
4. IF (nextRandomNumber() < $f_n$) $n.\text{delete}()$
5. NEXT $n$
6. return $N$

Table 5.2: The Procedure getNeighbours in Pseudo-Code

5.3.3 Implementation of Search-Strategies

I compare the performance of different strategies to order the baits. The procedure of a virtualPullDown is identical for all the strategies analysed. The strategies merely differ in the way the next bait protein is chosen from the set of remaining nodes that have not been used as bait yet.

- random: A bait is picked at random from the set of proteins that have not been used as a bait before

- greedy seen/confirmed: Given the overall topology of the network, for every node the number of adjacent edges not seen yet ($k1$) and the number of edges seen ($k2$) can be determined at any stage of coverage. By selecting the node with the highest $k1$ this strategy maximises the number of edges seen, while selecting the node with the highest $k2$ maximises the number of edges confirmed. Note that 'seen' means incoming edges to prey-proteins. In both cases, the process starts with the most highly connected node. In the greedy seen strategy there are no more edges to detect after about 30-40% of nodes have been used as bait. After this, the greedy seen strategy resorts to optimise $k2$.

- degree guided: All proteins are sorted according to their number of interactions in descending order. The nodes are used as bait in the resulting order.

- pay-as-you-go: The next bait is determined by dynamically estimating the degree of each node based on what is known about the network from experiments carried out so far. Details follow.
5.3.4 Implementation of the pay-as-you-go Strategy

The graph-theoretic principles of the pay-as-you-go strategy are based on the observation that during the entire screening process every protein (node) falls into one of the following three subsets:

- **P**: the proteins \( p \in P \) have been used as bait. For this subset of proteins the identity of their interacting partners (prey) and hence their degree \( k(p) \) is known from experiments.

- **Q**: the proteins \( q \in Q \) have been detected in experiments as prey and were not used as bait yet. Although the exact number of interactions (degree, \( k(q) \)) of these proteins is unknown at this point, it is known how often each of these proteins occurred as prey so far (\( \text{indegree} = k_{in}(q) : q \in Q \)). This forms a lower bound on the degree such that \( k_{in}(q) \leq k(q) \).

- **R**: the rest of the proteome where no interaction information is available yet.

In the beginning when there is no information about the network available, \( P \) and \( Q \) are empty and all proteins are contained in \( R \). As more and more interactions are revealed in successive experiments, \( P \) and \( Q \) contain ever larger fractions of the network. Finally, after every protein has been used as bait, \( P \) contains all proteins while \( Q \) and \( R \) are empty.

The bait-prey relationships resulting from the experiments can be modelled as directed edges leading from \( P \) into \( Q \). So the information derived so far is stored as a directed graph. While there are no edges leading into \( R \), for all the prey-proteins contained in \( Q \) the number of times they have been detected is known (see Fig. 3). The next bait is determined by selecting the node with the maximum indegree (=number of times this protein was detected as prey) and the maximum \( \Delta K \) (see below).

At any stage, the average indegree \( \langle K_Q \rangle \) of all proteins contained in \( Q \) can be calculated. If there is no protein \( q \in Q \) in with an indegree \( k_{in}(q) > \langle K_Q \rangle \), then the strategy resorts to choosing the next bait at random. The condition \( \text{indegree} \leq \langle K_Q \rangle \) for choosing the next bait at random is satisfied towards reaching full coverage of the network and in the initial stage. Since at the start \( Q \) is empty and \( \langle K_Q \rangle = 0 \) the whole process is automatically kick-started this way with a few randomly chosen virtualPullDown experiments.
Figure 5.2: Illustration of the subsets $P$, $Q$, and $R$ during the screening process. The proteins used as bait are denoted as $P = \{p_1, p_4\}$. The proteins detected as prey at this point are in the set $Q = \{q_1, q_4\}$. In this example (left) $q_2$ is selected as the next bait as it has the highest *indegree*. The situation resulting from using $q_2$ as bait in the next virtualPullDown (Table 5.1) is depicted on the right. Now $q_2$ becomes a member of $P$ and is denoted as $p_5$. All its previous incoming interactions are now *confirmed* while two new prey proteins ($q_5, q_6$) are detected and inserted into $Q$. In this situation, where 2 nodes ($q_i, q_j$) have the same *indegree*, the additional measure of $\Delta K$ is employed to break the tie.
5.3. METHODS

Figure 5.3: The anti-correlation of the degree of interacting proteins. Similar to the $P(k)$ distribution in Figure 5.2, the x-axis denotes the number of interactions (degree) of a protein. On the y-axis the average degree of the neighbours of all proteins with degree $x$ is plotted. This illustrates that proteins with few interactions tend to have highly connected interacting partners, while hubs tend to interact with many less connected proteins. The fact that the slopes of the linear trendlines are very similar for all datasets indicates the generality of this observation. The different offset of the trendlines corresponds to the different density of the datasets analysed (see Figure 5.2).
5.4. RESULTS

The number of times a protein \( q \) has been seen as prey \( (k_m(q)) \) is an integer number. Hence at any given stage there can be several proteins in \( Q \) with the same maximum \( \text{indegree} \). In order to break the tie, the node with the highest value of \( \Delta K \) is chosen. Both \( \text{indegree} \) and \( \Delta K \) are initially set to 0 for all nodes. \( \Delta K \), a 'distributed' weight-factor that is, is defined as follows. Every time a pull-down is performed the number of neighbours (or \( \text{degree} \) \( k \) of the bait protein is being determined as well. This degree is 'distributed' over all its' prey proteins by adding \( 1/k \) to the respective \( \Delta K \) of every prey, while the \( \text{indegree} \) of every prey protein is increased by 1.

The rationale behind this comes from the observation that hubs are less likely to be linked to other hubs (see Figure: hubs). As an additional local measure I developed \( \Delta K \) such that prey proteins interacting with less connected proteins get higher \( \Delta K \) values than prey proteins which are connected to hubs. So \( \Delta K \) is an empirical measure that guides the strategy towards choosing baits which are linked to lowly connected proteins and hence these baits are more likely to be hubs. It is noteworthy that \( \Delta K \) as a single indicator of degree without the \( \text{indegree} \) also performs significantly better than random. However, the combination of \( \text{indegree} \) with \( \Delta K \) as described above is the most efficient local strategy I have found so far.

5.4 Results

The main idea is to measure how different orderings could have revealed the interactions contained in various available datasets in the most efficient manner. There are two ways to define edge-coverage:

1. an interaction has been seen or detected once by using one of its interactors as bait, or
2. an interaction has been confirmed or detected in two directions after using both interactors as bait.

The latter definition directly addresses the problem that current experimental datasets contain a significant number of false positives and false negatives, such that an interaction is usually not regarded as safe knowledge unless it has been detected at least twice in independent experiments. The performance of different strategies in seeing (Figure: seeing) and confirming interactions (Figure: confirming) was measured on four different datasets, namely DIP, CORE, CZ and MDS (see Section: datasets for a more detailed description of the strategies and datasets used).
Figure 5.4: Performance of different strategies in seeing interactions.
Network coverage in terms of edges seen achieved by different strategies on the four different data-sets CORE, DIP, CZ and MDS. The fraction of the interaction network seen is plotted over the fraction of the proteins used as bait in the virtualPullDown experiments (Table). Each plot contains the performance curves of the following different strategies: greedy seen, greedy confirmed, degree guided, pay-as-you-go and random. The average of 10 independent simulations is shown for each strategy, together with the maximum and minimum deviation. An upper limit is formed by the greedy seen strategy, which uses the information on the global network topology. The degree-guided strategy roughly realises the so-called ‘80/20’ rule by covering about 80% of the interaction network using less than 20% of the proteins as bait. The pay-as-you-go strategy matches the performance of the greedy confirmed strategy without prior information on the network topology. At the same, the pay-as-you-go strategy performs significantly better than random.
Figure 5.5: Relative performance advantage in *edges seen* vs. *random* measured on MDS, DIP, CZ and CORE dataset. The relative performance advantage of *pay-as-you-go* vs. *random* plotted over the fraction of nodes used as bait (right) demonstrates the huge initial coverage provided by the *pay-as-you-go* strategy. Plotting the ratio of baits required in both strategies in order to reach a certain fraction of edges seen (right) shows a speed-up of up to a factor of 9 in the MDS data. Up to 80% coverage is reached by the *pay-as-you-go* strategy about 1.8 times faster on the CZ data, 2.5-3 times faster on the DIP and CORE data, and 4 times faster on the MDS network. Overall, the performance of the *pay-as-you-go* strategy seems to improve with increasing density (=number of edges/number of nodes) of the interaction networks, being CORE:2.046, CZ:2.115, DIP:3.150, MDS:4.743.
5.4. RESULTS

5.4.1 Using Global Information to order Baits

The time-weight efficiency depends on the order in which the baits are utilized (Figure [1]). A lower limit on the usefulness of all proposed strategies is defined by a random ordering of baits. Using the random strategy, the same average number of interactions is being seen at each step, corresponding roughly to the density (=number of interactions / number of proteins) of the network (Figure [2]). At the same time, edges are being confirmed by this strategy (Figure [3]) in all datasets at a rate quadratic to the fraction of nodes used as bait. This results from the increased probability of confirming an interaction by chance as an increasing fraction of edges has already been detected.

An upper limit on the performance of any strategy was determined by the greedy strategies, which assume complete information about the network topology. At each time step $t$, the bait that will yield the highest number of edges either seen or confirmed is calculated, based on the known topology. The greedy seen strategy already reaches 100% of edges seen with as few as 30% of the nodes used as baits. At the same time, the greedy strategies are significantly better than random in confirming interactions.

Following the insight that the highly connected nodes (hubs) are the most useful baits, the prior information requirements can be reduced to information solely about the degree of each node (i.e., the total number of interactors of each protein, but not knowing what the partners are). The degree guided strategy sorts the nodes by their number of adjacent edges, attacking the highly interacting proteins (hubs) first. In contrast to the greedy strategies, it does not take the exact network topology into account but blindly follows a pre-determined ordering. This method obtains the same amount of information about the network topology/connectivity several times faster than the random strategy. It does not perform quite as well as a greedy seen or greedy confirmed in seeing or confirming interactions, respectively. Still, the degree guided strategy yields an efficient approximation to the minimum covering set. In terms of time-course efficiency, it represents a good compromise in the apparent trade-off between seeing and confirming interactions, as it maximises the sum of the integrals of both performance curves (seen + confirmed).

At this point, the conclusion is that the question of achieving completeness in functional genomics seems to hinge on establishing a reliable ab-initio estimate of the degree of all proteins. However, up to now no such reliable
5.4. RESULTS

Figure 5.6: Network coverage in terms of confirmed interactions (=detected from both interactors) achieved by different strategies on the four different data-sets CORE, DIP, CZ and MDS.

The fraction of the interaction network confirmed is plotted over the fraction of proteins used as bait. The strategies shown here are the same as in Figure 5.3. The main point demonstrated here is that the pay-as-you-go strategy consistently performs as well as the greedy confirmed strategy in confirming interactions, even though the pay-as-you-go strategy has no prior information about the networks topology. It is noteworthy that the pay-as-you-go strategy outperforms the degree guided strategy in confirming interactions. The greedy seen strategy performs less well in terms of edges confirmed. The 'dent' in the performance curve (between 30%-40% of nodes used as bait) coincides with the point when all edges have been seen (see Figure 5.3 for comparison) by this strategy and there are only interactions left to be confirmed (see Section 5.1 for details).
5.4. RESULTS

Figure 5.7: Relative performance of in pay-as-you-go vs. random strategy in edges confirmed measured on MDS, DIP, CZ and CORE dataset. The initial coverage and performance advantage of the pay-as-you-go strategy is even more pronounced in confirming interactions. On all datasets, about 10 times more interactions are confirmed by the pay-as-you-go strategy at 20% of proteins used as bait than by the random strategy (left). The relative performance in confirming a certain fraction of the edges of the networks is plotted on the right. The pay-as-you-go strategy reaches 50% coverage in confirming interactions at least twice and up is to 4 times faster than random, while in the initial stages it is almost up to 6 to over 13 times faster.
measure seems to be available (see Section 5.4.1). Below, I present a novel strategy that is entirely different from the ones discussed above in that it requires no prior information about the network topology whatsoever. All decisions about bait selection are made on the fly as the graph is gradually revealed by successive experiments.

5.4.2 Using Local Information: pay-as-you-go Strategy

The pay-as-you-go strategy is based on the following observation: Given the scale-free distribution it is unlikely that an arbitrary node will be detected as prey several times, unless it is one of the few relatively highly connected nodes. This relationship between the number of times a protein was detected as prey (indegree) and the probable number of interactions is what the algorithm is primarily based on. Here I choose the protein that has been seen as prey most often so far as the bait for the next experiment. So by using the most recurrent prey our knowledge is extended on the fringes of the network that has been explored so far. Obviously, this procedure has to be kick-started by a few randomly selected virtualPullDown experiments (Table 5.3) (or otherwise by external information) to gain some initial information. At the same time, the strategy is insensitive to the starting points chosen at random in the initial phase. The pay-as-you-go strategy performs significantly better than random in seeing interactions (see Figure 5.5) but is slightly inferior to a degree guided strategy (Figure 5.6). However, the performance seems to increase with the density of the network used (Figure 5.7). Apart from being able to efficiently cover interaction space without prior knowledge, the real strength of the method lies in its ability to generate confirmed interaction information. The performance in confirming interactions is almost identical to the greedy confirmed and even slightly outperforms the degree guided strategy (see Figure 5.8).

5.4.3 Systematic Error Analysis

The fact that pay-as-you-go performs similarly on all analysed large-scale datasets, despite their differences in reliability and coverage, points to the general applicability of the pay-as-you-go strategy in the context of any currently available high-throughput technology. This is further substantiated by
5.4. RESULTS

Figure 5.8: The effect on systematic error on the relative performance in confirming interactions of the pay-as-you-go versus the random strategy. Here I concentrate the CORE data-set since it represents the most reliable subset of the yeast interactome. The random and pay-as-you-go strategy are compared. For practical reasons I consider only the fraction of interactions confirmed, as an interaction is usually not regarded as safe knowledge unless it has been detected at least twice independently.

Different error rates of false positive ($f_p$) and false negative ($f_n$) were simulated in the virtualPullDown experiments (Table 5.1) as follows. False positives were generated by replacing $f_p\%$ of the prey by randomly selected proteins from the CORE data. False negative errors were simulated by deleting $f_n\%$ of the set of direct neighbours of the bait, as defined in the CORE data (Table 5.1).

The plot on the left represents the performance curves of the pay-as-you-go and random strategies in confirming interactions (see Figure 5.8). For the sake of clarity, only a few selected combinations of false positive ($f_p\%$) and false negative ($f_n\%$) between 0% and 30% each are shown. It demonstrates that the shape of both performance curves stay practically identical under error, the curves are simply ’scaled down’ as less of the network is (dis)covered under increasing error rates.

For comparing the relative performance of both strategies comprehensively, the ratio of the areas under the pay-as-you-go and random performance curves is measured. The plot on the right shows the resulting overview for confirming interactions under different error conditions. All combinations of false positive and false negative error rates ranging from 0% to 90% in steps of 10 percentage-points were examined. As pay-as-you-go always performs better than random the ratio of the performance curves is above 1. In fact, the performance advantage of the pay-as-you-go strategy stays consistently around a factor of 2 over wide areas of noise. Only at very high error rates the relative performance deviates from this ratio. This deviation is due to the fact that at high error rates only a small fraction of the original network is detected. Hence small random variations give rise to large fluctuations.
a comprehensive error analysis (see Figure...). In the previously described comparative analysis of the different strategies I assumed that for any given bait all possible neighbours are detected as prey. In contrast to this assumption, here I investigated how different levels of systematic noise (rates of false positive and false negative) affect performance of the pay-as-you-go strategy.

Using the most reliable set of interactions, the CORE set, I added increasing levels of false positive and false negative information (see getNeighbours() in Figure...). Figure... shows that the relative performance of the pay-as-you-go strategy as compared to a random strategy stays constant over wide areas of noise.

In conclusion, the relative performance advantage of the pay-as-you-go strategy is not seriously affected by systematic error of the experimental technique. This opens the possibility to repeat experiments using the highly interacting proteins determined in the initial exploratory stages of the pay-as-you-go strategy in order to enhance the reliability of the resulting data even further.

### 5.4.4 Dependency on the Hubs

It has been argued, at least for the yeast two hybrid data, that a large fraction of the interactions adjacent to the hubs are due to systematic errors like 'sticky' proteins or auto-activation and hence do not represent physiologically relevant data. For example JSN1 (YJr091C) has a total of 282 interactions reported in DIP, while the validated CORE subset contains only 6 of these interactions.

This has been addressed by executing the pay-as-you-go simulations on the full DIP dataset but only counting the coverage of edges within the more reliable CORE subset. Figure... clearly show that the interactions denoted as reliable by the CORE dataset are seen and confirmed at a similar rate to the performance in the overall DIP network. Obviously, the pay-as-you-go strategy does not have a tendency to pick out less reliable information from the DIP network.

Since the pay-as-you-go strategy uses the scale-free distribution, I analysed how resilient the method is to the removal of hubs from the data (Figure...). From the full DIP dataset all proteins with more than 80 (40, 20) interactions are 'cut down' by random deletion of their interactions until their degree matched the number of interactions from a randomly chosen node.
Figure 5.9: Comparing the performance of the pay-as-you-go strategy for the DIP data-set from which highly connected nodes (hubs) are removed. While the full data-set contains all interactions from DIP for yeast for reference, the sets CT80, CT40 and CT20 contain the original data but no hubs with more than 80, 40 or 20 interactions, respectively. This was achieved without distorting the overall degree distribution by deleting interactions of the hubs until their degree matched the degree of a randomly selected node with fewer than 80 (40, 20) interactions. As expected, the performance in edges seen drops steadily in performance as more and more of the 'broad tail' of the scale-free distribution is eliminated. The performance in terms of interactions confirmed is less affected. Because hubs tend to interact with proteins of small degree (see Figure 5.8), the removal of hubs from the network leaves less edges to be confirmed by using lowly interacting proteins as bait. This analysis demonstrates that the pay-as-you-go strategy has the ability to make efficient use of the remaining tail of the power-law distribution even if the hubs in the network are removed.
with a degree of less than 80 (40, 20) respectively. Obviously the pay-as-
you-go strategy still uses the remaining broad-tail of the distribution very
efficiently.

The pay-as-you-go strategy uses local information to select baits from the
right-hand side of the degree distribution. Even on a random network the
pay-as-you-go strategy works more efficiently than a random ordering of the
baits in confirming interactions, as it tends to use above average connected
nodes first (see Figures [Figure 5.2] and [Figure 5.3]). A scale-free distribution boosts the
performance of the method because, due to the broad tail, the right-hand side
of the distribution is much more pronounced than in a random distribution.

5.4.5 Batch Processing and Parallel Tracks

In practice usually not a single but a whole batch of experiments is set-up at
a time. Hence I analysed running pay-as-you-go simulations in batch-mode.
A batch of up to 200 nodes to be used as bait was determined at a time. Then
virtualPullDown() experiments (Table [Table 5.5]) were conducted on this batch
of baits before the next batch was selected. These batch simulations did
not have any effect on the overall performance of the pay-as-you-go strategy
(Figure [Figure 5.4]).

The pay-as-you-go strategy extends the coverage on the fringes of what
has been detected so far of the interaction network. Thus, the presence of a
number of disjoint components in the graph slows down the overall process.
One of the reasons for the weaker performance of the pay-as-you-go strategy
on the CZ dataset may be that this dataset is, albeit quite reliable, less dense
and contains several connected components which are not interlinked.

The problem of multiple components in the interaction graph is easily
overcome by starting parallel efforts in key pathways (i.e. as determined by
orthology) simultaneously. At the same time, the current datasets probably
represent an underestimate on the density of the true interactome (Sali et al.,
2005).
Figure 5.10: Analysis of the effect of the batch size on the performance. Here the relative performance of the pay-as-you-go strategy is plotted as a function of increasing batch size, ranging from 10-200 baits at a time. The relative speed of the pay-as-you-go strategy over the random strategy is measured as the ratio of the integral of their respective performance curves. Overall, the performance ratio stays virtually identical over the whole range, with a slight decrease at larger batches. The decrease is more pronounced only on the smallest dataset (CZ), where a batch of 200 baits already represents about 1/6 of the nodes.
5.5 Discussion

I have shown that completeness in functional genomics is achievable by efficient coverage of interaction space. The degree guided strategy answers in general how much potential speed-up is attainable if any reliable indicator of interaction degree for every protein is available prior to experimental screening for interactions. Such an indicator with a monotonous correlation to the number of interacting partners of a protein has been proposed on the basis of various sources such as co-expression (Mrowka et al., 2007; Woicik and Schachter, 2001), or lethality on gene knock-out (Fraser et al., 2002). However, the question of how these various sources could be integrated to yield a practical indicator of interaction degree remains open. The debate whether sequence conformation can serve as a reliable indicator is still ongoing (Fraser et al., 2002). The use of such information within the framework of the pay-as-you-go strategy might yield even better results than what I have demonstrated so far. This question remains to be investigated.

The pay-as-you-go strategy exploits the scale-free and small-world properties of protein interaction networks. 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 from the starting point. The 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 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. Given the limitations of current experimental techniques, an interaction has to be repeatedly detected (at least twice) before it can be regarded as safe knowledge. The general principle of fast coverage in scale-free networks has been described in ...

To my knowledge the algorithmic strategy proposed here is the first one to actively exploit the scale free distribution in order to tackle a biologically relevant problem. The pay-as-you-go strategy could be applied to any network with a scale-free distribution. Therefore the implications go be-
yond proteomics and functional genomics. For example, protein structures are found to form scale-free networks (Vendruscolo et al., 2005), which suggests that such an algorithm could help to improve or speed-up calculations for the prediction and comparison of protein structures.

5.5.1 Implications for Functional Genomics

Following the proposed pay-as-you-go strategy it is possible to generate an overall scaffold of interactions that covers most of the interactome with minimal experimental effort. Once a 'highway' through the interactome is established, functional characterisation of the rest of the proteome is relatively straightforward: a few interactions for every protein will allow individual research groups to connect their protein of interest to the established functional framework. Based on the performance data presented here, the pay-as-you-go strategy could detect up to 90% of the human interactome with less than a third of the proteome used as bait (see DIP data, Figure 3). This means that about 10,000 TAP-MS experiments should suffice to provide a scaffold for functional prediction for the rest of the human proteome. At the same time, our strategy would yield over 50% of the interactome as confirmed interactions (see Figure 3). A random, uncoordinated strategy would require between 2-4 times the experimental effort to yield this amount of coverage in terms of interactions detected and confirmed. As illustrated in Figures 4 and 5, the maximum impact of the strategy is realised in the initial explorative stage of covering the interactome. It follows that it is crucial for public large-scale proteomics efforts to co-ordinate the community effort from the start. This could be achieved by publishing a continuously updated list of the most promising targets with the highest projected return of interaction information. Such a list can be computed based on the methods presented here and the data collected in public repositories, such as BIND or IntAct (Orchard et al., 2005). Grafting a publicly funded central body (e.g. the HUman Proteome Organisation HUPO, http://www.hupo.org/) could distribute the workload effectively among associated research institutions.

Once the coverage has reached a turning point (i.e. after 20-30% of the proteome have been used as baits) a random strategy will be just as useful. At this stage, the rest of the proteome can be covered in an uncoordinated fashion by individual research labs.
Figure 5.11: Effect of increasing randomisation on the relative performance of the *pay-as-you-go* strategy.

This analysis was performed on the MDS dataset. The fraction of edges *seen* (left) and *confirmed* (right) is plotted over the fraction of nodes used as bait. Each curve denotes a different level of randomisation in the dataset. Between 0% and 100% of all the edges were randomized such that both interactors were replaced with randomly selected nodes from the dataset (see Figure 5.3). The performance of both the *pay-as-you-go* and the *random* strategy were measured.

The performance of the *random* strategy remains virtually identical under any level of randomisation in the dataset and was plotted just once for reference. The performance of the *pay-as-you-go* strategy drops steadily with increasing randomisation. Finally, on the completely randomized dataset, the *pay-as-you-go* performs identical to the *random* strategy in terms of edges *seen*. In edges *confirmed* the performance of the *pay-as-you-go* stays ahead of the *random* strategy because it still tends to use the above-average connected nodes first.
Figure 5.12: The Linear Relationship of Randomization with Loss of Performance of the pay-as-you-go Strategy
The integral of each performance curve of the pay-as-you-go strategy (see Figure 5.11) divided by the integral of the performance curve of the random strategy was calculated. The resulting performance ratio was plotted over the increasing levels of randomisation ranging from 0% to 100% of edges randomised. The maximum and minimum deviation from the average (10 simulation runs) is indicated by the error bars. Surprisingly, the relative performance drops in a linear fashion down to 1 in edges *seen* under increasing randomisation of the network. Also in edges *confirmed* the performance drops linearly with increased randomization, but here the relative performance stays ahead at a factor of over 1.25 at 100% randomisation.
5.5.2 Evolutionary Interpretation

Besides the obvious benefits of the pay-as-you-go strategy in unravelling interaction networks in a fast and reliable manner, this work may also shed some light on the evolutionary origins of the scale-free architecture found in all cellular networks.

The view that cells can be seen as computational units is not new (Bray, 1996). Obviously, the living cell has to perform many decisions based on the internal state and perceived changes of the environment, like expression of a specific transporter protein in the presence of a nutrient. Hence being able to make such decisions effectively is a clear evolutionary advantage.

It is important to note that from the viewpoint of complexity theory any given NP-hard problem can be reduced to another, such that if ever there is an efficient solution found for one problem in NP the solution for any problem in this complexity class can be directly deduced in an efficient manner. In addition, any computational problem can be formulated as a decision problem. The main point here is that in terms of complexity theory the actual decision problem itself does not matter, as long as it is NP-hard.

Put generally, life has to solve a number of computationally very hard problems. The protein folding problem is an example of a computationally unsolved problem that is ‘solved’ in nature very effectively. In nature, all these processes are based on local rules. Put simply, proteins do not have the global information on the status of the cell available to them. Hence they have to ‘act’ according to local rules (binding).

It seems relatively straightforward to imagine how the local principles of the pay-as-you-go strategy described here could be implemented by local recognition via the specific binding of protein-interfaces. In summary, interpreting the cell as a computational mechanism renders increased computational capacity of the cellular network a clear evolutionary advantage. Obviously, a NP-hard problem can be efficiently approximated on a scale-free distribution using local rules only. As more and more of the network is randomised the performance advantage of the pay-as-you-go over the random strategy drops to zero (see Figure 1).

So Figure 1 demonstrates the performance advantage stems from the non-random, scale-free distribution of the underlying cellular network. The emergent power-law degree-distribution is observed at the phase-transition of complex systems between order and chaos. In an evolutionary context,
the performance of the pay-as-you-go strategy could be seen as a measure of fitness of the overall network. At this point the resulting question, whether the performance of the pay-as-you-go strategy could indeed be used as an objective function to describe fitness, remains a hypothesis to be further investigated.

So far the scale-free distribution has been interpreted as a by-product of self-organisational principles at work or explained by the principles of preferential attachment and the robustness of scale-free networks against errors (Liu et al., 2002).

An alternative hypothesis was previously formulated by S. Kauffman: “The reason complex systems exist on, or in the ordered regime near, the edge of chaos is because evolution takes them there.” The randomisation results (Figure 12) support this hypothesis and suggest that if the computational capacity of a cellular network is related to the fitness of the organism then the scale-free distribution is an emergent property that should indeed be actively maintained by evolutionary processes. The fact that complex problems can be effectively approximated in the special case of scale-free networks by using local rules only may explain why the scale-free distribution has been observed on all cellular networks so far.
Conclusions

"'You’re only putting off the inevitable', he said.
'That’s what being alive is all about.'"

Terry Pratchett, 'The Last Continent'

The main challenge in the post-genomic era is to characterize the function and structure of the enormous amount of genes which are being discovered in the on-going sequencing efforts. To this end, I have developed novel algorithms for interaction networks. Following the central guiding hypothesis that protein function equals interaction helped to define the notion of protein function in a clean and computationally graspable way. I have utilized interaction information to assemble the jigsaw puzzle of proteins into functional maps, building on recent advances in experimental and computational techniques to study interactions among all proteins on a large-scale.

Modeling interaction networks as graphs enabled me to integrate protein-protein interaction data from different datasets and various organisms into a single accessible resource for further studies. I have formulated the concept of up-casting incomplete interaction information. The results are abstracted interaction maps which are generic and organism independent. Subsequent graph-theoretic analysis has revealed common underlying architectural principles of interaction networks, namely scale-free and small-world properties.

The insight that these networks are non-random lead to the idea to use the signal they contain for predictive purposes. The knowledge-base for further inference is built on existing biological classifications of protein structure (i.e. SCOP) and function (i.e. GO). I have formulated a novel algorithmic principle - *EMBed*. This non-homology method does not require any sequence or structure similarity. It enables accurate predictions beyond the limits of current bioinformatics methods which are based on the concept of homology.
However, the implications of this work do not stop there: the core-logic of the method allows for the development of a numerical index of functional similarity that is applicable across a wide range of biological data. This work puts functional genomics on firm mathematical and computational ground.

The most exciting prospect is mapping causal pathways of disease states through interaction networks. Other researchers have demonstrated that there is a correlation between centrality and lethality in biochemical networks, and that knock-outs in clusters of interacting proteins also form phenotypic clusters. Hence interaction networks provide a link from genotype to phenotype, and their analysis should form an excellent basis for target identification in the drug-discovery process.

In this context, moving on to higher level definitions of protein function, I addressed the question of how complex networks can be decomposed into meaningful subsets. In order to formulate experimentally verifiable hypotheses the functional maps must be made accessible to human interpretation. I have developed a method that reliably extracts whole signal-transduction pathways. Here complex sets of protein associations derived from text-mining the biological literature are used as an example of the complex and noisy environment information is processed within the cell. Although the overviews generated by my algorithm match human expectations, it raises the interesting question of how far our reasoning about molecular biology in terms of pathways meets the biological reality.

By building a concise model of the overall information gain in the field of proteomics I formulated an algorithmic strategy that enables the proteomics community to build a reliable scaffold of the interactome in a fraction of the time compared to un-coordinated efforts. This work opens the door to a new breed of methods and strategies which actively use the scale-free properties of biological networks to their advantage. As a result, the methods are able to approximate \( NP \)-hard problems in bioinformatics effectively by relatively simple computational means. Potential applications of this strategy to structure prediction and docking are being discussed. That \( NP \)-hard problems like edge-coverage can be approximated effectively, in the special case of scale-free interaction networks, leads to various kinds of new hypothesis. My current interpretation is that this problem reflects on the \textit{computational capacity} of such networks. Whether such a concept can indeed explain the common features of all biological networks in an evolutionary context is currently investigated.
5.5. DISCUSSION

It has been quite a journey from my first attempts to integrate various experimental sources and databases to insights about protein function, structure and evolution. I hope this work contributes to a more integrated view and holistic understanding of living systems.

Let me conclude with an intriguing quote from the philosopher Georg W.F. Hegel, who already described scientific logic as processes in the context of complex networks: “Within this network eventually tighter nodes establish, forming the anchor- and target points of his life and conscience, owing their strength and power to the very fact they are, brought before his conscience, the constituting terms of his very being.”

In diesem Netze schuerzen sich
hin und wieder festere Knoten,
welche die Anhalts- und Richtungspunkte
seines Lebens und Bewusstseyns sind,
sie verdanken ihre Festigkeit und Macht eben dem,
da sie vor das Bewusstseyn gebracht
an und fuer sich seyenden
Begriffe seiner Wesenheit sind.

The Project Gutenberg, EBook of
Wissenschaft der Logik' Vol.2
by Georg Wilhelm Friedrich Hegel
Appendix A

Publications During this Work

The following list details work published during the course of this thesis in chronological order.


Bibliography


