FINDING FUNCTIONAL GROUPS OF GENES USING PAIRWISE RELATIONAL DATA: METHODS AND APPLICATIONS

by

JOCHEN BRUMM

Diplom, Rheinische Friedrich-Wilhelms Universitaet Bonn, 1997

M.Sc., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2008

Abstract

Genes, the fundamental building blocks of life, act together (often through their derived proteins) in modules such as protein complexes and molecular pathways to achieve a cellular function such as DNA repair and cellular transport. A current emphasis in genomics research is to identify gene modules from gene profiles, which are measurements (such as a mutant phenotype or an expression level), associated with the individual genes under conditions of interest; genes in modules often have similar gene profiles. Clustering groups of genes with similar profiles can hence deliver candidate gene modules.

Pairwise similarity measures derived from these profiles are used as input to the popular hierarchical agglomerative clustering algorithms; however, these algorithms offer little guidance on how to choose candidate modules and how to improve a clustering as new data becomes available. As an alternative, there are methods based on thresholding the similarity values to obtain a graph; such a graph can be analyzed through (probabilistic) methods developed in the social sciences. However, thresholding the data discards valuable information and choosing the threshold is difficult.

Extending binary relational analysis, we exploit ranked relational data as the basis for two distinct approaches for identifying modules from genomic data, both based on the theory of random graph processes. We propose probabilistic models for ranked relational data that allow candidate modules to be accompanied by objective confidence scores and that permit an elegant integration of external information on gene-gene relationships.

We first followed theoretical work by Ling to objectively select exceptionally isolated groups as candidate gene modules. Secondly, inspired by stochastic block models used in the social sciences, we construct a novel model for ranked relational data, where all genes have hidden module parameters which govern the strength of all gene-gene relationships. Adapting a classical likelihood often used for the analysis of horse races, clustering is performed by estimating the module parameters using standard Bayesian methods. The method allows the incorporation of prior information on gene-gene relationships; the utility of using prior

information in the form of protein-protein interaction data in clustering of yeast mutant phenotype profiles is demonstrated.

Table of contents

List of tables

List of figures

Acknowledgements

I would like to thank my supervisors Jennifer Bryan and Wyeth Wasserman, who have both been instrumental in getting me excited and active in the field of statistics in molecular biology. Thanks for the excellent support, opportunities and the freedom to explore my own research, which made for a great PhD experience.

Thanks to my committee members, Paul Gustafson and especially John Petkau, who has been a great mentor over the years. Thanks to Liz Conibear, who was very generous with her time and data and fun to work with.

My academic environment has been great; many thanks to the people at the Department of Statistics and the Centre for Molecular Medicine and Therapeutics (CMMT), in particular the Wasserman lab. Special thanks to Shannan Ho Sui and Debra Fulton for organizing all the sushi lunches and for housing me on my visits in Vancouver.

Thanks to Hugo Dominguez for the generous accomodation, all the trips to the airport; also thanks to Hugo and Matias Salibian-Barrera hanging out with me often on Friday night.

I've interacted with many more people during the course of my PhD, academically and personally, and most interactions were very positive and nice. I want to thank all of these people; UBC is a great place to study.

Last but not least, I want to thank my family: my mom, who has been supportive during all these years; my partner Lukpla (Somrudee Sritubtim), who encouraged me to do a PhD in the first place and has been a source of inspiration throughout, and Lukpla's family who has been very welcoming.

Co-authorship statement

Chapter 2: The collaboration with the Conibear laboratory was initiated by Jennifer Bryan and Wyeth Wasserman. The analysis was conceived and executed by Jochen Brumm and Jennifer Bryan, with feedback from Wyeth Wasserman and Elizabeth Conibear. The manuscript was written jointly by Jennifer Bryan, Wyeth Wasserman and Jochen Brumm. Elizabeth Conibear read and approved the manuscript.

Chapters 3 and 4: The analysis was conceived and executed by Jochen Brumm, with feedback from Jennifer Bryan, Wyeth Wasserman and Elizabeth Conibear. The manuscript was written by Jochen Brumm, with suggestions and edits from Jennifer Bryan and Wyeth Wasserman.

Chapter 1: Introduction

1.1 Outline of the thesis

A brief overview of the motivation, objectives and content of the thesis precedes presentation of a review of the scientific literature pertinent to the thesis. The overview will highlight themes and techniques common to different chapters and provides a link between the contents of the thesis and the diverse topics in the literature review. To keep the overview short, we provide details and references later in the literature review, where the topics are discussed in more detail.

An ultimate aim of the thesis research was to develop applied statistical methods appropriate for data generated by a biological researcher who has a particular interest in a cellular system. Examples of such cellular systems are DNA repair, vesicle transport and chromosome segregation. These systems are usually composed of modules of genes acting in concert. The common goal to our methods is to predict functional relationships between genes (or more precisely the proteins they produce) through identification or expansion of modules – groups of interacting genes or proteins – involved in such systems.

The data is derived from assays that query an objectively selected set of potentially relevant genes. Genes were selected because they showed interesting behavior in prior assays relevant to the system of interest (we typically deal with less than 500 genes per study). We collaborated closely with the laboratory of Dr. Elizabeth Conibear (UBC), which studies the vesicle transport system in yeast. The Conibear data and other emergent data from the yeast research community capture a quantitative contribution of individual genes or combinations of genes to phenotypes or other quantifiable outcomes. There are two types of measures for relationships: the indirect measurement of relationships via gene profiles that query the similarity of behavior of genes under various conditions, and the direct measurement of relationships, where pairs of genes are subjected to a query and the outcome is seen as a measure of the relationship. Data producing assays can be classified as high-throughput and gold standard. The procedure followed by the Conibear lab and increasingly within the research community is to identify candidate relationships from high-throughput assays, which are then validated in time-consuming, gold standard follow-up experiments.

We typically know few genes that are members of a system/module, which makes prediction methods relying on training data difficult to use. We can, however, assume that pairs of genes in the same functional module will behave more similarly than and have stronger measured relationships than pairs of genes where the genes are in different modules. Methods that try to exploit this assumption to identify candidate modules are typically called clustering procedures.

There are many clustering procedures available, developed for many different applications. In genomics, the arrival of high-throughput assays for both indirect (e.g. microarrays) and direct measurement of relationships (e.g. yeast-two-hybrid and co-immunoprecipitation) has led to a proliferation of clustering methods.

Data on indirect measurements of relationships is most important for this thesis. This data can be represented in different ways: either as data points in a d-dimensional space (where d is the number of conditions measured), as a n-by-n matrix of distances or similarities (where n is the number of genes in the study), or as an n-by-n matrix with entries being 0 or 1, where 1 indicates a relationship and 0 indicates absence of a relationship. This last structure can be viewed as the adjacency matrix of a graph, where the nodes are the genes and edges are placed if the data indicated a relationship. This graph can be derived from the matrix of similarity values by 'thresholding' the data, classifying each value below threshold as 0 and each value above as 1. For this reason, the resulting graph is often called a threshold graph.

The matrix of similarity values is an attractive starting point for clustering in genomics. The similarity matrix is the input to the popular hierarchical agglomerative clustering algorithms. The drawback, however, is that methods based on similarity values tend to be purely algorithmic in nature. That is, a typical application of an algorithm like the hierarchical agglomerative clustering used on its own delivers a clustering, but the algorithm gives no guidance on which clusters should be considered candidate modules, there is little protection from false positive predictions in noisy data and it is not clear how to take advantage of other data sources that are available.

These problems make the thresholding option of similarity values attractive. It turns the problem of clustering indirect relational data into a problem of clustering direct relational data;

a problem that has received much attention in fields ranging from physics to image processing to social network analysis. There are a number of papers in which this methodology is successfully applied to the analysis of genomic data sets.

This thesis is concerned with generalizations of the threshold graph clustering procedure and their applications in genomics. Our approach is based on the application of a sequence of similarity thresholds, leading to a process of threshold graphs. The process starts with an empty graph (corresponding to a very strong threshold), and then acquires edges one-by-one according to less-and-less stringent thresholds until finally all edges have been added to the graph.

Using such a graph process as a data structure provides broader analysis options than a single threshold graph. For example we can examine the evolution of groups of genes during the process. In Chapter 2, we generalize a clustering method based on graph processes originally introduced by Ling [1] and apply it to profiles derived from yeast mutant phenotype profiles. The method considers the evolution of singly connected subgraphs within the process. We introduce a measure of external isolation of these groups in the process, and derive a p-value for this measure based on a null model of random evolution. We show that groups that are exceptionally well isolated correspond to protein complexes in the data sets to which we apply the method.

High-throughput genomics data can be difficult to interpret for multiple reasons. The data may be flawed due to technical limitations of the laboratory assays or instrumentation. Alternatively, the data may be an accurate reflection of the biological reality - there is considerable complexity in the biological system that is not adequately described by existing conceptualizations and therefore data may appear flawed, at our current level of understanding. In this thesis the term "Noisy" is used to describe data that fails to define well-isolated modules - such noise may be a reflection of either of the above points.

The measure of evolution developed in Chapter 2 is simple and has a nice correspondence to single linkage clustering. Noisy data (in the sense that genes in different modules are more similar to each other than genes in the same module), however, can lead to a total breakdown of the procedure (even for our generalized measure of isolation). Also, for the model of random evolution it is difficult to take advantage of other available information (prior knowledge),

which could be used to overcome noise in the data. Both of these points are major concerns in the analysis of yeast mutant profile studies, where the laboratory data can be noisy and relevant high-throughput data in the form of observed protein-protein physical interactions is readily available.

For this reason, we sought a simple model that could represent the data generation of the graph process. We were inspired by stochastic block models used in the social sciences. The model postulates that each gene belongs to a block, but we have lost the label indicating this membership. The probability that a relationship is observed between genes of the same block is higher than the probability of observing a relationship between genes from different blocks.

To make this model applicable to graph processes, one must have a likelihood for drawing an edge in the process. For this purpose, we recognized that a graph process is equivalent to a ranking of the edges of a graph. Likelihoods for rankings of items are well-established; the model that we found most attractive is that of a stagewise ranking model. In this model, each edge has its own unobserved 'ability' (this type of model is often used for races, where the ability is an inherent quality of an individual racer). At each step of the graph process, the probability of choosing an edge (in other word, winning the race) depends on the abilities of the edges remaining at this stage. The chosen edge is then eliminated from all subsequent stages (competitions).

In the context of the stochastic block model, the abilities of edges within a block are higher than the abilities of edges between blocks. This means that in the ideal, noise-free process all 'within block' edges appear before the 'between blocks' edges. This allows us to search for a labeling of genes that gets as close as possible to this 'ideal' process, since each labeling of genes implies abilities of the edges between genes.

This model can be viewed within the Bayesian paradigm, viewing the missing gene labels as parameters of interest. This approach has the advantage of making standard Bayesian optimization procedures available to search for the optimal labeling, and it enables us to systematically integrate additional information into the clustering procedure through a prior distribution on the labeling parameters.

Chapter 3 develops the stochastic block model in detail and demonstrates its utility on two data sets of yeast mutant phenotype profiles.

Chapter 4 shows how to apply the general model in the case of integrating protein-protein interactions with yeast mutant phenotype profiles.

1.2 Introduction to genomics research

Molecular biology is concerned with the study of properties and relationships of molecules to explain biological (mostly cellular) phenomena. Examples of such cellular functions include cell division, DNA repair, chromosome segregation, and regulated transport of molecules within cells. To inquire about these cellular processes, molecular biology draws upon techniques from different fields, most notably biochemistry and genetics, and techniques involving the manipulation and investigation of cells. A key focus point in molecular biology is the protein. Proteins themselves are translated from molecules called (messenger) RNA, which in turn is transcribed from DNA. The stretch of DNA encoding such RNA is called a gene, and is the most fundamental unit of interest in a cell since it provides both the basis for cellular function as well as inheritance. We will in this thesis talk about the "function of genes" or "genes that interact", which often refers to the function of the derived protein of a gene – we generalize the term gene to include both the DNA and its products (RNA and protein). There may be more than one protein derived from a gene, which is not important for the purpose of this thesis.

Research in molecular biology has exploded since the completion of the first genome sequencing projects, most notably for our purposes the genome of the model organism *S. cerevisiae* (baker's yeast). Knowledge of the entirety of the genes in an organism has motivated the automation of experiments, as researchers increasingly seek to obtain data for many genes at once. Experiments done in this high-throughput style are often referred to as "genomic" studies.

Model organisms play a vital role in translating the foundations laid by the genome sequence projects into insights into cellular function, in particular those relevant for human disease. Some key model organisms related to humans include, besides yeast, the pufferfish (*F. rubripes*), a

worm (*C. elegans*), a fruitfly (*D. melanogaster*) and the mouse (*M. musculus*). Data obtained in the study of these organisms allow researchers to make inferences about cellular functions of human genes, since many of the key cellular mechanisms are evolutionarily conserved. Since many labs work on the same organism, creation of technology and integration of effort becomes possible.

In this thesis, we only consider systems in yeast. Yeast is a well-studied organism where high-throughput data is readily available. Also, functional annotations of many genes are available, making the evaluation of computational methods easier. Our methods do not rely on annotations, making them attractive for lesser understood organisms, but our evaluation of performance takes advantage of the available annotations.

1.3 Relationships of genes and functional groups in cells

The cataloguing of genes through genome sequencing projects is an important foundation for molecular biology, but it typically doesn't allow researchers to understand the role of genes in human disease and development.

Key to the understanding of gene function is the recognition that genes typically do not act alone to accomplish cellular tasks. This became obvious when it was discovered that the number of genes in humans was only in the order of about 30,000, much less than originally anticipated given the complex nature of human biology and variety of human cell types. Since then, many studies have confirmed that in fact genes are often organized in 'assembly lines' called pathways, where genes act in sequence to achieve cellular function and in 'hubs' called protein complexes, where genes act jointly and are often linked by physical interaction [2-4].

Either structure can be defined abstractly by relationships between genes, where we define two genes to be related if they are interacting either in a protein complex or a pathway. This is a popular view of cellular systems in genomics, where authors will often summarize their findings from a sometimes bewildering array of experiments in a diagram displaying these perceived relationships; see [5] for an attempt to formalize this approach. A tool that has become very popular for displaying relationships between genes is Cytoscape [6]. Identifying these biological relationships from data is the key objective of this thesis.

1.4 Experimental data on relationships

It is typically difficult to establish biological relationships between genes from experimental data for a variety of reasons. Establishing true functional relationship usually requires a series of tedious, time-consuming 'gold standard' experiments that can only be performed for a few candidate pairs of genes.

In contrast, genome sequencing projects and subsequent assay development has led to the creation of many so-called "high-throughput" assays that query many genes at once. In this sense, we call our data genomic data. These genomic measurements are usually not seen as sufficient to prove function; follow-up gold standard experiments are needed. Hence it is important to keep the number of false positive predictions low because follow-up is expensive and can usually only be done for a few genes at a time.

One type of experimental relationship data captures a direct measure of physical interactions between proteins. Examples of such assays are the physical recovery of protein complexes extracted from cells (e.g. co-immunoprecipitation assay) [7-9] or a genetic testing system that produces a visible result if a physical interaction occurs between proteins produced by two gene sequences (i.e. yeast-two-hybrid assay) [10, 11]. In the co-immunoprecipitation approach, after cross-linking (creating covalent atomic bonds between proteins in close proximity) and thereby fixing the otherwise temporary protein-protein interactions, an antibody which specifically binds to a known protein can be used to extract all proteins interacting with the target protein. (The antibody can be physically attached to a bead and thus the bead-antibody-protein complex can be separated from the rest of the proteins which remain in solution.) The proteins recovered are then identified by one of many methods, for instance mass spectrometry. The yeast-two-hybrid system is based on the creative use of a well-known transcription factor (GAL4). Transcription factors are proteins that bind to a piece of DNA and catalyze the initiation of transcription of a nearby gene. In the two-hybrid assay, the GAL4 transcription factor is split into two separate proteins and each of these half-proteins is fused with one of two potentially interacting proteins. The transcription factor can only initiate transcription if the two half-proteins are physically interacting, thus the GAL4 protein activity will only be restored if the two proteins form a stable complex. To provide a measurable property, the target DNA

sequence for the GAL4 protein is placed adjacent to a "reporter gene" – a gene that produces a protein that has quantifiable activity. Hence increased detection of the reporter gene, typically beta-galactosidase for yeast two-hybrid assays, indicates physical interaction between the pair of proteins.

These direct measurements of interactions can be effective, but not all protein interactions can be detected in this way. The interactions in yeast-two-hybrid, for example, are forced within the artificial fusion protein within the cellular nucleus with only two proteins at a time. This is not the natural interacting environment the proteins find themselves in typically. Also, the transient 'assembly line' interactions within pathways are often not detectable by this assay. The same goes for co-immunoprecipitation experiments; here also the availability of suitable antibodies represents an obstacle. Both methods are plagued by considerable false positive rates and are difficult and expensive to perform on a high-throughput basis. For a review, see [12]. A number of databases are available (BioGrid [13], SGD [14], MIPS [15] and BIND [16]) that make the existing interaction data easily available.

When it is not feasible to obtain results from direct assays, researchers resort to indirect measurements of biological relationships. The key is to measure the 'phenotype' caused by a gene under a given condition. The most prevalent type of phenotype currently used is the transcript abundance of a gene under a certain condition ("gene expression"), which can be measured simultaneously for many genes by a microarray. In this thesis, we are concerned with a loss-of-function phenotype for a gene under a given condition. This is achieved by disabling a given gene in the organism, in our case yeast. To achieve this, a yeast strain is created that has a normal genome, except that for one chosen gene one of the two alleles is mutated, leading to half the "gene dosage" (often referred to as haploinsufficiency). If the strain is a diploid, that is if there is only one allele for each gene, the deletion eliminates the production of the protein completely. These strains may have impaired function, which may be assessed by measuring growth or other phenotypes associated with the strains. The difference from wildtype for these measurements is then associated with the mutated gene, providing a measure of the importance of the gene under the given condition.

If we now collect measurements of a given gene under various conditions, we obtain a feature vector that describes the 'behavior' of the gene. In the case of the yeast mutant phenotype

features, we refer to such a vector as a yeast mutant phenotype profile (YMP). To obtain a measure of the relationship between two genes, we score how similar their respective profiles are. This can be accomplished by applying one of many available distance or association measure to all the pairs of profiles. Common choices are measures based on Euclidean distance, (ranked) correlation or mutual information.

The yeast community has created a resource of yeast mutant strains, one for each gene in the yeast genome (except for the roughly 1500 mutants that do not survive under ideal growth conditions, corresponding to the so-called essential genes) [17]. Since the creation of this resource, studies with these strains have proliferated [18, 19].

In this thesis, we mainly focus on indirect measures of relationships. We use the direct measures of relationships for validation purposes in Chapter 3, and as prior information in Chapter 4. We do not discuss the choice of similarity measure in this thesis, although we recognize that it does have an impact on the results.

1.5 Finding functional groups of genes from data

Before the advent of high-throughput measures, biological researchers would only consider the most obvious relationships observed in data, which would be directly interpreted as biological relationships. This type of direct interpretation of data leads to many false predictions when applied to high-throughput data. Direct measures of association for example are known to produce spurious interactions due to the nature of the assay. Indirect measures of relationship on a continuous scale need to be converted into predicted true biological relationships. It is assumed that if genes are biologically related then their profiles are similar, but how similar does a pair of genes have to be before we call them biologically related?

To reduce number of false positive predictions of relationships, we take advantage of the fact that genes are organized in modules and look for groups of related genes (clusters). Clustering predicts functional relationships by effectively solving a discriminative problem, where each gene is associated with the group of genes that it is most closely related to. We assume for the sake of simplicity that each gene is associated with exactly one module and that the

experimental conditions are all relevant for the discrimination of modules, so that automatic feature selection is not necessary. Future work may address generalizations of these assumptions.

Clustering has been an important part of applied data analysis in many different fields, and many different methods for both indirect and direct measures of relationships are available. Clustering is generally described as the attempt to find groups of similar items in data in an "unsupervised" fashion (that is without using training data). It is a very large area of research and applied in many diverse areas of science, ranging from image analysis to linguistics. We will only discuss selected highlights from the literature here, and even for these selected items the number of available research articles is large. For the interested reader, some well-known books on clustering methods include [20-22].

We have already introduced the importance of groups of related genes in genomics research. A lack of training data is still common, even in well-studied organisms such as yeast, meaning that clustering methods will remain important in genomics for the long-term. Also, functional groups can be different in different situations, meaning that the actual associations of genes may vary between conditions investigated.

For the investigation of indirect measures of relationships, the data comes to us in the form of gene profiles. One possible type of clustering is based on viewing each profile as a data point in d-dimensional space, where d is the number of conditions under investigation. There are many popular methods based on this representation, like k-means [23] and model-based clustering procedures [24]. These common approaches have been successfully applied in genomics; see for example [25].

A different representation of the data starts by converting the matrix of gene profiles into a matrix of pairwise similarity values. For each pair of profiles, a measure like the Euclidean distance or correlation is used to produce a measure of similarity. Sometimes these similarities are derived as distances; we will keep referring to these measures as similarities (and assume that an appropriate transformation has been applied where necessary). The choice of the basic form of the distance matrix to best capture biological similarity depends on the nature of the relationship expected. Similarity based on correlation, for example, is less dependent on the

magnitude of the measured values than similarity based on Euclidean distance. A related issue is the choice of variable weighting; some papers advocate the use of weights applied to the values derived under a given condition. However, current research doesn't support a universal recommendation; see [26].

Hierarchical agglomerative linkage algorithms are extremely popular for clustering of a matrix of similarity values. These algorithms start by placing each gene in its own cluster and then proceed to join the 'closest' pair of clusters at each step until all genes are in the same cluster. This leads to a hierarchy of partitions, which can be represented by a tree-like structure called a dendrogram, where the leaves of the tree are the individual genes and the cluster joins are represented by joining the corresponding branches. Finally, groups are extracted from the dendrogram (so the dendrogram is typically not of interest by itself). The definition of 'closest pair of clusters' varies, but is always based on a summary measure of the similarities between all the pairs of genes in the two disjoint clusters. Popular choices are "single linkage", where the maximum similarity is taken as the summary statistic, "average linkage" (mean similarity) and "complete linkage" (minimum similarity).

The average linkage clustering algorithm was popularized in genomics by the seminal paper [27], which remains one of the most influential papers in genomic data analysis. Hierarchical agglomerative clustering remains popular [19, 28]; other popular clustering procedures are based on self-organizing maps [29]. Clustering has been successfully applied to yeast mutant phenotype profiles [18, 30]; for a review of clustering microarray data, see [31].

The matrix of similarity values is a flexible representation with proven utility in genomics (for a different use of the similarity matrix, see for example [32], where the matrix is used to find a set of representative examples to represent clusters). We will take this representation as the starting point for the methods in this thesis.

There is a class of methods that turn the matrix of similarity values into an adjacency matrix for a (weighted) graph [33, 34]. A graph has so-called nodes which in our context are genes and it has edges representing pairwise relationships. An adjacency matrix displays the value of the edge weight for all gene pairs (so both its rows and columns represent genes, and a value greater than zero in any cell means that an edge is placed between the respective genes. A

graph provides a very flexible and well-studied framework that is mainly meant to provide a language and representation for a wide variety of problems. By phrasing the clustering in this language, algorithms developed for graphs in general can be applied to this clustering problem. The basic problem in graph clustering is to find groups of nodes that either have many edges placed within the group or few edges placed to nodes outside the group, or both.

A simple and commonly used method is to apply a threshold to the similarity matrix to turn it into an adjacency matrix, so that an edge is placed between genes with a similarity higher than the threshold [35-37]. Methods of this nature are closely related to the methods in this thesis. We call a graph derived in this way a *threshold graph*. For a different way to construct a graph using a similarity matrix, see [38].

For indirect measurements, representing the data as a graph derived from a similarity matrix, although proven effective, is not the most immediate choice. For direct measures of relationships, however, graphs are the most obvious framework for data analysis. The fact that both types of data can be represented in such a way opens the door to graph based integration approaches, as we will see in Chapter 4.

Graph-based clustering in genomics has been mostly applied to direct measures of relationship. Clustering of direct relational data in genomics became popular and necessary with the generation of high-throughput interaction data. Previously, in high quality data, measured relations between genes were automatically considered 'real'. The high throughput data produces many false positives, making it necessary to process the data. It became clear that the clustered relational data could help reduce false positives in the data and hence reveal protein complexes [10, 39]. This has led to the development of many specialized clustering algorithms [40], and algorithms originally developed for other purposes have been successfully applied [41, 42]. Graph clustering is by no means confined to networks in genomics and has attracted much attention in diverse fields [43-45].

Our approach in Chapters 3 and 4 is inspired by stochastic block models for social networks [46, 47]. In these models, nodes are assumed to be labeled with a fixed number of labels, but the labels are not observable. Relationships are more likely to occur between nodes with the

same labels. This leads to a likelihood which can be optimized with standard statistical techniques, notably Gibbs sampling.

There are also approaches that take relational data and map it into a feature-space representation. An approach that maps a similarity matrix into a two-dimensional space is called multi-dimensional scaling [48]. Recently, there have been developments in the social sciences that map a network into an unobserved feature space [49].

1.6 Random graph processes

The methods in the previous section are, as viewed in our context, mostly for the analysis of a single threshold graph. While the use of a single threshold graph is convenient, because standard methods for graph clustering are immediately transferable, we propose to use a sequence of thresholds to derive a graph process as the basis for the analysis instead. This circumvents the choice of a threshold, and allows for the examination of the evolution of groups rather than groupings observed in a single graph.

The theory of evolving graph processes is in fact very old and famous and predates much of the modern graph analysis [50, 51]. Erdős and Rényi recognized that the evolving random graph process could be used as a tool for the analysis of individual graphs, where the graph under investigation is compared to graphs that are plausible to occur under the random evolution model. For a recent review of this branch of graph theory, see [52].

A generalization of this type of graph evolution has been successfully used in the analysis of complex networks like the internet and networks in ecology and sociology. In these observed networks, it was found that the distribution of degrees of nodes (the number of edges hitting a node) is different from the type of distribution an Erdős and Rényi graph process could produce. Networks are instead modeled as the result of an adaptive process, where the placing of a new edge depends on edges already placed. This type of model has been used to analyze global properties of protein interaction networks as well [44, 53-58].

Graph processes have also been used to judge the output of clustering algorithms [1]. It turns out that certain types of hierarchical agglomerative linkage algorithms can be viewed within the framework of graph processes; see for example [59] and [60]. We elaborate on this connection in Chapter 2. Ling's method is important for our work since it introduces measures of evolution for groups of nodes in the graph process. This type of scoring is at the heart of all methods developed in the scope of the thesis research.

The measures of evolution proposed by Ling in his work rely heavily on properties of subgraphs. In [1], he proposed to score the evolution of singly connected components in the process. Singly connected components are subsets of nodes where each node can be reached from any other node by traversing edges, and no other nodes outside the component can be reached. This can be effective, as we show in Chapter 2, but the composition of these singly connected components can be heavily affected by noise in the data used for the graph process.

To circumvent this problem, we chose a different route to score evolution in graph processes that is less affected by noise. The key to this approach is the observation that a graph process obtained by adding an edge at each step is mathematically equivalent to a ranking of edges. For this reason, we sometimes refer to this data structure as *ranked relational data*. We are not aware of any work previous to this thesis that exploits this connection for the purpose of clustering. Framing the problem in terms of ranks gives us access to a body of literature concerned with the analysis of ranks and permutations, of which we now highlight some relevant parts.

1.7 Likelihood based clustering of ranked relational data

The analysis of ranks of items, or viewed differently, permutations of items, has a long tradition. Ranks of items arise naturally in many applications, including among others in problems concerning the analysis of personal preferences [61] and horse races [62]. We will use the racing analogy to explain our genomic application.

Likelihoods for permutations are often based on the 'ability' of each horse. It reflects an unobserved quantity relating to the probability of the horse to win a race against other horses.

This quantity can be inferred if the horse is observed in multiple races, and once determined can be used to predict the outcome of a race.

Our analysis is based on the Luce-Plackett model [63]. We give a detailed description of this model in Chapter 3. There are many approaches to the analysis of permutations. Most closely related to our approach, besides the Luce-Plackett class, is the Thurstonian approach [61, 64]. For a book about the analysis of ranked data, see [65].

In our approach, the items to rank are edges representing relationships between genes. To apply the Luce-Plackett model, we need to equip each edge with an unobserved ability to beat other edges in the graph process race. Here is where the link to the stochastic block model for binary relational data comes in: the ability of edges connecting genes in the same block is higher than the ability of edges connecting genes in different blocks. An assignment of genes to blocks (labeling) hence leads to an assignment of abilities to edges. The likelihood scores how well these abilities reflect the observed graph process. The solution to the problem can now proceed in analogy to the binary problem, with the binomial likelihood replaced by the Luce-Plackett likelihood.

In Chapter 3 we show how to use standard statistical optimization techniques to solve the problem of finding a 'most likely' labeling of genes. Note that, with the assumption that the 'between blocks' abilities are lower than the 'within block' abilities; the likelihood is maximal if all 'within blocks' edges arrive first in the process, in agreement of our intuitive notion of a useful graph process.

1.8 Cluster validity

It has been long recognized that clustering algorithms produce clusters even when random data (without structure) is used as input. Also, even in the presence of clusters the assignment of items to clusters may not be optimal due to peculiarities of the algorithm.

For these reasons, a large number of tools are available to either protect the user from erroneous clusterings or to post-process the clustering output in order to improve the items to clusters assignments. The phrase "cluster validation" is sometimes used to refer to some of these tools. There are also many tools which validate clusterings using external reference data, which we will not consider here. A comprehensive overview of some of cluster validation in applications to gene expression is given in [66].

Related to the methods in this thesis are probabilistic tests that are applied to judge if data contains more than one cluster and post-processing tools that assess a data partitioning derived from a clustering algorithm. Probabilistic tests for the existence of a clustering structure in the data need to specify a model of 'structure-less' data. This model can be based on a feature space representation or a relational representation. Our p-values in Chapter 2 are based on the model of random evolution of the graph process. We also use as a competitive method for cluster selection a test derived by [67], which relies on a feature-space representation of the data. For reviews of such methods, see [68, 69].

Post-processing tools that assess a partitioning are plentiful, since they can use inherently intrinsic scores (based for example on within-cluster and between-cluster similarity) as an optimization criterion and do not have to specify a probabilistic model. A few such tools are described and assessed in a simulation study in [69]. Among the most popular are gap-statistics [70] and silhouette scores [71]. See also [72] for a general reference on how to compare partitions.

Another criterion to assess the validity of a cluster is stability, either across multiple clusterings obtained from bootstrapping the data set, or by combining clusterings obtained from various algorithms [73].

1.9 Graphs as models

It is important to distinguish between graphs that represent data and graphs that represent true biological relationships between genes. In this thesis, we only consider models for the true relationships that correspond to 'complete blocks'; that is a gene is either related to *all* other

genes in a module or to *none*. Such constraints are often relaxed in genomics, where this assumption may not be realistic. Some genes participate in more than one module, for example, and the heterogeneity of gene modules can lead to data structures where a gene is only closely related to a subset of its module.

Graphs as representations of true biological relationships are used in many methods in genomics, for an application to protein interaction networks, see [74]. Even more general network structures are considered in [75]. Graphical models are also used in this context, where the graph represents dependence structure between measurements across the gene set. For an example related to our applications, see [76] .

1.10 Integrated clustering of heterogeneous relational data

A diverse range of assays have emerged in genomic research. Since we cannot observe molecules in their native state, each assay needs to exploit a particular cellular process or phenomenon, leading to built-in biases of the methods. Co-immunoprecipitation for example, relies on the availability of suitable antibodies to the protein of interest. Yeast mutant phenotypes rely on the correlation between molecular interaction and the similarity in phenotype. Furthermore, predicting gene relationships from one assay alone can lead to many false positive relationships. For this reason, researchers strive to take advantage of the various data sources available, since integration may improve the predictions [77-79].

Methods for the integration of genomic data are actively researched. We are most concerned with approaches that aim to use multiple data sources to find modules, in contrast to approaches that are more directly aimed at data integration [80-82], which could be used in conjunction with our methods. Methods to use multiple data sets to predict gene modules are often discussed in the context of integrating gene expression and protein- protein interaction measurements [40, 83]. We will limit ourselves to the discussion of yeast mutant phenotype data and direct protein interaction data. This applied situation is simpler than in other cases, since both data sources try to explore the same underlying biological truth. For a paper that attempts to integrate phenotypic and protein-protein interaction data, see [84].

There are two simple strategies to integrate this type of indirect relational data with direct relational data. By clustering, it is possible to convert the indirect relational data into a graph, which is subsequently overlaid with the graph obtained from the direct measurements. Alternatively, we can incorporate the direct relational data into the similarity measure which is subsequently used for clustering [85, 86].

Our approach can deal with the two sources of information independently: by representing the direct relational data as prior information in our Bayesian approach, we can perform integrated clustering. Our approach is similar in spirit to work on the model-based clustering of gene expression data [87] and to approaches to clustering with soft constraints [88].

1.11 Contributions of this thesis

This thesis introduces graph processes as a useful tool for data analysis and presents methods to analyze this data structure. Methods built on graph processes retain benefits of the relational data model while using more information than a single graph. We demonstrate that following the evolution of groups of genes in this process can be used to identify functional modules in genomic data.

Chapter 2 builds on Ling's theoretical work for cluster scoring based on external isolation, extends it and applies it to genomic data. We generalize Ling's method to allow for situations with higher noise levels, and demonstrate that this generalization indeed improves sensitivity.

Chapter 3 introduces a stochastic block model for the analysis of graph processes. This is a novel parametric model for graph processes, and we demonstrate its utility in genomics. By modeling the graph process with a likelihood for a ranking, we introduce a novel model for relational data in genomics. Our parametrization of the parameters for this likelihood is inspired by similar models for binary data in the social sciences. Our model allows the use of standard optimization techniques to solve difficult clustering problems by delivering a clustering solution with confidence scores. It also provides a basis for the integration of other data sources into the clustering.

In Chapter 4, we use the model developed in Chapter 3 to demonstrate the utility of our novel method for data integration to overcome noise in the data. We address the important problem of integrating protein-protein interaction data and yeast mutant phenotype profiles.

1.12 Related contributions

The choice and development of methods in this thesis were strongly influenced by several applied projects conducted in parallel to this thesis:

1) Global analysis of yeast endosomal transport identifies the Vps55/68 sorting complex [89].

This paper uses the method developed in Chapter 2 to predict protein complex membership using yeast mutant phenotype data and experimentally validates a novel interaction predicted by our method as biologically relevant.

2) Ulysses - an application for the projection of molecular interactions across species [90].

This paper demonstrated that reliable biological relationships can be obtained by overlaying multiple networks of protein-protein interactions. It also shows that the networks suffer from low coverage of protein interactions.

3) Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response [91].

Our thesis demonstrates the utility of the derived methods on yeast mutant data. Gene expression measurements are another important data source where the prediction of gene modules is of interest. In this paper, we analyzed a time-course of microarray measurements during the fermentation of grape juice. We used a semi-supervised clustering approach to identify gene modules involved in the stress response during fermentation.

4) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes [92].

As an alternative to intrinsic evaluation of candidate modules, sometimes it is possible to score based on an independent source of data. In this paper, the overrepresentation of transcription factor binding sites in a candidate gene module (derived from a gene-expression study, for example) is used to identify meaningful gene regulation relationships.

5) Gene characterization index: assessing the depth of gene annotation [93].

This paper explores the nature and depth of annotations in the human genome by establishing an automated score using available data sources designed to capture a researcher's perception of 'depth of annotation' of a given gene. Applying the score to the genome reveals trends in research and potential drug target genes.

1.13 References

1. Ling RF: **Probability theory of cluster analysis**. *J Am Stat Assoc* 1973, **68**(341):159-164.
2. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12123-12128.
3. Rives AW, Galitski T: **Modular organization of cellular networks**. *Proc Natl Acad Sci U S A* 2003, **100**(3):1128-1133.
4. Pereira-Leal JB, Enright AJ, Ouzounis CA: **Detection of functional modules from protein interaction networks**. *Proteins* 2004, **54**(1):49-57.
5. Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes**. *Science* 2004, **306**(5701):1555-1558.
6. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.
7. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**(6868):180-183.
8. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al*: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440**(7084):631-636.
9. Krogan NJ, Cagney G, Yu HY, Zhong GQ, Guo XH, Ignatchenko A, Li J, Pu SY, Datta N, Tikuisis AP *et al*: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae**. *Nature* 2006, **440**(7084):637-643.
10. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast**. *Nat Biotechnol* 2000, **18**(12):1257-1261.
11. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins**. *Proc Natl Acad Sci U S A* 2000, **97**(3):1143-1147.
12. Vidal M: **Interactome networks.** *Dev Biol* 2005, **283**(2):579-579.

13. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Res* 2006, **34**(suppl_1):D535-539.
14. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res* 1998, **26**(1):73-79.
15. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast**. *Nucl Acids Res* 2006, **34**(suppl_1):D436-441.
16. Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV: **BIND--The Biomolecular Interaction Network Database**. *Nucleic Acids Res* 2001, **29**(1):242-245.
17. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al*: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418**(6896):387-391.
18. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH *et al*: **Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast**. *Cell* 2006, **126**(3):611-625.
19. Lee W, Onge RPS, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G: **Genome-wide requirements for resistance to functionally distinct DNA-damaging agents**. *PLoS Genet* 2005, **1**(2):235-246.
20. Kaufman L, Rousseeuw PJ: **Finding Groups in Data: An Introduction to Cluster Analysis**. New York: Wiley; 1990.
21. Hartigan JA: **Clustering Algorithms**. New York: John Wiley & Sons; 1975.
22. Everitt B, Landau S, Leese M: **Cluster Analysis**, 4th edn. London: Arnold 2001.
23. Hartigan JA, Wong MA: **Algorithm AS 136: a K-means clustering algorithm**. *Appl Stat* 1979, **28**(1):100-108.
24. Fraley C, Raftery AE: **Model-based clustering, discriminant analysis, and density estimation**. *J Am Stat Assoc* 2002, **97**(458):611-631.
25. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**(10):977-987.
26. Gnanadesikan R, Kettenring JR, Tsao SL: **Weighting and selection of variables for cluster-analysis**. *J Classif* 1995, **12**(1):113-136.

27. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.

28. Measday V, Baetz K, Guzzo J, Yuen K, Kwok T, Sheikh B, Ding HM, Ueta R, Hoac T, Cheng B *et al*: **Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation**. *Proc Natl Acad Sci U S A* 2005, **102**(39):13956-13961.

29. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation**. *Proc Natl Acad Sci U S A* 1999, **96**(6):2907-2912.

30. Dueck D, Morris QD, Frey BJ: **Multi-way clustering of microarray data using probabilistic sparse matrix factorization**. *Bioinformatics* 2005, **21**:i144-i151.

31. Quackenbush J: **Computational analysis of microarray data**. *Nat Rev Genet* 2001, **2**(6):418-427.

32. Frey BJ, Dueck D: **Clustering by passing messages between data points**. *Science* 2007, **315**(5814):972-976.

33. Sharan R, Maron-Katz A, Shamir R: **CLICK and EXPANDER: a system for clustering and visualizing gene expression data**. *Bioinformatics* 2003, **19**(14):1787-1799.

34. Xu Y, Olman V, Xu D: **Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees**. *Bioinformatics* 2002, **18**(4):536-545.

35. Hartuv E, Schmitt AO, Lange J, Meier-Ewert S, Lehrach H, Shamir R: **An algorithm for clustering cDNA fingerprints**. *Genomics* 2000, **66**(3):249-256.

36. Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B, Chesler EJ, Branstetter LK, Langston MA: **Extracting gene networks for low-dose radiation using graph theoretical algorithms**. *PLoS Comput Biol* 2006, **2**(7):757-768.

37. Rougemont J, Hingamp P: **DNA microarray data and contextual analysis of correlation graphs**. *BMC Bioinformatics* 2003, **4**:-.

38. Getz G, Levine E, Domany E, Zhang MQ: **Super-paramagnetic clustering of yeast gene expression profiles**. *Physica A* 2000, **279**(1-4):457-464.

39. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**:2.
40. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data**. *BMC Syst Biol* 2007, **1**(1):8.
41. Van Dongen S: **Graph clustering by flow simulation**. *PhD Thesis.* Centre for Mathematics and Computer Science (CWI), University of Utrecht; 2000.
42. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics* 2006, **7**:488.
43. Wu Z, Leahy R: **An optimal graph-theoretic approach to data clustering - theory and its application to image segmentation**. *IEEE Trans Pattern Anal Mach Intell* 1993, **15**(11):1101-1113.
44. Newman MEJ, Watts DJ, Strogatz SH: **Random graph models of social networks**. *Proc Natl Acad Sci U S A* 2002, **99**:2566-2572.
45. Hartuv E, Shamir R: **A clustering algorithm based on graph connectivity**. *Inform Process Lett* 2000, **76**(4-6):175-181.
46. Nowicki K, Snijders TAB: **Estimation and prediction for stochastic blockstructures**. *J Am Stat Assoc* 2001, **96**(455):1077-1087.
47. Snijders TAB, Nowicki K: **Estimation and prediction for stochastic blockmodels for graphs with latent block structure**. *J Classif* 1997, **14**(1):75-100.
48. Kruskal JB: **Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis**. *Psychometrika* 1964, **29**(1):1-27.
49. Hoff PD, Raftery AE, Handcock MS: **Latent space approaches to social network analysis**. *J Am Stat Assoc* 2002, **97**(460):1090-1098.
50. Erdős P, Rényi A: **On random graphs I.** *Publ Math - Debrecen* 1959, **6**:290-297.
51. Erdős P, Rényi A: **The evolution of random graphs**. *Magyar Tud Akad Mat Kutató Int Közl* 1960, **5**:17-61.
52. Bollobás B: **Random Graphs**, 2nd edn. Cambridge ; New York: Cambridge University Press; 2001.
53. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**(6804):651-654.
54. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks**. *Nature* 1998, **393**(6684):440-442.

55. Barabasi AL, Oltvai ZN: **Network biology: Understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.
56. Albert R: **Scale-free networks in cell biology**. *J Cell Sci* 2005, **118**(21):4947-4957.
57. Farkas I, Jeong H, Vicsek T, Barabasi AL, Oltvai ZN: **The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae**. *Physica A* 2003, **318**(3-4):601-612.
58. Barabasi AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**(5439):509-512.
59. Van Cutsem B, Ycart B: **Indexed dendrograms on random dissimilarities**. *J Classif* 1998, **15**:93-127.
60. Godehardt E: **Graphs as Structural Models**. Braunschweig/Wiesbaden: Vieweg; 1988.
61. Thurstone LL: **A law of comparative judgment**. *Psychol Rev* 1927, **34**:273-286.
62. Henery RJ: **Permutation probabilities as models for horse races**. *J Roy Stat Soc B Met* 1981, **43**(1):86-91.
63. Critchlow DE, Fligner MA, Verducci JS: **Probability-Models on Rankings**. *J Math Psychol* 1991, **35**(3):294-318.
64. Stern H: **Models for distributions on permutations**. *J Am Stat Assoc* 1990, **85**(410):558-564.
65. Marden JI: **Analyzing and Modeling Rank Data**. London: Chapman & Hall; 1995.
66. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset**. *Genome Biol* 2002, **3**(7):0036.0031 - 0036.0021.
67. Duda RO, Hart PE: **Pattern Classification and Scene Analysis**. New York: Wiley; 1973.
68. Bock HH: **Probabilistic models in cluster analysis**. *Comput Stat Data An* 1996, **23**(1):5-28.
69. Milligan GW, Cooper MC: **An examination of procedures for determining the number of clusters in a data set**. *Psychometrika* 1985, **50**(2):159-179.
70. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic**. *J Roy Stat Soc B* 2001, **63**:411-423.
71. Rousseeuw PJ: **Silhouettes - a graphical aid to the interpretation and validation of cluster-analysis**. *J Comput Appl Math* 1987, **20**:53-65.
72. Hubert L, Arabie P: **Comparing partitions**. *J Classif* 1985, **2**(2-3):193-218.

73. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data**. *Mach Learn* 2003, **52**(1-2):91-118.

74. Scholtens D, Vidal M, Gentleman R: **Local modeling of global interactome networks**. *Bioinformatics* 2005, **21**(17):3548-3557.

75. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data**. *Proc Natl Acad Sci U S A* 2004, **101**(9):2981-2986.

76. Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP: **A latent variable model for chemogenomic profiling**. *Bioinformatics* 2005, **21**(15):3286-3293.

77. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.

78. Gerstein M, Lan N, Jansen R: **Proteomics - integrating interactomes**. *Science* 2002, **295**(5553):284-+.

79. Ge H, Walhout AJM, Vidal M: **Integrating 'omic' information: a bridge between genomics and systems biology**. *Trends Genet* 2003, **19**(10):551-560.

80. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: **A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8348-8353.

81. Kiemer L, Costa S, Ueffing M, Cesareni G: **WI-PHI: a weighted yeast interactome enriched for direct physical interactions**. *Proteomics* 2007, **7**(6):932-943.

82. Jansen R, Yu HY, Greenbaum D, Kluger Y, Krogan NJ, Chung SB, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**(5644):449-453.

83. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction data**. *Bioinformatics* 2002, **18**(Suppl 1):S233-S240.

84. Walhout AJM, Reboul J, Shtanko O, Bertin N, Vaglio P, Ge H, Lee H, Doucette-Stamm L, Gunsalus KC, Schetter AJ *et al*: **Integrating Interactome, Phenome, and Transcriptome Mapping Data for the C. elegans Germline**. *Current Biology* 2002, **12**(22):1952-1958.

85. Huang DS, Pan W: **Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data**. *Bioinformatics* 2006, **22**(10):1259-1268.

86. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data**. *Bioinformatics* 2002, **18**(S1):S145-S154.

87. Pan W: **Incorporating gene functions as priors in model-based clustering of microarray gene expression data**. *Bioinformatics* 2006, **22**(7):795-801.

88. Law MHC, Topchy A, Jain AK: **Clustering with soft and group constraints**. *Lect Notes Comput Sc* 2004, **3138**:662-670.

89. Schluter C, Lam K, Brumm J, Wu B, Saunders M, Stevens T, Bryan J, Conibear E: **Global analysis of yeast endosomal transport identifies the Vps55/68 sorting complex**. *Mol Biol Cell* in press.

90. Kemmer D, Huang Y, Shah SP, Lim J, Brumm J, Yuen MMS, Ling J, Xu T, Wasserman WW, Ouellette BFF: **Ulysses - an application for the projection of molecular interactions across species**. *Genome Biol* 2005, **6**(12):-.

91. Marks VD, Ho Sui SJ, Erasmus D, van der Merwe GK, Brumm J, Wasserman WW, Bryan J, van Vuuren HJJ: **Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response**. *FEMS Yeast Res* 2008, **8**(1):35-52.

92. Sui SJH, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes**. *Nucleic Acids Res* 2005, **33**(10):3154-3164.

93. Kemmer D, Podowski RM, Yusuf D, Brumm J, Cheung W, Wahlestedt C, Lenhard B, Wasserman WW: **Gene characterization index: assessing the depth of gene annotation**. *PLoS ONE* 2008, **3**(1):e1440.

Chapter 2: Discovery and expansion of gene modules by seeking isolated groups in a random graph process[1]

[1] A version of this chapter has been submitted for publication. Brumm J, Conibear E, Wasserman WW, Bryan, J: Discovery and expansion of gene modules by seeking isolated groups in a random graph process.

2.1 Background

Much of systems biology research aims to identify biologically meaningful relationships between genes or their products, such as protein-protein interactions or co-membership in a biological pathway. This undertaking can be viewed as moving from the "parts lists" produced by genome sequencing projects to the assembly instructions for a complex system.

The combination of entities and their relationships is often described as a network, which can represent diverse biological systems such as cellular or signal transduction pathways [1, 2]. A common assumption made in the analysis of networks is the existence of biologically defined subnetworks commonly referred to as modules. Examples of such a module are a protein complex or a gene expression regulon.

Quantitative data from diverse genome-scale experiments can be exploited for the identification of new modules and the expansion of known modules. Correlation of expression levels or, more relevant to this study, loss of function phenotypes across multiple conditions provides an indirect measure of gene-gene relationship. Other assays such as yeast two-hybrid or genetic interaction screens using double knockouts, provide direct measures of these relationships. Early approaches to such studies were limited by a binary representation of the observations, but increasingly more powerful analysis is enabled by quantitative readouts [3-5].

While the quantitative data can be highly reproducible and informative, identifying the relevant functional relationships can still be a challenge. In noisy data there is great risk of predicting a spurious relationship between any pair of genes. An analytical approach based on modules, however, moves the focus from individual to connected sets of relationships. To invoke a concept from social network analysis, there is greater evidence for a relationship reinforced by common associations than for an individual, seemingly strong pairing. This principle is the basis for many algorithmic approaches for network identification [6-8].

The two main paradigms for module finding utilize different representations of the relationships: (i) a graph is obtained by applying a global threshold to the relationship data; or (ii) a hierarchy such as a dendogram (or tree) is produced by a clustering algorithm. In graph-based approaches, nodes represent genes and edges represent relationships. A 'threshold graph' is obtained from continuous relational data by classifying all pairs with similarity above the chosen threshold as related, and all other pairs as not related. The graph is subsequently processed, for instance based on the density of intra-group relationships, to produce candidate modules. In tree-based approaches, genes appear as leaves connected by branches, where branch height corresponds to some measure of relationship strength. Gene groups are obtained by pruning the tree, often by invoking a global height threshold.

In both approaches, the specification of a global threshold is fundamentally limiting. Modules in genomic data can be dissimilar: they can vary greatly in size, in internal cohesion (how related two genes within a module are) and external isolation (how unrelated the genes in the module are to genes in other modules). No single threshold graph or pruning of a tree reveals all of the modules in a heterogeneous biological system. Both methods are limited in their ability to perform well for the simultaneous analysis of all modules and are extremely sensitive to the selection of the threshold parameter.

We develop a novel approach for the detection of modules in relational genomic data. Our approach is fundamentally based on the *ranking of the relationships* between genes. Viewed in terms of the graph paradigm introduced above, we work with the entire sequence of threshold graphs that result from sliding the global threshold from stringent to permissive, Modules in this sequence of graphs are identified as groups that appear and persist as cohesive subgraphs. This approach for the detection of module isolation, which we refer to as the Miso method, permits the identification of modules with differing internal cohesion and determines the statistical significance of each candidate module. Extending a theoretical method introduced by Ling [9], the Miso method can also be used to score clusters in any single linkage dendrogram. In application to two collections of yeast mutant data [10, 11], we show that our method successfully identifies known protein complexes. Furthermore, our method predicted a new module which was subsequently experimentally confirmed [11]. A comparative study establishes that the Miso method performs very well relative to several alternative methods based on the post-processing of threshold graphs or dendrograms. Additionally, this comparison

underscores the practical advantage offered by the tuning parameter of the Miso method. Its natural interpretation as a measure of stringency provides external guidance when selecting a value appropriate for a specific application and, more generally, implies a predictable relationship between its value and classical measures of performance.

2.2 Results

2.2.1 Dissimilar biological modules in relational data

We assume that genomics data arrives in the form of ranked pairwise relationship scores (e.g. derived from Euclidean distance or correlation). While such data can be generated by many approaches and take many forms, for the purpose of this report we analyze only yeast mutant phenotype data in which the modules we seek are protein complexes. In Figure 2.1, we compare the intra-module and extra-module (genes not known to be in the module) relationships for known protein complexes associated with vesicle transport in yeast (description to follow below). We present both ranked relationships and the associated underlying continuous association measures. For all these complexes the intra-module relations generally are stronger than the extra-module relations. However, the threshold that provides the best modular distinction varies noticeably between complexes. Summarizing, there is no global threshold that is ideal for the recovery of all network modules.

2.2.2 The graph process captures evolving relationships across a spectrum of threshold

Graph processes are a useful representation of pairwise relationships. In contrast to a single graph, a graph process is an ordered set of graphs generated by incrementing a parameter. Conceptually within the process this parameter can be thought of as time. As illustrated in Figure 2.2, the process is initiated with a graph that has all genes but no edges. The next graph is obtained by placing an edge between the pair of genes with the highest relationship rank. Subsequent edges are added in the order of gene-gene relationship scores. This results in a

sequence of graphs that starts with an empty graph and ends with a complete graph. When a global threshold (i.e. one value of the parameter) is applied to relational data, the entire analysis is based merely on a single graph.

2.2.3 Candidate modules are subgraphs of significant persistence

It is our thesis that modules will appear and persist within this graph process for a period of time as identifiable subgraphs.

The most straight-forward, identifiable subgraphs are the 'singly connected' components that arise during the graph process. These subgraphs have the defining property that every gene pair is linked by a sequence of edges within the subgraph and no edge connects to this subgraph from the outside. Figure 2.3 presents an example of a graph process produced by the ranked yeast vesicle transport data. The emergence (Figure 2.3 A) and disappearance (Figure 2.3 B) of subgraphs corresponding to modules (protein complexes associated with the vesicle transport system) can be observed. The set of all singly connected components appearing in the graph process can be enumerated and form a set of candidate modules.

2.2.4 Figure of merit for candidate modules based on survival time

To facilitate interpretation, the candidate modules must be assessed with a quantitative measure of significance. Such a score ranks candidates for expensive validation studies and provides an objective measure of confidence. Our measure of significance for a candidate module is based upon the length of time it survives within the graph process as an identifiable subgraph. We say a subgraph is born when the associated set of nodes first becomes singly connected and dies upon the placement of the first edge connecting a node in the subgraph to a node outside the subgraph. The survival time is the difference between death and birth. Figure 2.2 illustrates birth, death, and survival of candidate modules in a simple example.

Following Ling [9], we assess the statistical significance of an observed survival time by comparing it to the distribution of survival times in a randomly evolving graph process. At the

birth of a specific candidate module, each remaining edge can be classified based on the two associated nodes; the edge is 'within' (both nodes in the candidate), 'between' (exactly one node in the candidate), or 'outside' (neither node in the candidate). The death of the candidate module occurs upon the placement of the first 'between' edge. The distribution of this waiting time under random evolution is easily obtained and, therefore, we can compute a p-value for the observed survival time. Intuitively, this method assumes that 'within' edges typically arrive before 'between' edges and that biological modules will often appear as identifiable subgraphs that enjoy unusually long survival times. We refer to this p-value as an isolation index.

2.2.5 Augmenting the list of candidate modules: removing high leverage edges

Noise in the data can lead to the premature placement of edges between genes belonging to distinct biological modules, which violates the assumption that 'within' edges arrive before 'between' edges. Such noise could arise from the limitations of the experimental assay or from true biological heterogeneity (e.g. a protein belongs to multiple modules). In our procedure, where candidate modules are singly connected subgraphs, such errors in edge order can affect survival times and even the composition of the list of candidate modules. Our method could fail to detect a true biological module if its survival time is truncated or if, when it first emerges, it is already embedded within some larger group of genes. These two problems arise when a mistimed edge arrives after or before, respectively, the birth of the module.

To make our Miso procedure robust to this sort of error, we extend it by considering the impact of high-leverage edges, i.e. the 'between' edges whose placement cause the death of a candidate module. To mitigate the effect of these high leverage edges that hit a module after birth, we compute the waiting time and associated p-value for the arrival of the k-th 'between' edge, for $k = 1,2, ..., K$, and define the extended k-isolation index as the minimum of these K p-values. To reduce the impact of edges that hit a module before birth, we consider parallel graph processes in which each individual high-leverage edge is postponed until the end. We extract candidate modules and associated p-values from these processes using the procedures described above. We form the list of candidate modules obtained from all one- and two-edge removed processes.

To introduce the nomenclature that follows - the number of allowed mistimed edges before and

after birth are given in parentheses. For example, Miso(0,1) refers to the modules that are not hit before birth, and for which isolation is assessed until the first edge hits after birth; Miso(2,6) refers to the isolated modules that were hit at most twice before birth and tracked until at most six hits have occurred after birth.

2.2.6 Relationship to single linkage clustering

Our approach, in which candidate modules are singly connected components, is related to single linkage hierarchical clustering. The candidate modules identified by Miso(0,1) are exactly the clusters arising in the dendrogram. Therefore one broadly useful application of our method is the selection of significantly isolated clusters from single linkage clustering (see [11]). While dendrograms are a useful representation of single linkage clustering, clusters that are significantly *k*-isolated with $k>1$ may not be detectable by visual inspection. Candidate modules detected via the removal of one or more high leverage edges may not even appear as clusters in the dendrogram.

2.2.7 Analysis of vesicle transport and DNA damage response in yeast

For the model organism *S.cerevisiae* the research community has created a collection of modified strains in which each member of a panel has a distinct gene disabled [10, 12, 13]. Using an appropriate assay, the phenotype of each strain in the panel is measured under a set of conditions. It is anticipated that for two genes within a module their respective mutants will display similar properties. We apply our methods to yeast mutant phenotype studies of two important systems - vesicle transport [11] and DNA damage response [14]. The modules within the vesicle transport system are well annotated, making this set suitable for the evaluation of our analytical method. Although the modules are less deeply annotated, we present an analysis of the DNA damage data as an independent validation.

We first applied our methods to a data set exploring vesicle transport. In eukaryotic cells, the directed movement of substances in membrane-bound sacs (vesicles) within the cell is called vesicle transport. Vesicle traffic is regulated by protein modules that select cargo for

incorporation into a forming vesicle and direct vesicle docking and fusion with the appropriate target membrane. The modules tend to be conserved between species, thus knowledge generated in studies of yeast can reveal the biochemical mechanisms by which defects in protein and lipid trafficking contribute to human disease.

Quantitative phenotypes were obtained under 14 conditions for 279 genes that displayed a strong phenotype in an initial, independent genome-wide screen [11]. The 279 genes include 137 genes known to belong to 25 modules. In the analysis reported in [11], we used the Miso(0,1) method with great success and the key results are displayed on top of a dendrogram. For example, the two largest candidate modules correspond almost exactly to two previously known modules – namely, the protein pump V-ATPase and the ESCRT subcomplexes (Table 2.1). Another high-scoring candidate module ("55-68") was subsequently validated in prospective experiments that confirmed a predicted protein-protein interaction. Here, in addition to the most conservative implementation [Miso(0,1), Table 2.1], we also apply our method in a more aggressive form [Miso(2,6), Table 2.2] to the vesicle transport data. We find that 78% [Miso(0,1)] and 63% [Miso(2,6)] of the predicted within-module relationships are, in fact, 'true', i.e. are implied by the prior knowledge, and that 48% [Miso(0,1)] and 53% [Miso(2,6)] of true relationships are successfully predicted. The Miso methods perform as well or, arguably, better than published alternatives at recovering and expanding modules in the yeast vesicle transport system (detailed further below).

We then applied our methods to the DNA damage response phenotype data described in [14]. DNA damage response pathways are relevant for cancer in humans, both for prevention and treatment. In [12] the authors analyze the phenotypic response of 140 deletion mutant strains in 36 conditions related to exposure to DNA damaging agents. From the average linkage dendrogram, interpreted in light of expert knowledge, the authors selectively identified the following six functional groups containing 23 genes:

- C1: NER (RAD2, RAD4, RAD10, RAD14, and RAD1)
- C2: error-prone TLS (REV1 and REV3);
- C3: PRR (RAD6, RAD18, and RAD5);
- C4: homologous recombination (RAD57, RAD51, and RAD54);
- C5: cell-cycle checkpoint control (RAD9, RAD24, RAD17, DDC1, and MEC3)

- C6: (SHU2, SHU1, CSM2, MPH1, and PSY3).

We recover these hand-picked modules in an objective fashion using our methods (Tables 2.3 and 2.4). The conservative Miso(0,1) output contains modules C1 and C2 perfectly, with partial recovery of C3 (2 of 3 predicted; no additional predictions), C4 (2 of 3; 2 additional genes included) and C5 (2 of 5; no additional). Two additional modules of 3 genes each were predicted. Analysis with Miso (2,6) recovers C1, C5 and C6 perfectly. Compared to the Miso(0,1) results, C3 is unchanged and both C2 and C4 have one additional gene. In addition, the Miso(2,6) method predicts only one other candidate module, with the noteworthy property that 4 out of the 5 genes are known to be involved in DNA repair.

2.2.8 Comparison of methods

To assess the relative performance of the Miso method, we applied it along with alternative methods to the vesicle transport data (Figure 2.4). The DNA damage response data is less suitable for comparative analysis due to the sparse annotations; the results are given for completeness in Figure 2.5, but will not be discussed in detail here. We selected representatives from the two broad categories described above: graph-based and dendrogram-based methods. We used MCL [6] as the representative graph clustering procedure because it performed well in a benchmark test [15]. For the identification of modules within a dendrogram, we consider both global cuts, including the one suggested by the Gap statistic, and local cuts. Objective local cuts, although not commonly used in genomics, are included because they are the most similar conceptually to the Miso method. Based on the work of Milligan and Cooper [16], we employ the local cut criterion introduced by Duda and Hart [17].

All of these procedures must be supplied with a tuning parameter to return a list of candidate modules. The tuning parameters of the methods we study are conceptually very different. For the Miso method, the tuning parameter is a p-value cutoff and therefore is a measure of stringency (i.e. the closer the parameter is to zero the higher the positive predictive value). For none of the other methods is there such a simple relationship between the tuning parameters and the performance of the method. MCL requires the specification of a similarity threshold which, in effect, corresponds to the selection of a single threshold graph. MCL proceeds to identify

candidate modules as dense subgraphs within the selected graph. In the context of a dendrogram, a global cut across a tree partitions the genes into clusters which are the candidate modules. This global cut can be viewed either as the selection of a similarity threshold, as in MCL, or more relevant in actual practice, as specifying the number of clusters. Local cuts applied to a dendrogram are implemented in a bottom-up manner to determine when to merge clusters based on their lack of separation.

Based on existing annotation of modules associated with vesicle transport, any relationship (edge) between two members of the same module is considered "true", all other relationships "false". Two metrics are computed to quantify performance. We use the positive predictive value (PPV), which gives the rate of true positive predictions among all positive predictions, and the sensitivity, which gives the proportion of true relationships predicted. Biological knowledge is incomplete, therefore a portion of the "false" predictions will be true – i.e. the reported PPV measures are conservative. Groups of greater than 50 genes were not considered as valid candidate modules, since the true biological modules of interest (protein complexes) are of much smaller size.

The results for our comparative study, given in Figure 2.4, show that the Miso methods are the best with respect to PPV and match the performance of the other methods with regards to sensitivity. In Figure 2.4 a), the Miso methods perform as expected with Miso(0,1) making a smaller number of higher quality predictions relative to Miso(2,6). The local cut methods perform uniformly worse with respect to PPV and exhibit maximum sensitivities that are comparable to those of the Miso methods (Figure 2.4 b). The most striking finding for the global cut methods is the volatile relationship between the tuning parameters and performance, especially for PPV (Figure 2.4 c). This volatility demonstrates the importance of the tuning parameter as well as the difficulty of choosing its optimal value, particularly in the absence of known annotations.

2.3 Discussion

Based on the analysis of graph processes, we have introduced a novel method for the identification of biological modules in ranked relational data. Building on a theoretical

foundation from Ling [9], the Miso methods accommodate the heterogeneity and noise that is inherent to genomic data and detect modules that vary widely in size, external isolation and internal cohesion. An objective measure of confidence -- a p-value -- is assigned to each candidate module, prioritizing candidates for further study. Because the isolation index is a measure of stringency, it is particularly attractive for applications in which there is little or no prior biological knowledge to guide the selection of tuning parameters.

In the ongoing effort to identify modules from genomic data, the most dominant methodological approaches are based on one of two representations of the data: graphs and hierarchical clusterings (dendrograms). Regardless of the analytical paradigm, a key challenge is to overcome the combined effect of biological heterogeneity and experimental variability. The distinctive, individual properties of real biological modules generally imply that there is no universal 'signature' that would enable module detection based on threshholding relationship strength or, by extension, some related summary measure. This reveals, therefore, a fundamental limitation of methods based on threshhold graphs or the global pruning of dendrograms. In the presence of diverse modules, the Miso methods are better able to perform well for many modules simultaneously, since each candidate module is evaluated in a distinct timeframe within an evolving graph process.

It is increasingly common to address the variability in genomic relational data by using probabilistic approaches to graphs [7, 15, 18]. When analyzing a single observed graph, noise can be acknowledged by recognizing the potential error associated with each observed edge (or lack thereof). The graph process paradigm for module finding, originally introduced by [9] and adapted and extended for genomic data analysis here, offers a natural extension of the probabilistic graph based analysis.

Data from genomic data tends to be very noisy (in that genes in different modules are more similar to each other than to genes in the same module). Our extensions of the original isolation index offer improved sensitivity by allowing for a few mistimed edges in the graph process. However, even the extended Miso method is sensitive to noise of the magnitude often observed in genomic data. While it will still maintain a high quality of predictions, its sensitivity will be diminished in datasets with less separation (due to lack of adequate assays or technical

variability). We expect that with continuing development of assays lack of separation will become less of an issue, whereas the heterogeneity of complexes is inherent in the biological systems.

The study of an evolving graph process offers a promising new direction for the discovery of biological modules. Building on the groundwork laid here, an intriguing approach to the analysis of noisy relational data is to move from a model of random evolution representing the absence of modules to a constructive model driven by the presence of modules. In such an approach, the probabilistic model for relational data is represented by a likelihood function that explicitly incorporates the properties of gene-gene relationships between and within modules. Such an extension provides a natural basis for solving even harder problems, such as the integration of relational data arising from distinct experiments or even different platforms.

2.4 Materials and methods

2.4.1 Data Sources

The vesicle transport data was reported in [11]. The data is obtained by plating yeast mutant colonies in 1536-array format on nutrient media. The growth of the colonies in the presence of various chemicals or the secretion of certain proteins (as determined by biochemical assays) is measured by quantifying images by densitometry. The measurement values were preprocessed by averaging across replicates, correcting for background intensity by subtracting the values of blank spots and converting the measurement of growth or secretion into a percentage relative to the wild-type strain. In an initial, independent genome-wide screen, 279 genes that displayed a strong phenotype were selected. Quantitative phenotype measurements can be arranged in array form with the rows being the gene knockout strains and the columns the conditions, the values of this array were processed by scaling each column by its standard deviation. Indirect measures of relationships were obtained by applying Euclidean distance to the rows of the preprocessed array.

The DNA damage data was taken from [10], where the authors report an analysis of 140 genes. We take the same genes here and apply our procedure to the data using a Euclidean distance measure.

2.4.2 Probabilistic model for the graph process and scoring survival times

A graph process is a representation of the ranking of quantitative, pairwise gene-gene relationships for n genes. Ties within the $N=n(n-1)/2$ gene-gene relationship ranks were randomly ordered and the resultant ranking was held constant for all analyses reported here. To analyze an observed graph process in a probabilistic fashion, the null model assumes that all rankings of pairwise relationships are equally likely. As noted by Ling [9], this assumption does not strictly hold when the rankings are derived from a distance measure, because of the constraint that the triangle inequality imposes on pairwise distances.

As introduced above, a candidate module is a simply connected subgraph and its survival time is the difference between the rank of the edge that established the subgraph and the edge that adds a new member. Our measure of significance is the p-value for the survival time in a random graph process. Since edges are drawn without replacement, the probability of choosing a particular edge at step t out of all possible N edges is $1/(N-t)$ (any of the $N-t$ edges left has equal probability).

For the purpose of scoring a specific candidate module, we define 'success' as the placement of an edge between two genes either both within or both external to the module. We define 'failure' as the placement of an edge between one gene within the module and one gene external to the module; this results in the death of the candidate module. We denote the probability of failure at step t as p_t, and note that the probability of success is simply $(1 - p_t)$. A survival time of r for a module of size c born at step b then implies there were $(r-1)$ successes at steps $b+1, b+2,..., b+r-1$ followed by a failure at step $b+r$. The probability of failure at step $b+j$ can be computed as follows: there are $N-(b+j)$ edges remaining, of which $c\ (n-c)$ constitute failures. Therefore the probability of failure at step $b+j$ is $p_{b+j} = c\ (n - c) / (N - b - j)$. The null distribution of the survival time S is given by

$$P(S = r) = (1 - p_{b+1})(1 - p_{b+2})\ldots(1 - p_{b+r-1})p_{b+r}$$

We approximate these probabilities by setting p_{b+j} to the constant p_{b+1}, leading to an approximating geometric distribution for *S*. Ling established that this approximation is good if *r* is small compared to *N-b*.

2.4.3 Generalized isolation

To address the problem of mistimed edges, we consider a more general set of survival times: the waiting times to the 1st, 2nd ,…, *k*-th failure, where *k* is a low number (we evaluated up to *k*=6).

To derive the null distribution of these survival times, we can use the same reasoning, except that the number of the number of edges that lead to failure is now $c\ (n\text{-}c) - k - 1$ so that $p_{b+j} = (c\ (n - c) - k - 1) / (N - b - j)$. We are now waiting for the *k*-th failure, so the approximating distribution is negative binomial rather than geometric. These two distributions are identical for $k = 1$.

The situation where a group of genes is hit before birth requires a different approach. To identify such groups, we analyze a modified graph process: we remove all *n-1* module-killing edges individually from the original graph process and apply our method to the modified process, leading to the Miso(1,*k*) method. Iterating this procedure by removing all *n-1* module-killing edges from all *n-1* modified processes leads to the Miso(2,*k*) method. For any candidate module, we utilize the minimum p-value for 1, 2, …, *k*.

2.4.4 Analyses of yeast mutant phenotype data

In all methods, we did not consider groups of size greater than 50 to be valid candidate modules since they are not suitable for follow-up experiments. For the Miso method, the results did not vary noticeably for a wide range of size cutoffs.

For the MCL method, we picked a sequence of threshold graphs with 500, 1000, … up to half

of the relationships available, at which point MCL only finds a few clusters. We also tried different settings of the granularity tuning parameter for MCL but found that while this parameter can improve the results for a single graph, the set of results for our sequence of graphs did not benefit from choosing different levels of granularity. For this reason we ran each MCL procedure with the default granularity setting.

For the global cuts of dendrograms, we chose cuts leading to 3, 10, 20, …, 100 clusters. We found that cuts leading to few clusters did not perform well, but included 3 since it was the value chosen by the gap statistic. To create Figure 2.4 and Figure 2.5, we retrieved the height corresponding to the chosen number of clusters and matched it to the rank in the process (number of edges with similarity less or equal to this height)

In contrast to the gap statistic and other methods that can be applied to any partitioning (from a hierarchical or any other clustering), local cuts (sometimes called "stopping rules") are restricted to hierarchical agglomerative methods. The cuts are conducted as formal tests if the cluster resulting from a merge step in the clustering contains one or two clusters (the joins are performed until there is evidence against the null-hypothesis of one cluster). For any chosen rejection value for the test statistic, the method delivers a set of candidate modules. To perform this test, assumptions about the distribution of the data have to be made, given in detail in [17].

All clustering except for MCL (version 06-058, default settings) was done in R [19].

2.5 Tables and figures

2.5.1 Tables

Table 2.1: Results for the Miso(0,1) method for the vesicle transport data
The table shows the composition of significantly isolated candidate modules (Bonferroni corrected p-value less than 0.05), and is a subset of the results in [11]. The "birth" column gives the step in the graph process when the module is first connected, the "death" column gives the first time an edge from outside hits the module, the p-value is computed with Miso(0,1). Most candidate modules can clearly be associated with a protein complex. The 55-68 cluster contains a relationship validated in [11]. Omitted are a candidate module of size 3 and 7 of size 2 with unknown annotations.

Candidate module	Size	Birth	Death	p-value	Composition
V-ATPase	18	4314	7922	6.82E-228	V-ATPase (18)
ESCRT	13	4348	4506	1.52E-05	ESCRT(13)
Retromer (I)	10	3224	3530	9.76E-09	Retromer(4),PI3K(2),ClassD *VPS*(1), ClassA/D *VPS* (1)
COG/YPT6	9	1252	1468	1.46E-04	COG(4),YPT6(4),ARF(1)
SWR-C	6	1073	1524	5.56E-07	SWR-C(6)
INO80	4	1413	2326	3.95E-10	INO80(2)
55-68	3	463	2072	1.51E-13	55-68(2),ClassD *VPS*(1) validated
ClassB	3	6451	7599	3.23E-11	ClassB *VPS* (3)
PI3KC	2	12208	21651	1.10E-84	PI3K(2)
ClassD	2	2644	6344	4.16E-23	ClassD *VPS* (2)
Garp	2	1029	2923	1.92E-10	GARP(2)
Retromer(II)	2	130	1844	2.70E-07	Retromer(2)

Table 2.2: Results for vesicle transport data with Miso(2,6).

Candidate modules derived from the Miso(2,6) method with a (Bonferroni-corrected) p-value cutoff of 0.05.

Cluster	Size	Remarks
Vatpase	21	Vatpase(18),ClassC *VPS*
Escrt	13	Escrt(13)
Retromer	12	Retromer(6),PI3KC(2)
SWR-C	11	SWR-C (8)
YPT/COG	9	YPT(4),COG(4),ARF(1)
55-68	4	55-68(2),ClassD *VPS*
EE	3	EE(2)
Glycosyl	3	Glycosyl(2)
PI3KC	3	PI3KC(2),ClassC *VPS*(1)
ClassD	3	ClassD *VPS*(3)
ClassB	3	ClassB *VPS*(3)
DNA	3	SWR-C(1), RSC(2)
Garp	2	Garp(2)

Table 2.3: Results for DNA damage data with Miso(0,1).

Candidate modules derived from the Miso(0,1) method with a (Bonferroni-corrected) p-value cutoff of 0.05.

Cluster	Size	Birth	Death	p-value	Remarks
RAD4, RAD2, RAD10, RAD14, RAD1	5	2233	6760	3.68E-145	5/5 from C1
RAD18, RAD5	2	995	5525	1.40E-52	2/3 from C3
MMS4P, YBR099C, MUS81	3	364	2838	3.73E-41	Not on list
REV1,REV3	2	13	2862	6.06E-31	2/2 from C2
RAD9, RAD24	2	252	795	1.34E-06	2/5 from C5
LTE1,BCK1,CLA4	3	219	490	4.03E-05	Not on list
RAD57,RAD55, RAD51, HPR5	4	162	353	7.64E-05	2/3 from C4

Table 2.4: Results for DNA damage data with Miso(2,6)

Candidate modules derived from the Miso(2,6) method with a (Bonferroni-corrected) p-value cutoff of 0.05.

Cluster	Size	Remarks
RAD4,RAD2,RAD10,RAD14,RAD1	5	5/5 from C1
RAD5,RAD18	2	2/3 from C3
REV1,REV3,RAD23	3	2/2 from C2
RAD59,MMS4P,YBR099C,PPH3,MUS81,SAE2	6	4/5 DNA Repair
SHU2,SHU1,CSM2,MPH1,PSY3	5	5/5 from C6
RAD9,RAD24,MEC3,RAD17,DDC1	5	5/5 from C5
RAD51,RAD57,RAD55,RTT101,HPR5	5	2/3 from C4

2.5.2 Figures

Figure 2.1: Smoothed histograms of the observed intra- (solid lines) and inter- (dashed lines) module relationships for selected protein complexes from yeast vesicle transport data [11].

The Euclidean distance is presented on a rank scale in A and on the original scale in B. The plots depict the heterogeneity in the internal cohesion and external isolation of protein complexes.

Figure 2.1 A

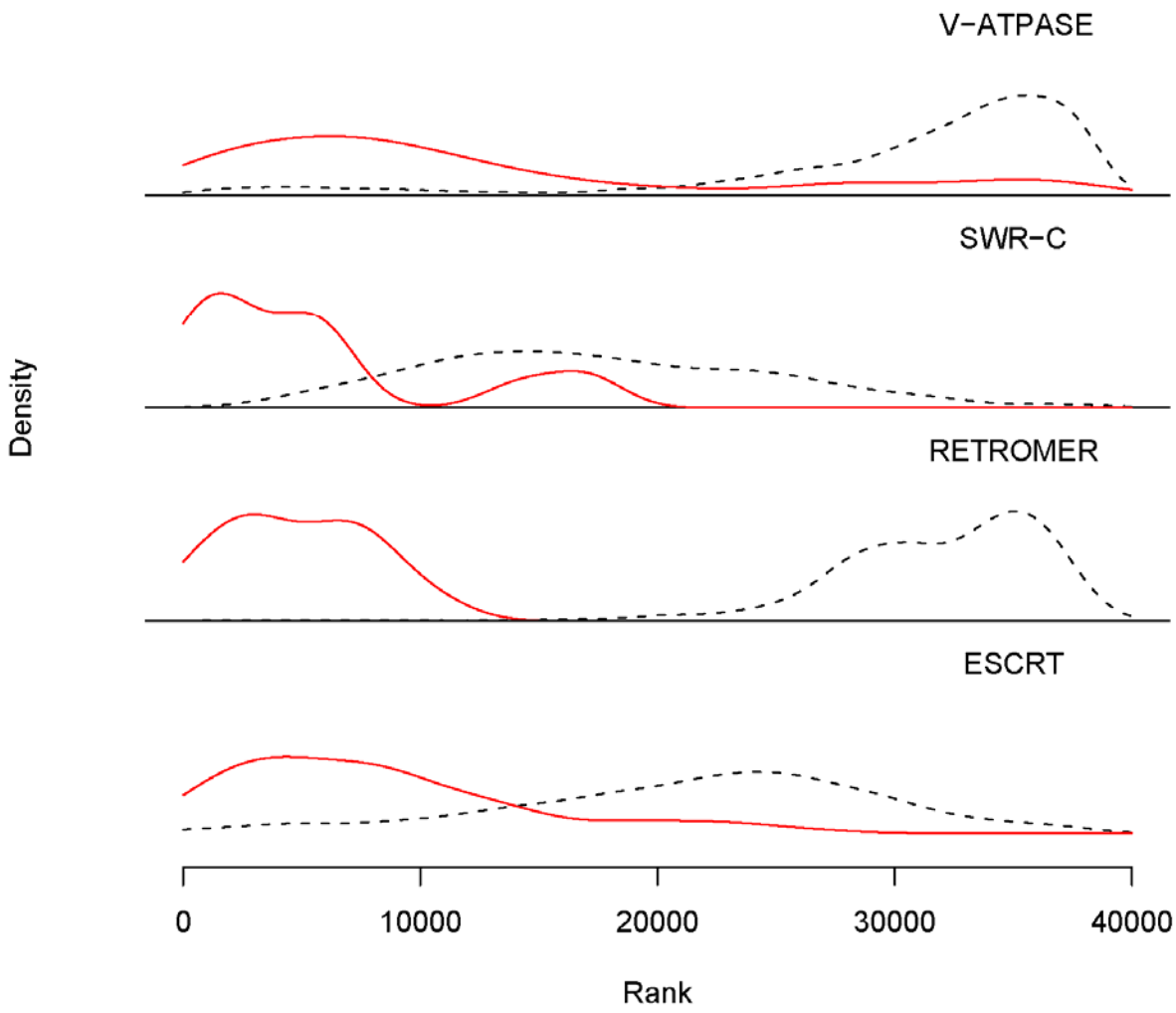


Figure 2.1 B

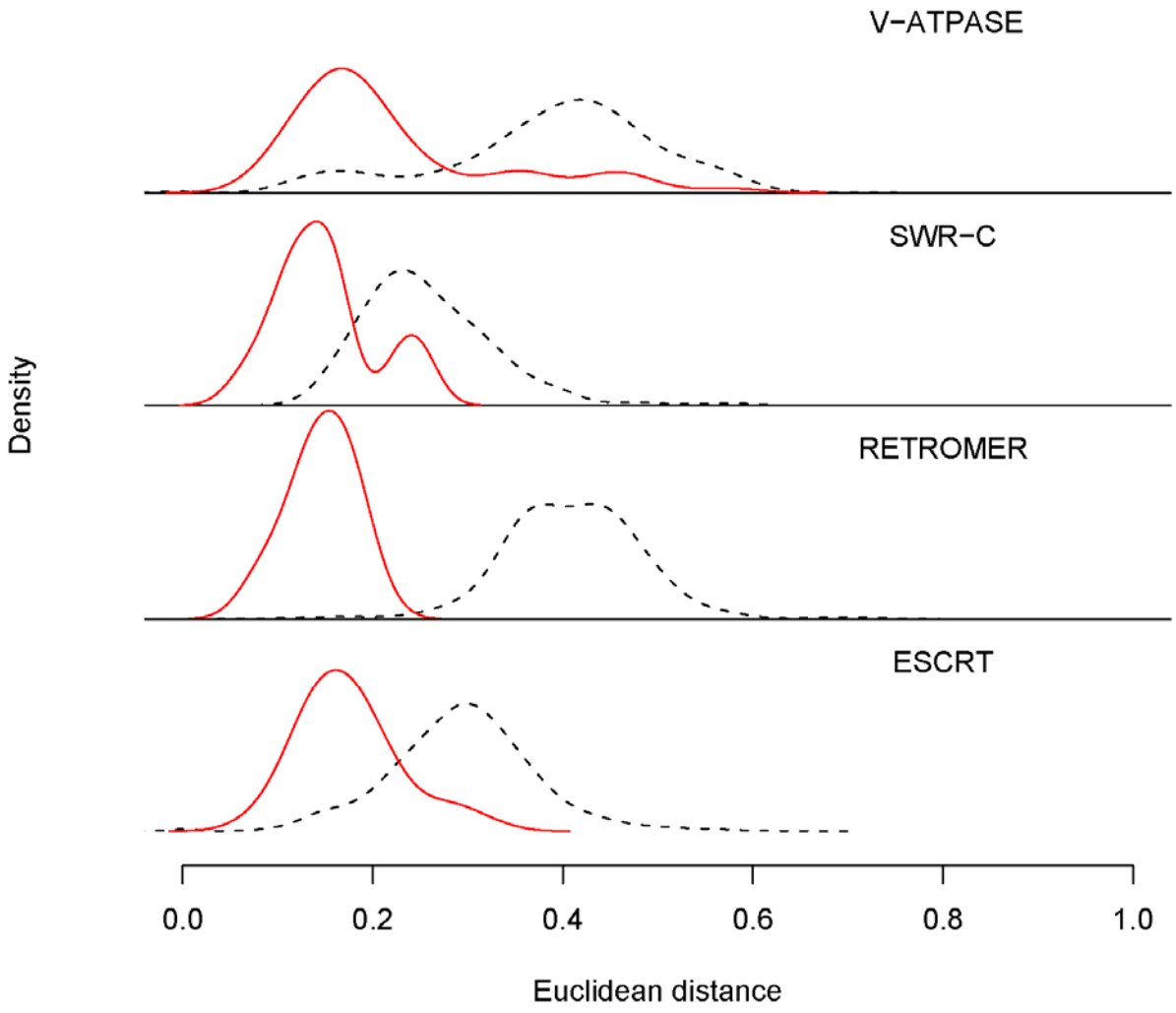

V−ATPASE
SWR−C
RETROMER
ESCRT
Density
0.0
0.2
0.4
0.6
0.8
1.0
Euclidean distance

Figure 2.2: Schematic illustration of a graph process and the birth and death of identifiable subgraphs, defined here as singly connected components.

A graph process proceeds by sequentially adding edges in rank order. When two subgraphs are joined, two candidate modules 'die' and a new candidate module is born. Survival time is defined as the number of edges added in the graph process between birth and death. We show steps 4,5,6 and 11 here in panels A, B, C and D, respectively. In B a 'between' edge joins subgraphs (2,3,5) and (4) into a new subgraph. In C, a 'within' edge is placed which does not affect subgraph membership. In D, the subgraph born in B dies resulting in a survival time of 11-6 = 5. Panel E provides the corresponding single linkage dendrogram. Note that the height of cluster merge events corresponds exactly to death and birth events of subgraphs.

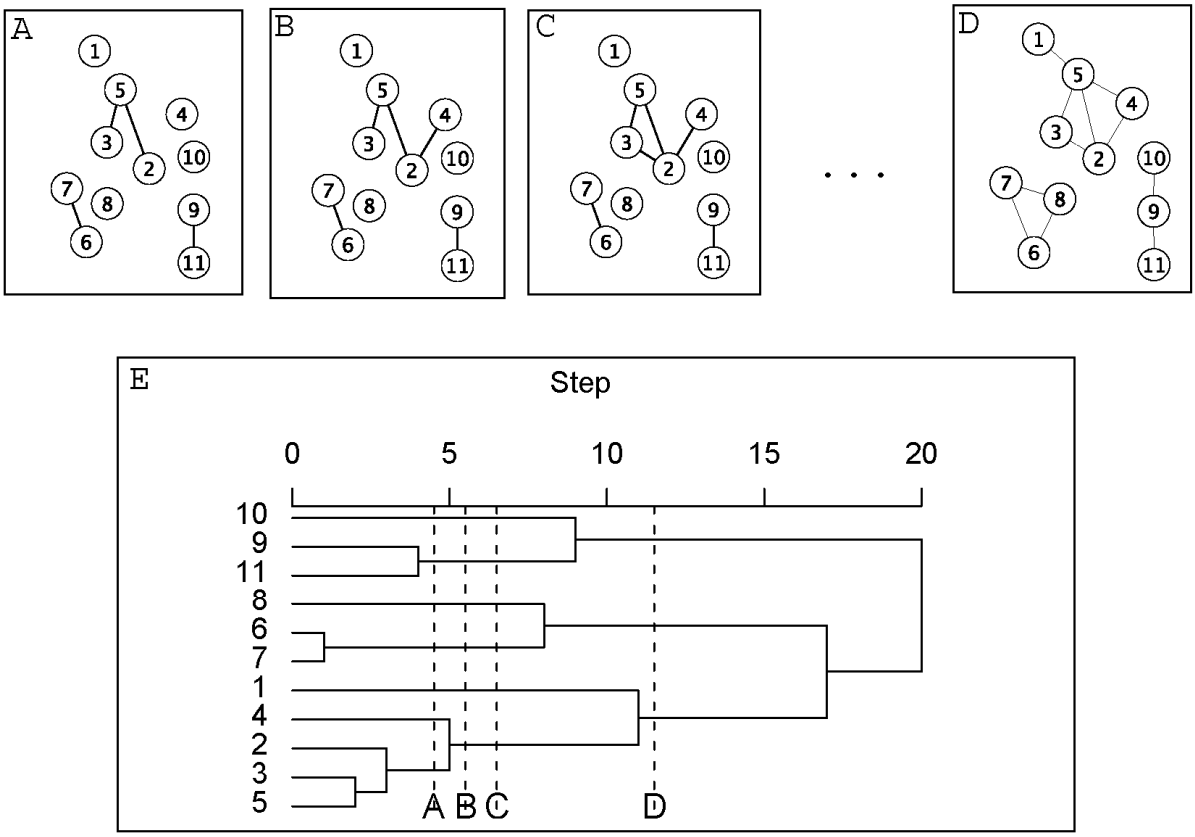


Figure 2.3: The observed graph process for yeast vesicle transport data at step 2000 (A) and 5000 (B).

Node color corresponds to protein complex membership; unannotated genes appear in grey. Specific identifiable subgraphs in panel A have been incorporated into larger subgraphs in panel B.

Figure 2.3 A:

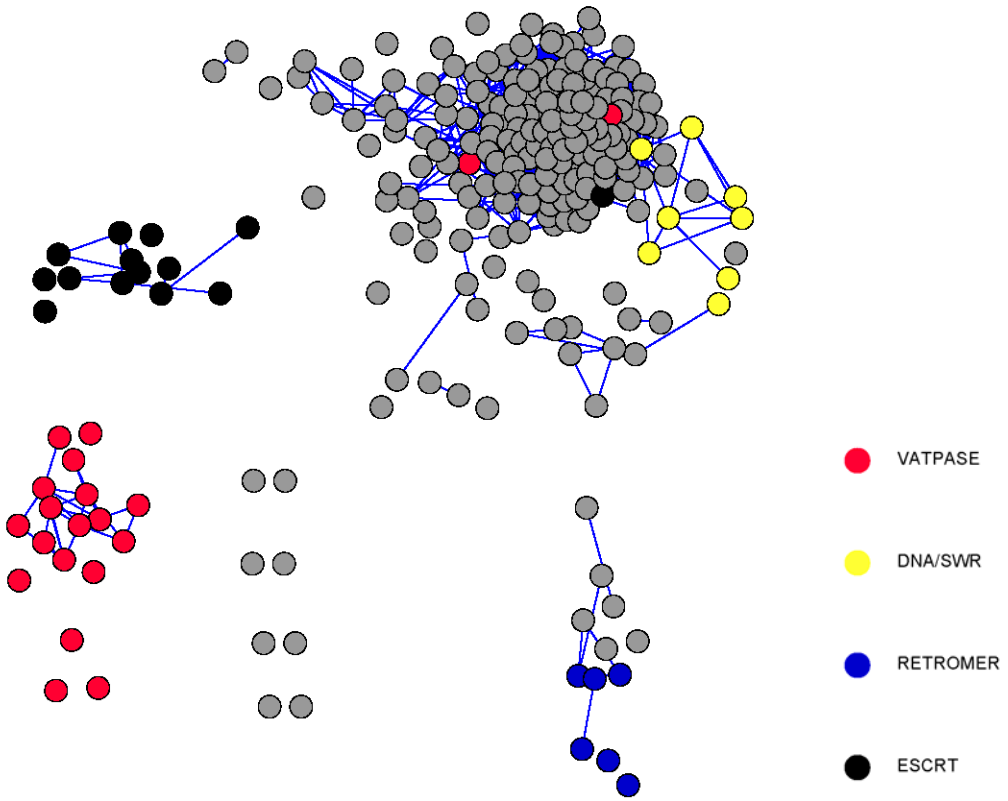


Figure 2.3 B:

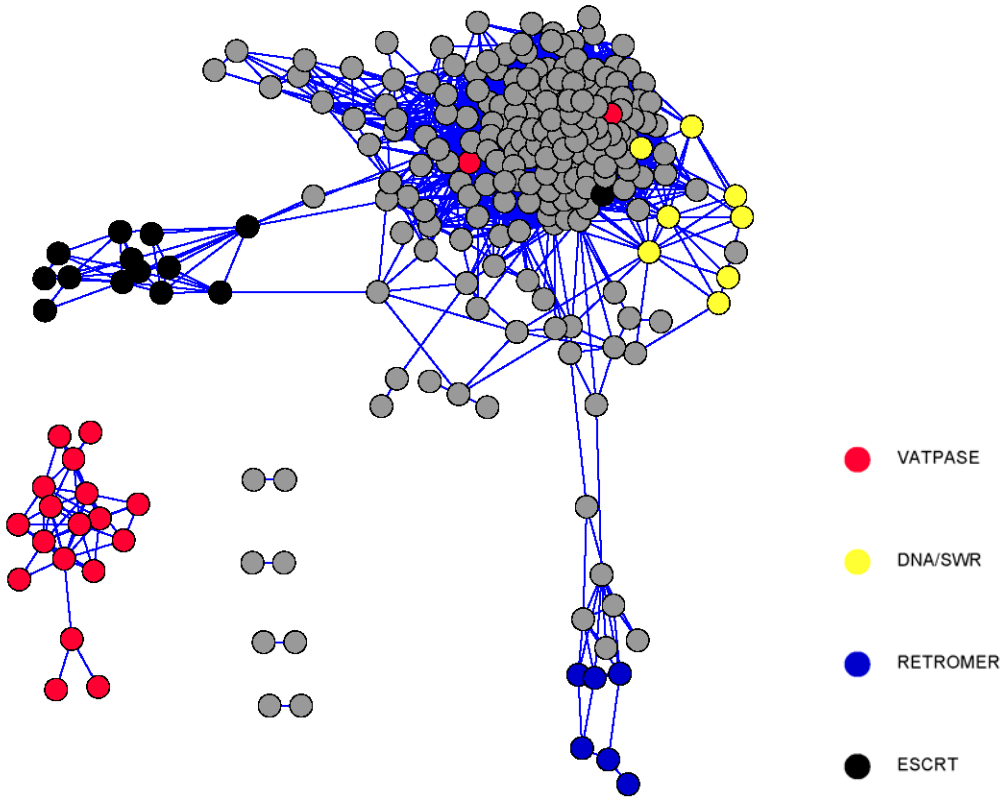


Figure 2.4: Relative performance of module detection methods applied to yeast vesicle transport data.

Displayed are the PPV (top row) and sensitivity (bottom row). The horizontal axes correspond to the tuning parameters specific to each class of methods; see Section 2.4: Materials and methods. For the Miso methods in column a), the tuning parameter is the threshold applied to module-specific p-values.For the local cuts in column b) the tuning parameter is the rejection value for the Duda-Hart test statistic. For the global methods in column c), the tuning parameter corresponds to a step in the graph process (see Materials and methods).
The puzzling behaviour of the single linkage PPV curve in column c) results from a late joining gene pair corresponding to a true biological relationship. Column d) summarizes the range of PPV and sensitivity values.

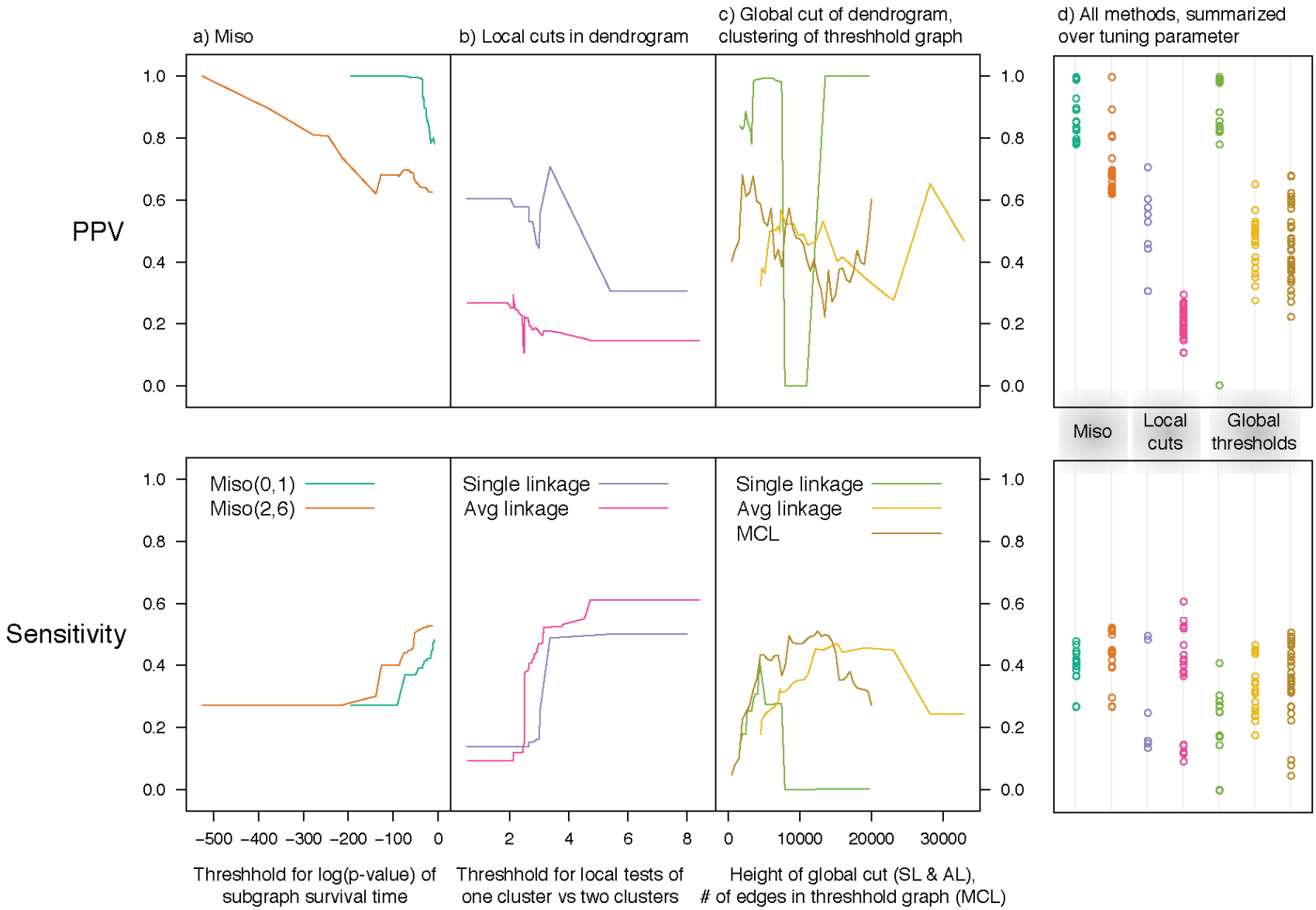
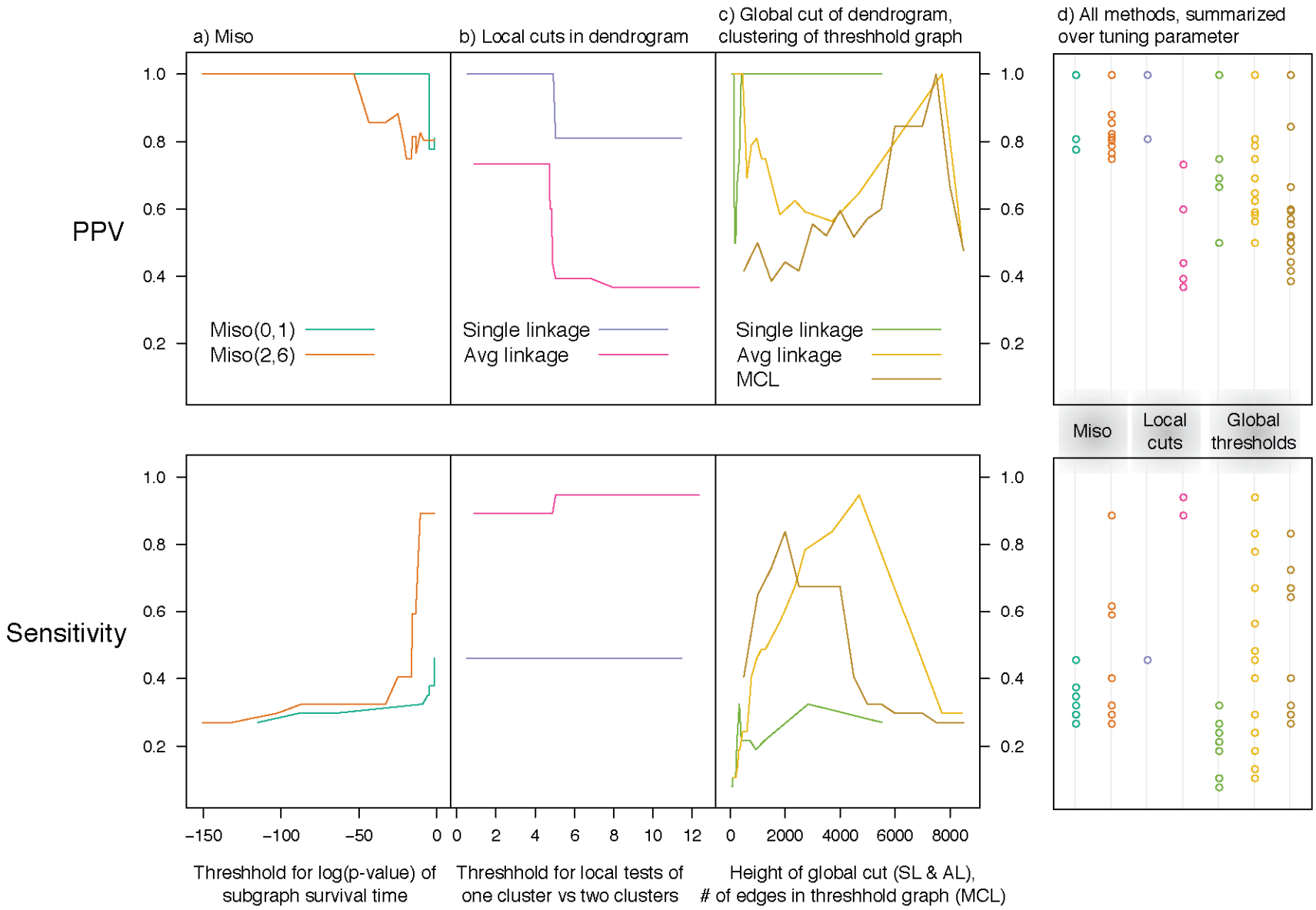

a) Miso
b) Local cuts in dendrogram
c) Global cut of dendrogram, clustering of threshhold graph
d) All methods, summarized over tuning parameter
PPV
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Miso(0,1)
Miso(2,6)
Single linkage
Avg linkage
Single linkage
Avg linkage
MCL
−500 −400 −300 −200 −100 0
2 4 6 8
0 10000 20000 30000
Threshhold for log(p-value) of subgraph survival time
Threshhold for local tests of one cluster vs two clusters
Height of global cut (SL & AL), # of edges in threshhold graph (MCL)
Miso
Local cuts
Global thresholds

Figure 2.5: Relative performance of module detection methods applied to yeast DNA damage response data.

Displayed are the PPV (top row) and sensitivity (bottom row). The horizontal axes correspond to the tuning parameters specific to each class of methods; see Section 2.4: Materials and methods. For the Miso methods in column a), the tuning parameter is the threshold applied to module-specific p-values, for the local cuts in column b) the tuning parameter is the rejection value for the Duda-Hart test statistic. For the global methods in column c), the tuning parameter corresponds to a step in the graph process (see Materials and methods).


a) Miso
b) Local cuts in dendrogram
c) Global cut of dendrogram, clustering of threshhold graph
d) All methods, summarized over tuning parameter
PPV
Sensitivity
Miso(0,1)
Miso(2,6)
Single linkage
Avg linkage
Single linkage
Avg linkage
MCL
Miso
Local cuts
Global thresholds
Threshhold for log(p-value) of subgraph survival time
Threshhold for local tests of one cluster vs two clusters
Height of global cut (SL & AL), # of edges in threshhold graph (MCL)

2.6 References

1. Beyer A, Bandyopadhyay S, Ideker T: **Integrating physical and genetic maps: from genomes to interaction networks**. *Nat Rev Genet* 2007, **8**(9):699-710.
2. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.
3. Kiemer L, Costa S, Ueffing M, Cesareni G: **WI-PHI: a weighted yeast interactome enriched for direct physical interactions**. *Proteomics* 2007, **7**(6):932-943.
4. Rinner O, Mueller LN, Hubalek M, Muller M, Gstaiger M, Aebersold R: **An integrated mass spectrometric and computational framework for the analysis of protein interaction networks**. *Nat Biotech* 2007, **25**(3):345-352.
5. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: **Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2007, **6**(3):439-450.
6. Van Dongen S: **Graph clustering by flow simulation**. *PhD Thesis.* Centre for Mathematics and Computer Science (CWI), University of Utrecht; 2000.
7. Hartuv E, Schmitt AO, Lange J, Meier-Ewert S, Lehrach H, Shamir R: **An algorithm for clustering cDNA fingerprints**. *Genomics* 2000, **66**(3):249-256.
8. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**:2.
9. Ling RF: **Probability theory of cluster analysis**. *J Am Stat Assoc* 1973, **68**(341):159-164.
10. Lee W, Onge RPS, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G: **Genome-wide requirements for resistance to functionally distinct DNA-damaging agents**. *PLoS Genet* 2005, **1**(2):235-246.
11. Schluter C, Lam K, Brumm J, Wu B, Saunders M, Stevens T, Bryan J, Conibear E: **Global analysis of yeast endosomal transport identifies the Vps55/68 sorting complex**. *Mol Biol Cell* in press.
12. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al*: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis**. *Science* 1999, **285**(5429):901-906.

13. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al*: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418**(6896):387-391.
14. Lee W, St Onge RP, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G: **Genome-wide requirements for resistance to functionally distinct DNA-damaging agents**. *PLoS Genet* 2005, **1**(2):e24.
15. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics* 2006, **7**:488.
16. Milligan GW, Cooper MC: **An examination of procedures for determining the number of clusters in a data set**. *Psychometrika* 1985, **50**(2):159-179.
17. Duda RO, Hart PE: **Pattern Classification and Scene Analysis**. New York: Wiley; 1973.
18. Huber W, Carey VJ, Long L, Falcon S, Gentleman R: **Graphs in molecular biology**. *BMC Bioinformatics* 2007, **8 Suppl 6**:S8.
19. Team TRDC: **R: A Language and Environment for Statistical Computing**; 2007.

Chapter 3: Stochastic block models for ranked relationships in genomics[2]

[2] A version of this manuscript will be submitted for publication. Brumm J, Wasserman WW, Bryan J: Stochastic block models for ranked relationships in genomics.

3.1 Background

Cells depend on genes (and their derived proteins) to perform functions such as signal transduction, intra-cellular transport and chromosome segregation. To achieve a function in a given cellular context, genes act in concert as part of gene modules; such gene modules are often protein complexes or molecular pathways. Finding gene modules and their member genes will remain an important task in genomics, in particular since the composition of modules can be different even in related cellular contexts.

An effective way to investigate gene modules is to collect data on the behavior of genes under conditions or treatments known to trigger the cellular context of interest. By measuring a gene-specific feature, such as a gene expression measurement or a loss-of-function mutant phenotype under each condition, a behavior profile of a gene is established. Genes in the same module often have similar gene profiles; conversely the similarity of gene profiles can be used to predict joint module membership.

The lack of known annotations of genes to modules makes methods for the unsupervised detection of modules, commonly referred to as clustering, important. Representing the data in the form of pairwise relationships, either as a matrix of pairwise similarity values or an adjacency matrix representing a graph has been used successfully in genomics [1-4]. In addition to its utility for clustering, the relational representation of data is natural for direct measures of functional relationship, such as protein-protein interactions and co-localization experiments. As we show in this paper, viewing gene profiling data within the relational paradigm allows for the straightforward incorporation of directly measured relational data such as protein-protein interactions into the analysis.

Relational data is often analyzed as a graph, where relationships between genes are either considered present or absent [2]. We have earlier introduced *graph processes* as a useful paradigm extending the graph-based search for gene modules (Chapter 2). In a graph process obtained from similarity values for gene profiles, edges are added between nodes (representing

genes) in the ranked order of similarity, starting with an empty graph and ending with a complete graph containing all edges between all genes.

Here we develop a data generating model of the graph process under the assumption that edges connecting genes within a module tend to arrive earlier in the graph process than edges between a genes in different modules (*module clustering assumption*). Assignments of genes to candidate modules (called *blocks*) that fulfill the module clustering assumption score a high probability, allowing us to search for a clustering solution by optimizing the likelihood function.

The likelihood is based on two key observations: 1) the candidate assignment of genes to blocks induces a labeling of edges as 'within block' and 'between blocks', and 2) the graph process of these labeled edges can be modeled as a standard stage-wise ranking (and scored with the corresponding likelihood) where within block edges are 'stronger' than between blocks edges. Likelihoods for rankings based on unobserved 'strength' parameters such as the ability of a horse to win a race have long been used [5-7]; typically the strength parameters are determined by analyzing multiple races. Here, the repeated placement of edges within and between modules allows their determination. The use of such a ranking model for the graph process is the key innovation of this paper.

The modeling of the module clustering assumption as a stochastic block model is inspired by methods for the analysis for binary data from the social sciences [8, 9]. Key to the stochastic block model, as for other latent class models, is the representation of the block label for each node as a random variable taking on values in a discrete set of block labels.

Our model is a principled approach to relational data analysis. Our method performs well in applications and is robust to noise, facilitates integration and gene-module assignments can be computed using standard optimization techniques. In this paper we lay the foundations for the stochastic block model for graph processes, a specific application will be considered in Chapter 4.

3.2 Results

3.2.1 A generative model for ranked relational data

To express our assumptions about gene modules in a model, we assume there is a true, unknown modular organization of the n genes under study, i.e. each gene belongs to exactly one of M underlying biological blocks. We refer to the hypothesized modules as blocks to distinguish them from the biologically defined gene modules, since some gene modules may not be recognizable in a given study. For a specific gene i, we refer to its block membership as its 'label', denoted by $c(i)$, where $i = 1, 2, \ldots, n$ and $c(i) \in \{1, 2, \ldots, M\}$. The complete organization of genes in blocks or, equivalently, the complete collection of gene labels, $c = (c(1), c(2), \ldots, c(n))$, is the parameter we wish to estimate. That is, we wish to find the (latent) gene labeling that is most compatible with the observed ranked relational data. In order to specify a likelihood for the observed data (ranks of gene-gene relationships) in terms of the parameter of interest (the latent gene labels), we employ a novel combination of stochastic block model theory from social network analysis and stage-wise ranking models initially developed for applications such as horse racing.

In a stage-wise ranking model, each item to be ranked is often assumed to have an underlying 'strength' parameter [5, 10], such as the propensity of a racehorse to win races. In our application, these items are the $N = n(n-1)/2$ edges that comprise the edge set of the complete graph on the n genes (nodes) and the strength of edge k, denoted $\lambda_k(c)$, is fundamentally determined by the modular structure captured by the labels c (detailed below). We define $\lambda(c) = (\lambda_1(c), \ldots, \lambda_N(c))$. The likelihood of the ranks can be constructed by considering a sequential selection procedure, hence the name 'stage-wise ranking model'. At the first stage, we will choose one item (edge, in our case) from the full set (N possible edges), where the choice probability of item k is given by

$$P(k \mid \lambda(c)) = \frac{\lambda_k(c)}{\sum_{r=1}^{N} \lambda_r(c)}.$$

The chosen item is then removed from all subsequent stages of the procedure.

The set of the not-yet-selected items (edges) will be denoted by S and, in a manner similar to the initial selection, we carry out a sequential selection procedure, in which one item is selected at each stage and which ultimately results in an observed ranking of the items. The general choice probability of item k, if it is contained in S, is given by

$$P_S(k \mid \lambda(c)) = \frac{\lambda_k(c)}{\sum_{r \in S} \lambda_r(c)},$$

and is zero otherwise. To continue the race analogy, a horse's probability of beating the other horses in S is given by its strength parameter relative the total strength represented in S. We assume that our ranked relational data arises from this type of stage-wise ranking model that is driven by the choice probabilities or, equivalently, by the underlying item strengths.

An observed ranking of the gene-gene edges, such as the observed graph process, is denoted by $\pi = (\pi_1, \ldots, \pi_N)$, meaning that $\pi_w \in \{1, \ldots, k, \ldots, N\}$ is the index of the edge with rank w. Observed tied ranks are broken at random. If we denote the set of edges remaining at stage l by S_l, the likelihood for the observed ranked relationships is

$$P(\pi \mid \lambda(c)) = P_{S_1}(\pi_1 \mid \lambda(c))\, P_{S_2}(\pi_2 \mid \lambda(c)) \ldots P_{S_{N-1}}(\pi_{N-1} \mid \lambda(c))\, P_{S_N}(\pi_N \mid \lambda(c))$$

$$= \frac{\lambda_{\pi_1}(c)}{\sum_{i=1}^{N} \lambda_{\pi_i}(c)} \; \frac{\lambda_{\pi_2}(c)}{\sum_{i=2}^{N} \lambda_{\pi_i}(c)} \; \cdots \; \frac{\lambda_{\pi_{N-1}}(c)}{\lambda_{\pi_{N-1}}(c) + \lambda_{\pi_N}(c)} \; \frac{\lambda_{\pi_N}(c)}{\lambda_{\pi_N}(c)}. \tag{3.1}$$

This likelihood is invariant under multiplication of $\lambda(c)$, that is

$$P(\pi \mid \lambda(c)) = P(\pi \mid \gamma\, \lambda(c)) \tag{3.2}$$

for any constant $\gamma > 0$.

Our most fundamental assumption is that $\lambda_k(c)$, the strength of the edge k, depends only on the labels of the two associated nodes. That is, if edge k connects the two nodes i and j, $\lambda_k(c)$ depends only on $c(i)$ and $c(j)$. As a result of our assumption that edge strength is solely determined by the module membership of the associated genes, the entire collection of edge strengths are generated by the entries of an M by M matrix Λ, where $\Lambda_{l,m}$ is the edge-strength parameter for edges connecting the block with label l to the block with label m, and we assume that $\Lambda_{l,m} = \Lambda_{m,l}$.

In this paper we will only use the *homogeneous block model*, in which all diagonal elements of Λ (corresponding to the strength of 'within block' edges) are equal to λ_W, and all off-diagonal elements (corresponding to the strength of 'between blocks' edges) are equal to λ_B. We assume, in accordance with the module clustering assumption, that $\lambda_W > \lambda_B$. Recall that the likelihood is invariant to multiplication of $\lambda(c)$ by a positive constant. This implies that, without loss of generality, we can redefine λ_W to be the ratio of within and between edge strengths, i.e. $r = \lambda_W / \lambda_B$, and assume that λ_B is equal to 1. With respect to edge strength parameters, the likelihood now only depends on r.

Note that in this model-specification, the parameters $\lambda_k(c)$ depend only on the n by n matrix $B(c)$, with $B_{ij}(c) = 1$ if $c(i) = c(j)$ and $B_{ij}(c) = 0$ otherwise. This matrix simply records if an edge is within a block or between blocks, and is a different way to express the influence of c on the likelihood which is convenient for incorporating prior relational information, as we show below. The matrix B is the adjacency matrix for the graph that represents the block structure, where edges are present between all nodes in the same block and no edges are present between blocks. In summary, we can express the likelihood as

$$P(\pi \mid \lambda(c)) = P(\pi \mid B(c), r).$$

3.2.2 Bayesian estimation for the block model

We analyze the problem using the Bayesian paradigm. The Bayesian approach allows for the incorporation of prior information, which is often available in genomic studies. Furthermore,

sequential conditional updating algorithms like Gibbs sampling can be used to resolve the problem of estimating the labels [8].

3.2.2.1 Incorporating prior information

The stochastic block method can be used effectively without prior information, as we show below; here we show how to incorporate relational data as prior information into the analysis. An application of the stochastic block model using prior information is given in Chapter 4. We restrict our attention to prior information in the form of binary relational data (such as protein-protein interactions). It is convenient to represent this type of prior relational data by an adjacency matrix A where $A_{ij} = 1$ if genes i and j are related in the prior data and zero otherwise. The key to incorporating this prior information into the analysis is the adjacency matrix B, representing the block-structure induced by the labeling c. It is now possible to 'overlay' these two graphs and evaluate which edges corresponding to the block-graph represented by B are present in the prior graph represented by A.

We assume that the prior probability for two genes having the same label, conditional on A, is given by

$$P(c(i) = c(j) \mid A_{ij} = 1) = P(B_{ij}(c) = 1 \mid A_{ij} = 1) = \alpha\,, \tag{3.3}$$

$$P(c(i) = c(j) \mid A_{ij} = 0) = P(B_{ij}(c) = 1 \mid A_{ij} = 0) = \beta\,. \tag{3.4}$$

To simplify formulas, we define the following counts for the number of concordant and discordant pairs between prior information and the relationships induced by the candidate labeling: $s_{11} = \#\{ i<j: c(i) = c(j), A_{ij} = 1\}$, $s_{01} = \#\{ i<j: c(i) \neq c(j), A_{ij} = 1\}$, $s_{10} = \#\{ i<j: c(i) = c(j), A_{ij} = 0\}$ and $s_{00} = \#\{ i<j: c(i) \neq c(j), A_{ij} = 0\}$, so that s_{11} is the number of gene pairs in the same block ($B_{ij}(c) = 1$) that are a priori related ($A_{ij} = 1$), s_{01} the number of gene pairs in different blocks ($B_{ij}(c) = 0$) that are a priori related ($A_{ij} = 1$) and so on. We also define $s_0 = \#\{ i<j: A_{ij} = 0\}$, the number of gene pairs a priori unrelated.

We assume that a priori all relationships are independent. With this assumption, the prior probability of the adjacency matrix *B(c)* of a labeling *c* is given by:

$$P(B(c) \mid A, \alpha, \beta) = \prod_{i<j} B_{ij}(c)\, \alpha^{A_{ij}} \beta^{1-A_{ij}} + [1 - B_{ij}(c)](1-\alpha)^{1-A_{ij}} (1-\beta)^{A_{ij}}$$

$$= \alpha^{s_{11}} (1-\alpha)^{s_{01}} \beta^{s_{10}} (1-\beta)^{s_{00}} .$$

The parameter α would typically be chosen close to 1, if a prior relationship represents strong evidence towards a true functional relationship given what is known about the assay used to derive the prior information. If coverage of these prior relationships is low, that is, only few relationships between genes in the same module are observed, (as in the case of protein-protein interactions), we set β to 0.5. In the latter case, the above equation simplifies to

$$P(B(c) \mid A, \alpha, \beta = 0.5) = \alpha^{s_{11}} (1-\alpha)^{s_{01}} 0.5^{s_0} ,$$

where the last factor is independent of the labeling *c*.

3.2.2.2 Estimating the labels using the Gibbs sampler

Our Bayesian inference relies on the joint posterior distribution of the parameters, given the data. The joint posterior distribution for the latent labels is proportional to the product of the likelihood and prior distribution:

$$P(c \mid \pi, r, A, \alpha, \beta) \propto P(\pi \mid B(c), r) P(B(c) \mid A, \alpha, \beta) . \qquad (3.5)$$

This posterior distribution is used to estimate the gene labels. To obtain samples from the posterior distribution, we use the well-known Gibbs sampling procedure. This procedure starts with random values for the gene labels and then updates the label for each gene *c(i)* in turn according to the following conditional probabilities. For any $m \in 1,2, \ldots, M$, this update probability is determined through the relationship

$$P(c(i) = m \mid \pi, c[-i], r, A, \alpha, \beta) = \\ P(\pi \mid B(\tilde{c}), r)\ P(B(\tilde{c}) \mid A, \alpha, \beta)\ K. \tag{3.6}$$

Here *c[-i]* is the vector of labels except for *c(i)*, $\tilde{c}$ refers to the labeling where *c(i) = m* (and all other values are unchanged) and *K* is the likelihood integrated over all possible values for *c(i)* (and hence is independent of *m*). Equation (3.6) allows us to compute the update probabilities for the vector *c(i)*, as these are multinomial probabilities, so that the probability that *c(i) = m* can be computed using the ratio

$$\frac{P(c(i) = m \mid \ldots)}{\sum_{l=1}^{M} P(c(i) = l \mid \ldots)} \tag{3.7}$$

even though *K* is unknown.

The updated gene label is drawn from the updated multinomial distribution for *c(i)*. Updating each gene label in turn and iterating through these updates yields samples for each gene label; these samples are approximately distributed like samples for the gene labels from the joint posterior probability distribution (3.5) for all genes. We call one round of updating of all gene labels one 'Gibbs iteration'.

3.2.2.3 Co-labeling probabilities as key parameters

The probability $p_{ij} = P(\ c(i) = c(j)\ )$ that nodes *i* and *j* are in the same block (referred to as co-labeling probabilities) are useful parameters for our method. The block labels are not identified in our model in the sense that an arbitrary switching of labels leads to the same value for the likelihood. It is well-known that if the labels are not identified, the Gibbs sampler might visit more than one of these equivalent labelings (this is often described as 'label-switching' [11]). The co-labeling probabilities are invariant under relabeling, facilitating the interpretation of the output of the algorithm [9]. The co-labeling probability p_{ij} can be estimated by the co-labeling rate of nodes *i* and *j* in the Gibbs iterations.

Aside from these technical considerations, the estimated co-labeling probabilities are very useful in applications. As they provide a confidence score for each gene-gene relationship, they allow for the prioritization of the output of our procedure for follow-up experiments, in which the pairs of genes with the highest estimated co-labeling probability are investigated for co-membership in the same gene module first.

3.2.3 Adjusting the likelihood

The block structure imposed by the labeling c determines the types of items available for the ranking (how many edges are within blocks, how many between blocks). As we show here, the number of within and between edges in the collection of items influences the magnitude of the likelihood for the ranking. Since we compute the likelihood with varying c, we need to consider this effect.

It is apparent from the formula (3.1) that the numerator of the likelihood does not depend on the order of the edges, but only on the number of within and between edges. As $\lambda_W > \lambda_B$, the numerator is maximized by taking the node labeling resulting in the maximal number of within edges, leading to an inherent tendency to re-assign a gene from a small block to a large block. This tendency leads to the accumulation of many genes in a giant block during the Gibbs iterations, except for genes in blocks that are well-separated from other blocks.

To counteract the influence of the numbers of items available, we recognized that only the relative magnitude of the likelihood is important in determining the ratio (3.7) for updating the label of a gene. If two labelings lead to approximately the same number of within and between edges, the numerators of the likelihood (3.1) cancel out approximately in the ratio. Consequently, we adjusted the likelihood by only calculating the denominator of (3.1); that is, we used

$$\frac{1}{\sum_{i=1}^{N}\lambda_{\pi_i}(c)}\frac{1}{\sum_{i=2}^{N}\lambda_{\pi_i}(c)}\cdots\frac{1}{\lambda_{\pi_{N-1}}(c)+\lambda_{\pi_N}(c)}\frac{1}{\lambda_{\pi_N}(c)}$$

instead of the likelihood in the computation of (3.7).

To investigate the impact of size of the blocks on the conditional marginal probabilities (3.6) (the probability of assigning a given node to the respective blocks) using either the original or the adjusted likelihood, we used the well-annotated CPY data introduced in Chapter 2 (see Section 3.2.5.2 below). To get a biologically accurate labeling, we labeled each gene with its biological annotation, making each block correspond to a gene module. We illustrate the impact of the adjustment using a gene (called gene X here; other genes could have been used for the illustration as well) from the COG/YPT6 module; this module fulfills the module clustering assumption well. Genes of unknown function were assigned to the same block; most genes in this block correspond to mutant strains with profiles similar to the wild-type strain profile. We then compute the update probabilities for the given gene X for each block; Figure 3.1A shows the size of the block plotted against the rank of the corresponding update probability using the original and adjusted likelihood. Using the original likelihood, the update probability with the highest rank is for the assignment of gene X to the largest block (corresponding to not annotated genes).

Ideally, the update probability for the assignment of gene X to the COG/YPT6 block would be highest. Figures 3.1B and 3.1C explore the relationship between update probability and the edge-ranks from gene X to each block in more detail. In both plots, the ranks of the edges from each block to gene X are given along the horizontal axis and the blocks are ordered from highest update probability (top) to lowest (bottom). While the early arrival of edges from the block to gene X generally leads to greater update probabilities (as desired), the update probability is sensitive to differing block sizes, rewarding small blocks in the case of the adjusted likelihood and large blocks in the case of the original likelihood.

We found that in the Gibbs sampling procedure, using the adjusted likelihood led to a vastly superior performance as measured by the ability of the algorithm to recover known gene modules. Hence all of our results are based on the adjusted likelihood.

3.2.4 Impact of tuning parameter selection on performance

Two key parameters that influence the performance of the algorithm are the number of blocks, *M*, and the ratio of within-block to between block edge strengths *r*. We could not find suitable estimators for these parameters (see Discussion). We also tried the internal quality score suggested in [9] (called *Hx* there) for selecting *M* or *r*, but found that the value of *Hx* is unfortunately of no help in choosing *M* (*Hx* decreased with *M*), and tends to select large values of *r* (in the analyses presented in Figure 3.4, the lowest value of *Hx* corresponds to a value of $r=100$). Since we could not find a way to automatically choose values for *r* and *M*, we instead treat them as tuning parameters to be chosen by the user. This section explores the impact of the tuning parameter choices on the estimators of the co-labeling probabilities.

To judge the impact of various parameter choices, we analyzed the CPY data with varying parameter choices. To evaluate the results from our procedure, we took advantage of the known membership of genes in biological modules. Since this data is well annotated, a comparison between the block structure revealed by our procedure and the true biological module structure can be used to judge the quality of the derived block structure as well as the impact of parameter choices.

We determined the number of true positive (TP) gene pairs (both members of the pair are in the same block that and are known to be in the same gene module), the number of false positive (FP) gene pairs (both members of the pair are in the same block that but they are not known to be in the same gene module) and the number of false negative (FN) gene pairs (the members of the pair are in different blocks but they are known to be in the same gene module). We used the positive predictive value PPV (= TP/(TP+FP)), the rate of correct predictions in all predicted interactions) and the sensitivity (= TP/(TP+FN), the percentage of available protein-protein interactions recovered). We prefer the PPV over alternative measures, because it measures the success rate of the scientists in their follow-up experiments (that are only performed on pairs of genes predicted to be in the same module; there are TP + FP such pairs). To obtain predictions, we applied a threshold to the estimated co-labeling probabilities for the gene pairs, classifying a pair as within-block if its estimated co-labeling probability was above the given threshold. We

varied the threshold from stringent (=1) to weak (=0) and recorded the PPV and sensitivity at each threshold.

We first evaluated the impact of the initial labelings on performance. Figure 3.2 shows the difference in performance for repeated runs of the procedure, which was started from a random label assignment. Figure 3.2A shows a case with unusual differences ($M=20$ and $r=10$) and Figure 3.2B a case with differences typically observed for other parameter settings ($M = 60$ and $r = 5$). As expected, the impact of the initial labeling is generally higher for low sensitivity (corresponding to stringent cutoffs for p_{ij}) and lower for cutoffs leading to higher sensitivity, because stringent cutoffs lead to few predicted co-labeled gene pairs.

Next we evaluated what impact the choice of M has on the resulting block assignments. Figure 3.3 shows selected runs of the algorithm for $r = 10$. It shows that the performances are comparable, if M is chosen large enough. We recommend to choose M large compared to the number of modules expected to be relevant in the data, keeping in mind that the running time of the procedure increases considerably with M (see Discussion).

The impact of the choice of r for given M is shown in Figure 3.4. Figure 3.4 A shows the results for $M = 40$ and Figure 3.4 B for $M = 60$. Intermediate choices as values for r delivered the best performance. We found that for this dataset, large values of r rewarded internally cohesive, but not necessarily externally isolated groups of genes. This leads to results with a considerably lower PPV, because subsets of larger gene groups get identified as gene modules, in particular from the set of gene profiles similar to wild-type strain profile.

3.2.5 Application to yeast mutant phenotype data

The methodological developments presented here were motivated by collaboration with the Conibear lab at the University of British Columbia, which aims to elucidate the mechanisms of vesicle transport in the cell. This machinery is used by the cell to facilitate regulated transport of chemicals or proteins within the cell. To study the gene modules relevant in the vesicle transport system, gene profiles in the form of yeast mutant phenotypes were obtained, using a collection of yeast strains with individual genes disabled [12]. Besides data produced in their

own lab, the Conibear lab also uses publicly available repositories of yeast mutant profiles. We apply our method to a subset of data derived in [13]. This data, measuring the growth of yeast mutants across a wide spectrum of drugs, is referred to as the CHS6 data here.

3.2.5.1 The stochastic block model performs well in noisy situations

The CHS6 data motivated the development of the stochastic block method, since it proved too noisy for our previously developed isolation index (Chapter 2). The graph process stopped after 315 steps (Figure 3.5) shows that the data is too noisy for the isolation index to uncover the important AP1 and Escrt gene modules. Genes in these modules (highlighted in color) are, even at this early stage in the graph process, already well-connected to genes outside the module. In this case the number of between-module connections exceeds the noise tolerance of the Miso method (Chapter 2), even of the Miso(2,6) method.

The stochastic block model succeeded in identifying suitable candidate modules even in this noisy data (Table 3.1). Of the six candidate modules with more than three genes, four correspond well to relevant gene modules. The AP1 candidate module (size 17), which contains 5 out the 7 known AP1 genes, contains several not yet annotated genes of interest for follow-up studies by the Conibear lab, particularly as some show promising behaviour in other assays (data not shown). The Escrt candidate module contains high proportions of relevant genes, known to be involved in vesicle transport. Note that the given list is almost the complete output of our procedure (small clusters are listed in the Table caption); no annotation-guided selection needed to be performed.

3.2.5.2 The stochastic block model identifies well-isolated modules

Our previous analysis of the CPY data using the Miso method (Chapter 2) was able to identify gene modules effectively as the relevant modules were well-separated, leading to a graph process with little noise. To assess the performance of the stochastic block method in a low-noise setting, we re-analyzed the CPY data. The results, given in Table 3.2, are comparable to the results we obtained earlier with the Miso method (Tables 2.1 and 2.2), recovering the

relevant V-ATPase, SWR-C, Retromer, and Escrt gene modules well. In particular, the output includes the gene module validated in [14]. These results show that the stochastic block method offers excellent performance in this situation with low noise.

3.2.5.3 Comparison to threshold graph clustering

We used the CPY data to assess the performance of our method. The CPY data is well-annotated, making it useful for method evaluation. We compared the performance of the stochastic block model, which is based on the graph process, to the collection of clusterings that can be obtained by clustering a chosen sequence of individually clustered threshold graphs. As a graph clustering method, we chose MCL [15] which has been shown to be an effective method [16]. We created threshold graphs every 500 steps up to 20000 edges; each such graph was clustered using MCL. We eliminated clusters of size greater than 50 from consideration, since they would not be used in an applied situation and unduly lower the PPV. For the stochastic block method, we varied the cutoff for the estimated co-labeling probabilities from stringent (threshold = 1) to weak (threshold = 0).

The estimated co-labeling probabilities are useful for the prioritization of follow-up experiments, since the sensitivity decreases and the PPV increases with increased probability threshold (Figure 3.6). The performance of the MCL-based method, on the other hand, depends on the choice of a good threshold graph, which is difficult (see Chapter 2 for more on the challenge of threshold selection). Overall, choosing a cutoff of about 0.4 for the co-labeling probabilities in the stochastic block model delivers performance comparable to the best available MCL-based procedure (Figure 3.6).

3.3 Discussion

This paper demonstrates the utility of graph processes as a useful data structure and offers an algorithm to predict functional relationships. Our method works in difficult situations with noisy data while retaining excellent predictive performance in situations with low noise. The co-labeling rates delivered by the Gibbs sampling procedure offer confidence values for

predictions of individual functional relationships that are useful for the prioritization of follow-up experiments. These rates could also be used as input to methods integrating across different platforms for gene function prediction [17].

Our probabilistic approach allows us to incorporate prior information in the form of pairwise relationships, such as protein-protein interactions. This is more difficult to achieve in a feature-space representation of the data (used by algorithms such as *k*-means), where each gene profile is represented as a point in a high-dimensional space. The utility of our method using prior information will be presented in Chapter 4.

Our likelihood for the ranking can be alternatively viewed as a Thurstonian model, which assumes that the ranking is obtained through the ordering of values of underlying, unobserved random variables with continuous values [18]. Our model can be obtained by assuming that the distribution of these latent random variables is exponential [6], revealing a potential limitation, since for exponential random variables the mean is equal to the standard deviation. This implies that if late edges (which tend to be edges between blocks) correspond to large values of the latent variables, the variability of the latent values corresponding to between-blocks edges is also large, an assumption that may not be realistic for the data. However, if a different distribution (such as a normal) is assumed for the Thurstonian model, the likelihood for the ranking is a multi-dimensional integral, making it more difficult to evaluate.

An evaluation of the procedure on data other than the CPY data would be desirable. However, experimental data often suffers from the lack of annotations, making evaluation difficult. Simulation studies, on the other hand, require the user to assume that the model used to generate the data is an appropriate reflection of the structure observed in experimental data of interest, an assumption that is often rejected by applied scientists, making the results hard to communicate to the potential users of the methods.

Automatic determination of the model parameters r and M proved difficult. Although it is possible to estimate r with the maximum likelihood estimator when the correct labels are known, we found that the maximum likelihood estimator performs poorly when the labels are incorrect, pushing the ratio r towards 1 during the Gibbs iterations; a value of r close to 1 leads to poor clustering results. The model selection statistic Hx did not perform very well on the

CPY data but, at least for choosing r, may work better in datasets where internal cohesion of clusters is the dominating feature. Choosing M automatically is less important in our method (in contrast to true partitioning methods like k-means, for example), since the ultimate clustering solution does not seem to be affected severely as long as M is chosen large enough.

A limitation of our method is the computational complexity. In its current form, each likelihood evaluation for n genes is on the order of $O(n^2)$ (meaning that there is a constant z such that the worst case performance in terms of computing time can be bounded above by $z\, n^2$). Each Gibbs step is hence of order $O(n\, M\, n^2)$, since each node label is evaluated for each of the M available blocks. This high demand on computing time limits the possibilities for Gibbs convergence diagnostics and limits the number of genes we may consider (our biggest example is the CHS6 data with n=329, for which, with M = 40, 120 Gibbs steps took several hours to complete on a regular computer with a 2 GHz CPU and 1 GB of RAM). The computational complexity of the algorithm may be reduced by using an approximation to the likelihood, as most informative events in the graph process happen quite early (the edges appearing at late stages of the process are typically only between blocks).

In general, the Gibbs sampling procedure needs to be monitored to assess convergence. This is not a simple problem in general, and we may not be able to guarantee convergence of our sampling procedure for our method, since the computational complexity for our method prohibits many Gibbs iteration steps. In our applications, starting from a random labeling, the Gibbs sampler converged rapidly towards a clustering solution close to the 'true' module assignments constructed from known annotations. In difficult situations, it may be necessary to start the Gibbs sampler from a labeling reasonably close to a desired solution. One possibility would be to construct a complete or average linkage clustering and extract M clusters from the corresponding tree as a starting labeling of the nodes.

Thresholding of the co-labeling rates may sometimes be desired, if a list of predictions rather than a ranking of follow-up experiments is desired. Finding a suitable threshold was difficult because the estimated co-clustering probabilities are severely affected by M and to a lesser degree by the choice of r.

The focus of the current paper is to introduce the stochastic block model for graph processes and establish its utility and limitations. A more detailed application to the integrated analysis of yeast mutant phenotype profiles and protein-protein interactions will be presented in Chapter 4. A possible extension of the stochastic block model is to use the block labels to integrate multiple gene profile data sets. Future work also includes the generalization of the simple block structure used in this paper; this is particularly attractive since the co-labeling rates are a flexible output which may accommodate structures such as a single gene belonging to more than one functional module which is often the case in biological systems.

3.4 Materials and methods

3.4.1 Data sources

The CPY data was introduced in Chapter 2 and [14]. This data set contains 279 yeast mutant profiles that were normalized and converted into pairwise similarities using Euclidean distance.

The CHS6 data was extracted from [13], where yeast mutant strains were exposed to 82 conditions; a subset of 329 genes was selected for showing a strong phenotype in an initial screen performed in the Conibear lab. We converted the profiles into pairwise similarities using the commonly used measure of $(1 - C)$ as the similarity score (where C is the Pearson correlation), after dividing each value by the standard deviation of the values for all 329 genes in the given condition.

3.4.2 Gibbs sampling

The Gibbs sampler was always initialized with a random labeling. All results presented here are based on 110 iterations; the first 10 iterations were discarded when we computed the co-labeling rates (the number of times a pair was in the same block, divided by the number of Gibbs sampling iterations)

3.4.3 Formula for Hx

The quantity for model assessment used by [9], adapted for our situation with undirected graphs, is

$$Hx = \frac{2}{n(n-1)} \sum_{i<j} \hat{p}_{ij}(1-\hat{p}_{ij}).$$

The motivation given in [9] is that if *Hx* is small, for each pair it is clear if they are in the same block ($\hat{p}_{ij}$ approximately equal to *1*) or not ($\hat{p}_{ij}$ approximately equal to 0), indicating a strong clustering structure.

3.4.4 Clusters derived from estimated co-labeling probabilities

After thresholding the estimated co-labeling probabilities, the gene-pairs predicted to be in the same block can be viewed as edges in a graph. If the the threshold is chosen stringent, the graph typically has several unconnected components. We use these components as clusters to summarize the results succinctly.

3.5 Tables and figures

3.5.1 Tables

Table 3.1: Results for the CHS6 dataset with the following parameter settings:

co-labeling rate threshold = 0.5, M = 40, r = 10 (see "Materials and methods" on how to obtain the clusters from estimated co-labeling probabilities). Not listed are 2 clusters of size 3 and 7 clusters of size 2 with unremarkable composition. The most notable gene modules found in a given candidate module are listed under 'Remarks' with the number of genes in the candidate module and the number of genes in the module given in parentheses. The table is ordered to show the most remarkable clusters on top.

Cluster	Size	Remarks
AP1	17	AP1 (5/7)
Escrt	17	Escrt (4/7), Class B/C VPS(4/9), Retromer (3/4), ClassD VPS(3/5), AP3 (2/2)
DNA	7	DNA (4/8)
Rho	5	Rho (2/2)
UB	3	UB (2/8)
Garp/YPT6	2	Garp/YPT6 (2/5)
?	10	Mixed annotations
Orfs	8	Mostly uncharacterized

Table 3.2: Clustering results for the CPY data (threshold of 0.5 for co-labeling rates) for *r=5*, *M=40*. The most notable gene modules found in a given candidate module are listed under 'Remarks' with the number of genes in the candidate module and the number of genes in the module given in parentheses. The table is ordered to show the most remarkable clusters on top. Not listed are clusters of size 2. The VPS 55-68 candidate module was validated in [14].

Cluster	Size	Remarks
V-ATPASE	14	V-ATPASE (14/20)
ESCRT	7	ESCRT (7/14)
SWR-C	7	SWR-C (7/9)
COG/YPT6	7	COG (3/5), YPT6 (4/5)
RETROMER	6	RETROMER (6/6)
VPS 55-68	3	VPS 55-68 (2/2),
V-ATPASE	3	V-ATPASE (3/20)
ClassD	3	ClassD VPS(3/6)
PI3KC	3	PI3KC (2/4)

3.5.2 Figures

Figure 3.1. Influence of size of the blocks on update probabilities using the original and adjusted likelihoods.

Figure A shows the rank of the update probabilities plotted against the size of the respective block. B and C show the ranks of edges hitting the gene in the COG/YPT6 module displayed along the x-axis; the y-axis splits the ranks by module. The plots are ordered by the update probability of assigning the gene to the group, with the first line corresponding to the group with the highest likelihood (so ideally the COG module would occupy the first line of the plot). B shows the edge-ranks in the order of update probabilities using the original likelihood; C in the order corresponding to the adjusted likelihood.

Figure 3.1A

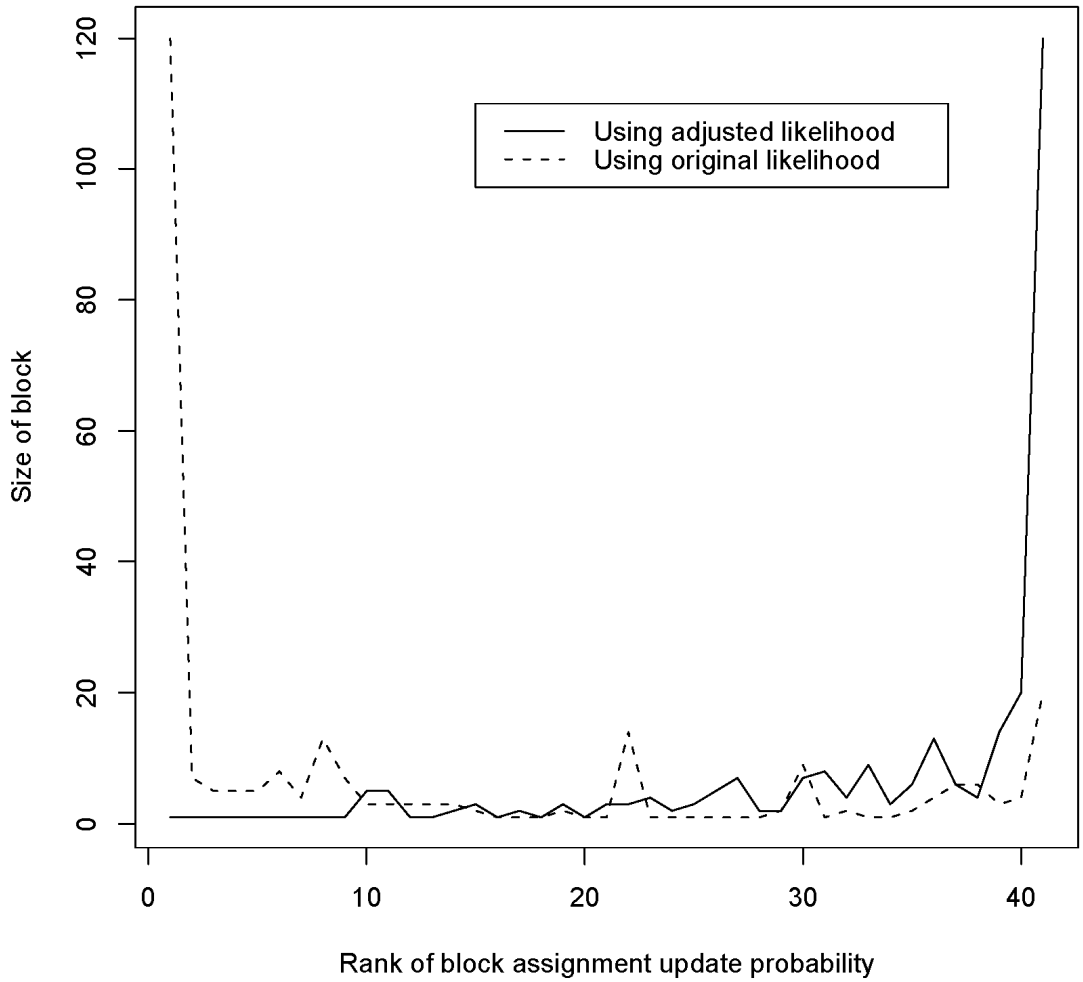

Using adjusted likelihood
Using original likelihood
Size of block
0
20
40
60
80
100
120
0
10
20
30
40
Rank of block assignment update probability

Figure 3.1B

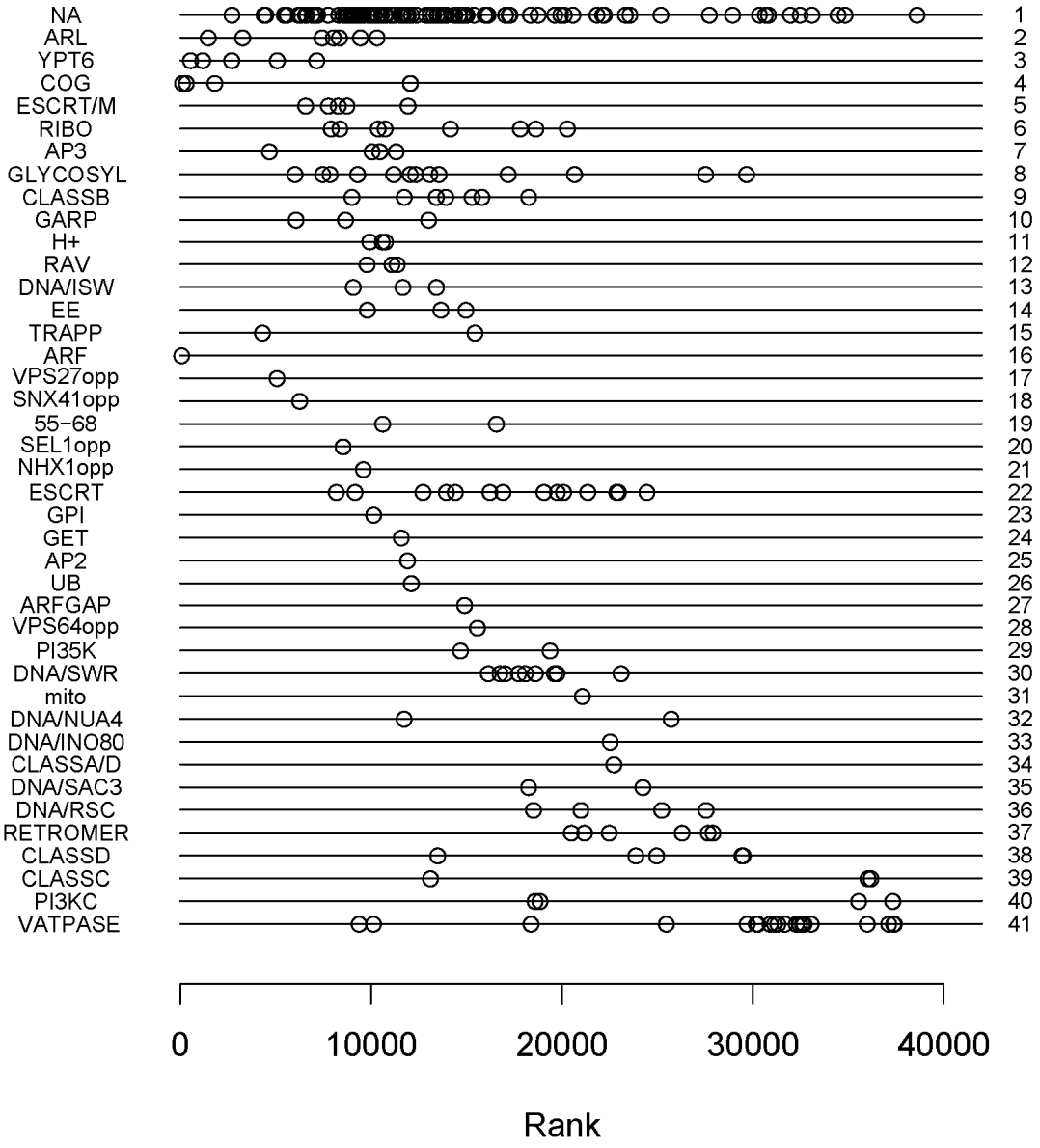

NA 1
ARL 2
YPT6 3
COG 4
ESCRT/M 5
RIBO 6
AP3 7
GLYCOSYL 8
CLASSB 9
GARP 10
H+ 11
RAV 12
DNA/ISW 13
EE 14
TRAPP 15
ARF 16
VPS27opp 17
SNX41opp 18
55−68 19
SEL1opp 20
NHX1opp 21
ESCRT 22
GPI 23
GET 24
AP2 25
UB 26
ARFGAP 27
VPS64opp 28
PI35K 29
DNA/SWR 30
mito 31
DNA/NUA4 32
DNA/INO80 33
CLASSA/D 34
DNA/SAC3 35
DNA/RSC 36
RETROMER 37
CLASSD 38
CLASSC 39
PI3KC 40
VATPASE 41
0
10000
20000
30000
40000
Rank

Figure 3.1C

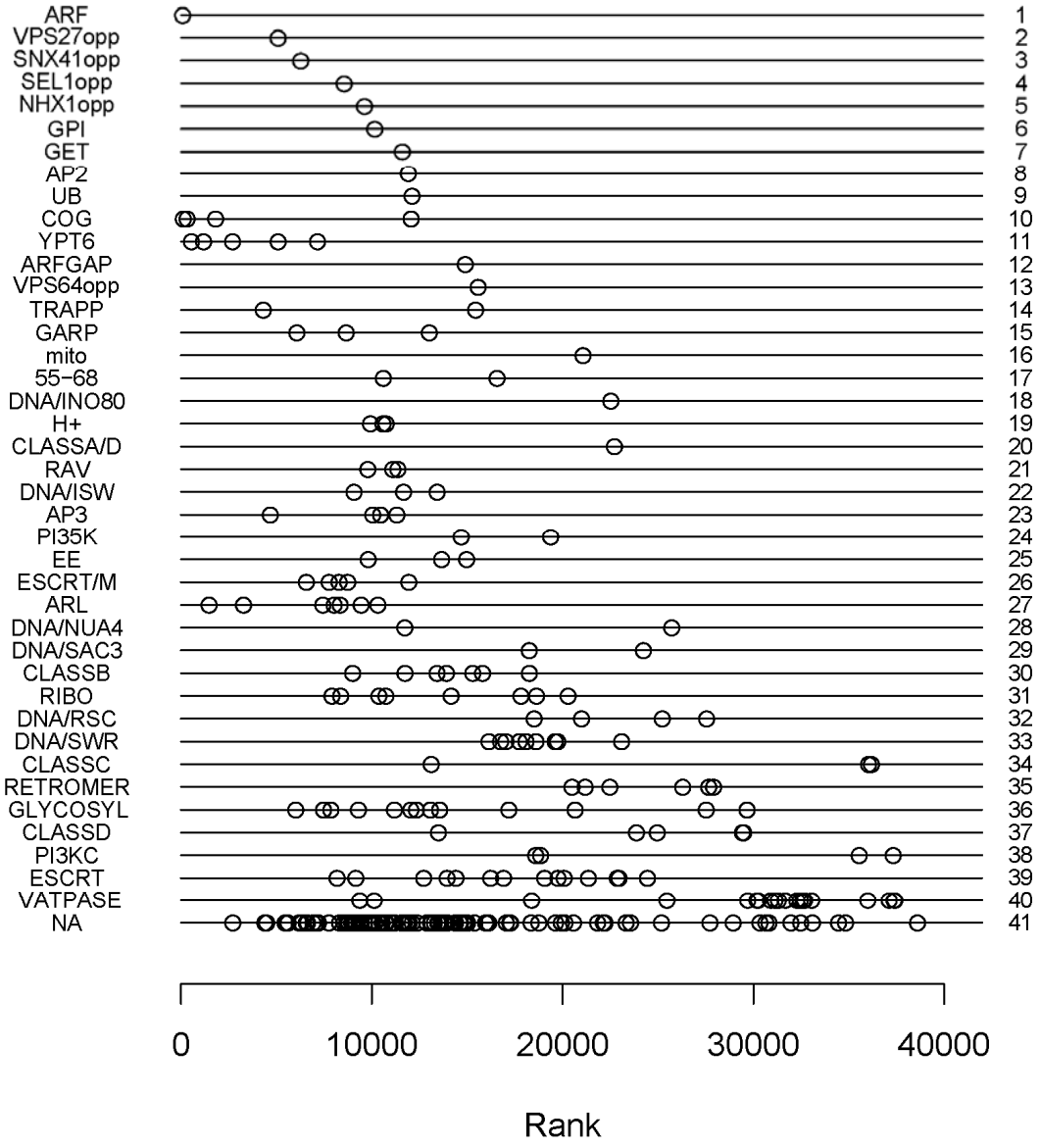

ARF 1
VPS27opp 2
SNX41opp 3
SEL1opp 4
NHX1opp 5
GPI 6
GET 7
AP2 8
UB 9
COG 10
YPT6 11
ARFGAP 12
VPS64opp 13
TRAPP 14
GARP 15
mito 16
55−68 17
DNA/INO80 18
H+ 19
CLASSA/D 20
RAV 21
DNA/ISW 22
AP3 23
PI35K 24
EE 25
ESCRT/M 26
ARL 27
DNA/NUA4 28
DNA/SAC3 29
CLASSB 30
RIBO 31
DNA/RSC 32
DNA/SWR 33
CLASSC 34
RETROMER 35
GLYCOSYL 36
CLASSD 37
PI3KC 38
ESCRT 39
VATPASE 40
NA 41
0
10000
20000
30000
40000
Rank

Figure 3.2: Impact of initial labeling on results.

The PPV (= TP/(TP+FP)) is plotted against sensitivity for repeated runs of the procedure from random starting points with the same parameter settings. Settings are: $M = 20$ and $r = 10$ for A and $M = 60$ and $r = 5$ for B.

Figure 3.2A

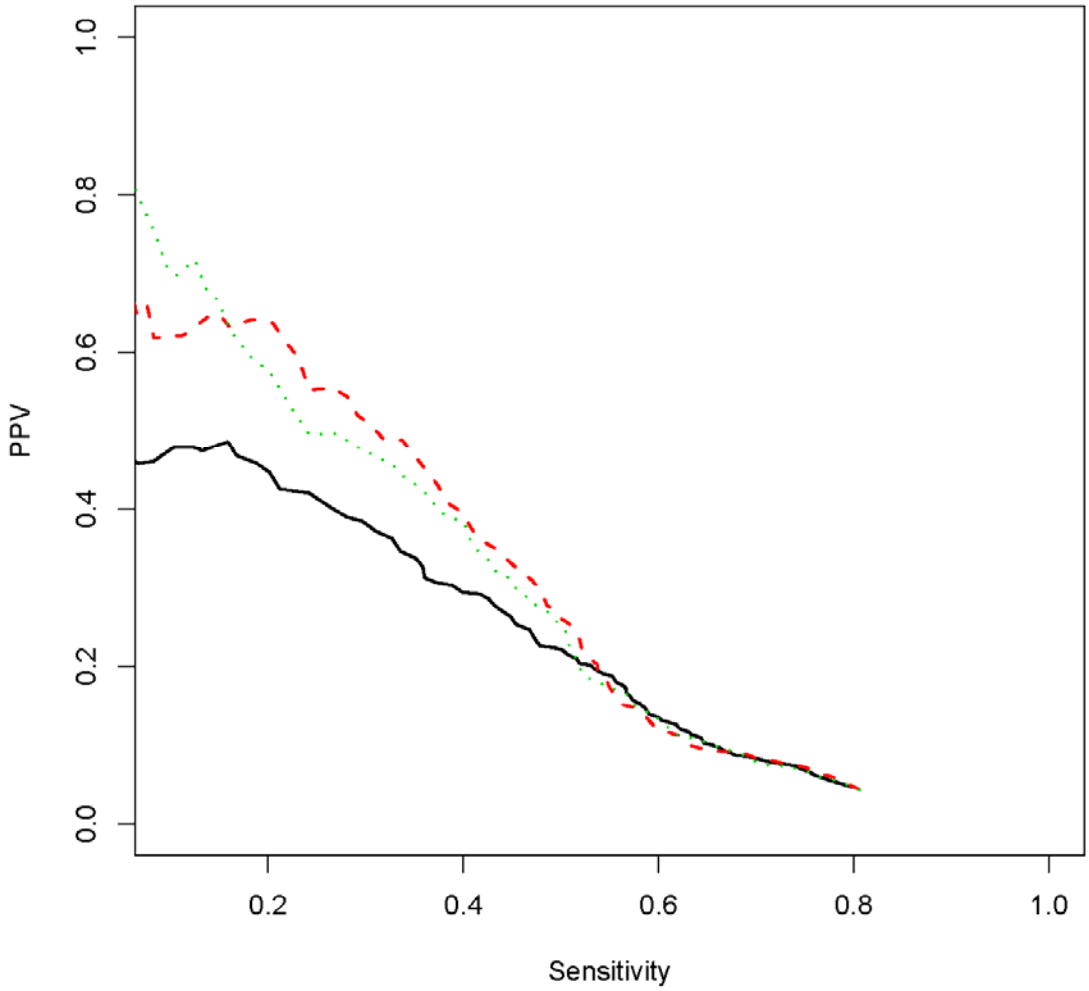


Figure 3.2B

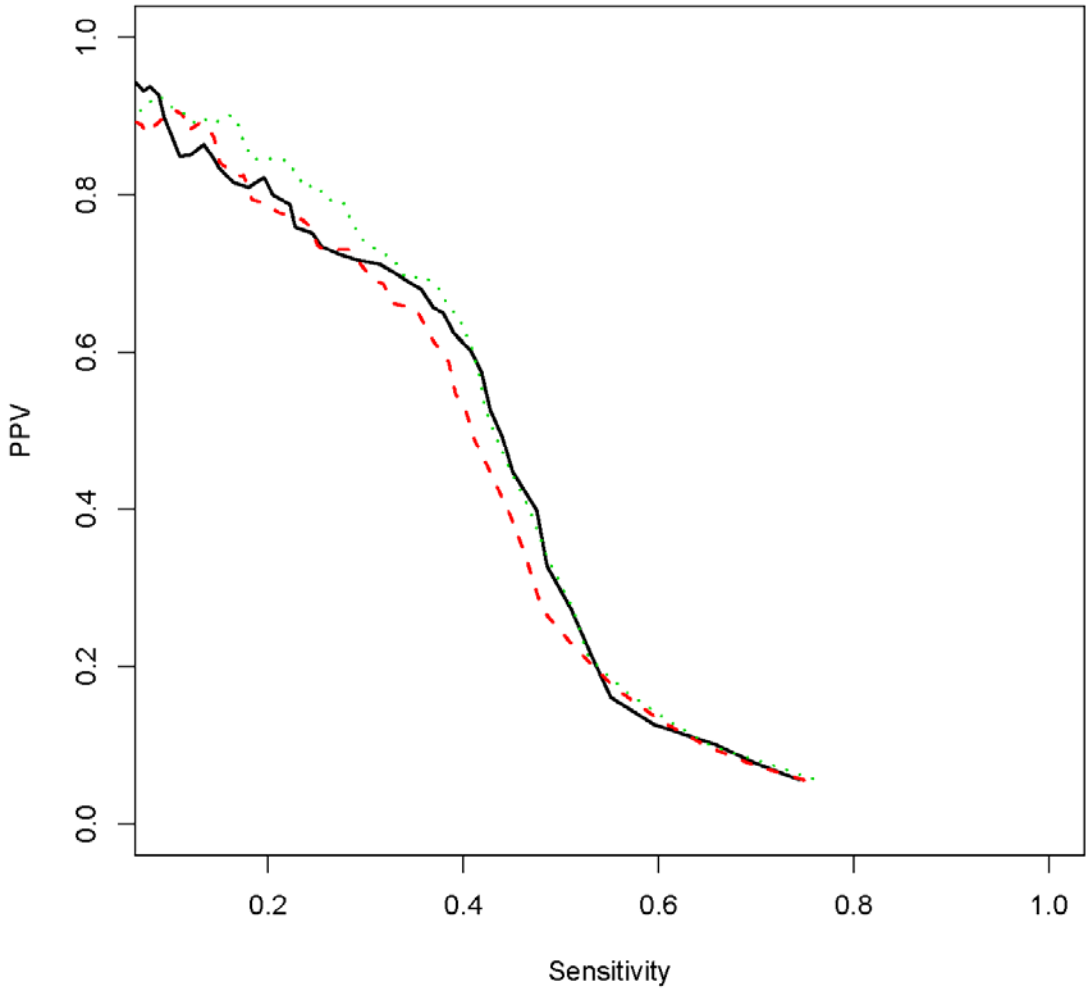


Figure 3.3: Impact of M on clustering results.

The number of blocks available for the Gibbs sampler, *M*, is a key parameter in the procedure. This plot shows the impact of different choices of *M* for *r* = 10.

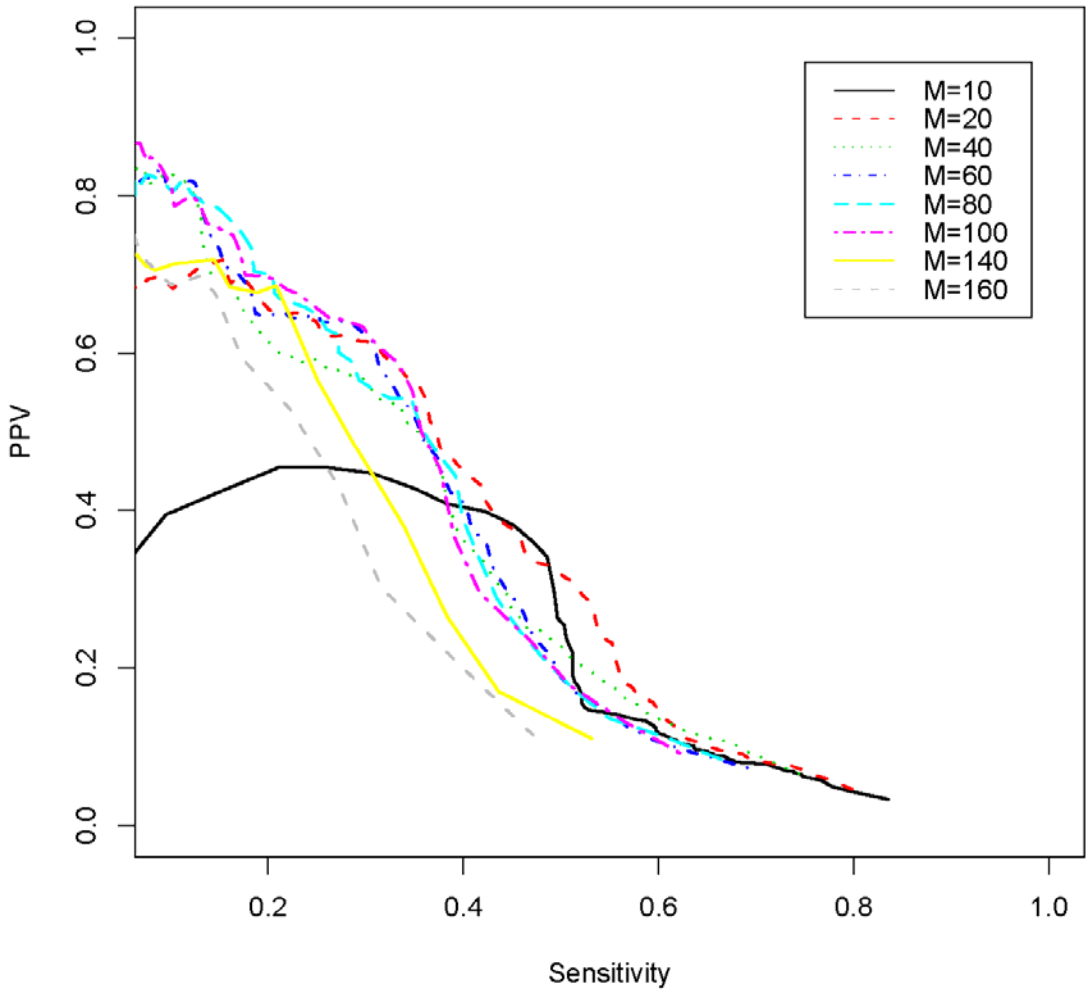


Figure 3.4: Impact of the ratio of 'within block' to 'between blocks' abilities on the clustering results.

The choice of $r = \lambda_W/\lambda_B$ influences the performance of the stochastic block model. A shows the results for varied r with $M = 40$; B shows the results for $M=60$.

Figure 3.4A

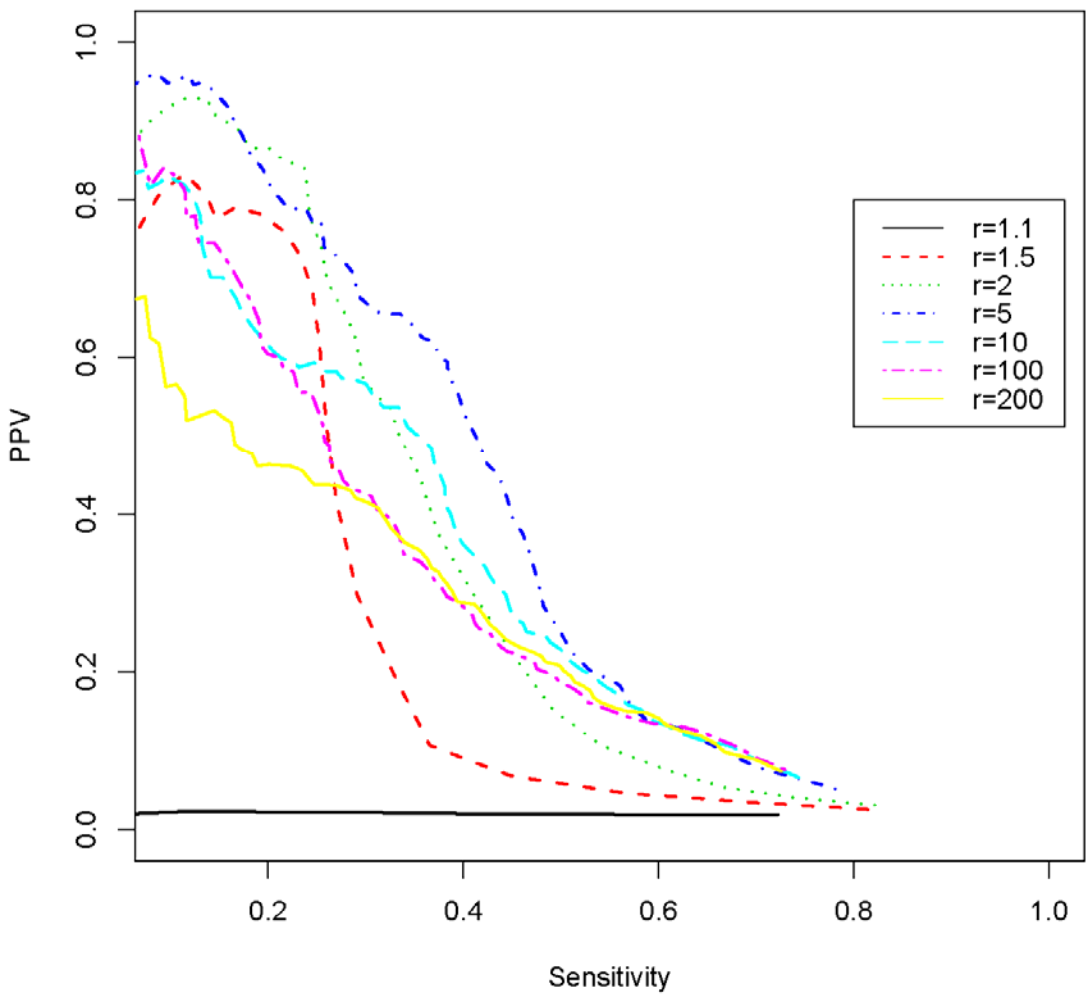


Figure 3.4B

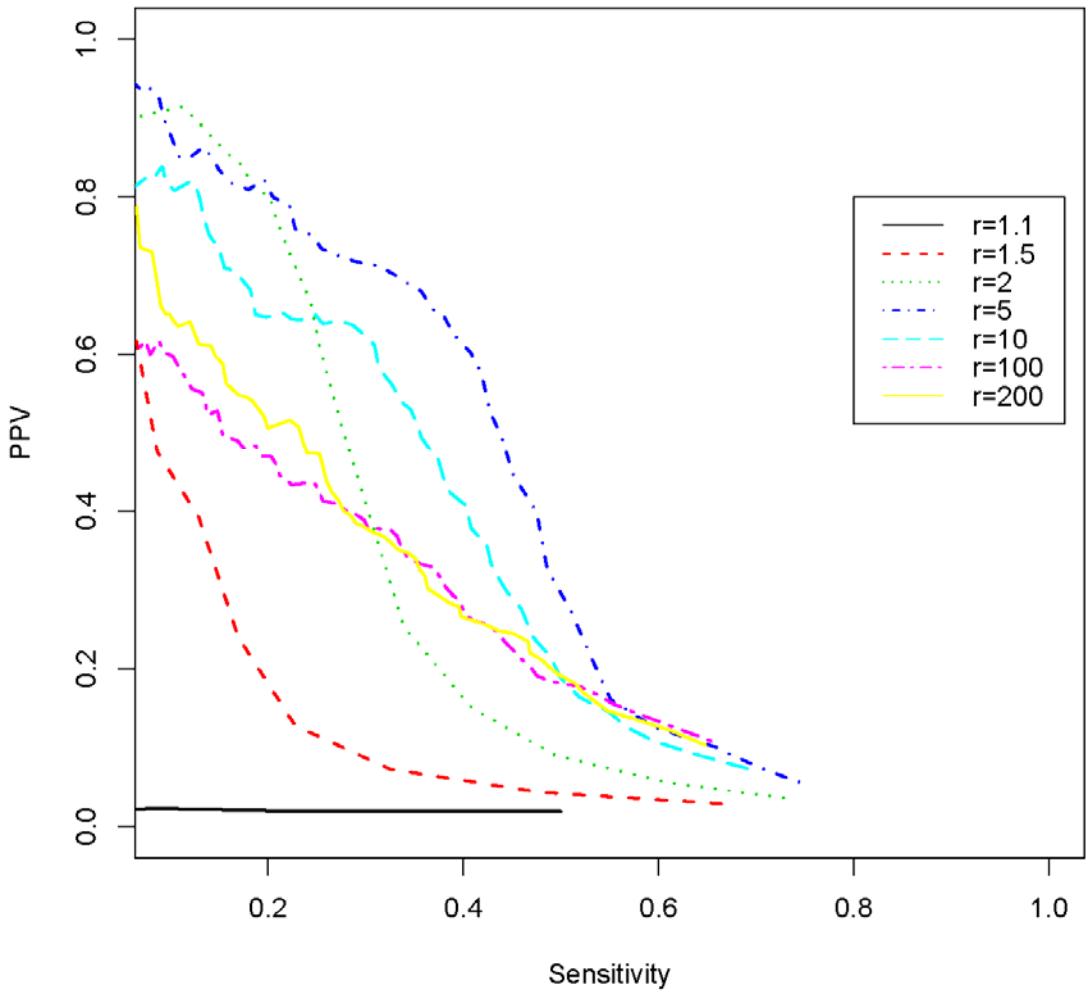

PPV
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
Sensitivity
r=1.1
r=1.5
r=2
r=5
r=10
r=100
r=200

Figure 3.5: Gene modules in CHS6 graph process after 315 steps.

Only one component in the resulting graph is shown. The colors correspond to genes in the AP1 module (red) and the Escrt module (blue). Even at this early stage, the number of edges connecting the two modules and to other genes is large.

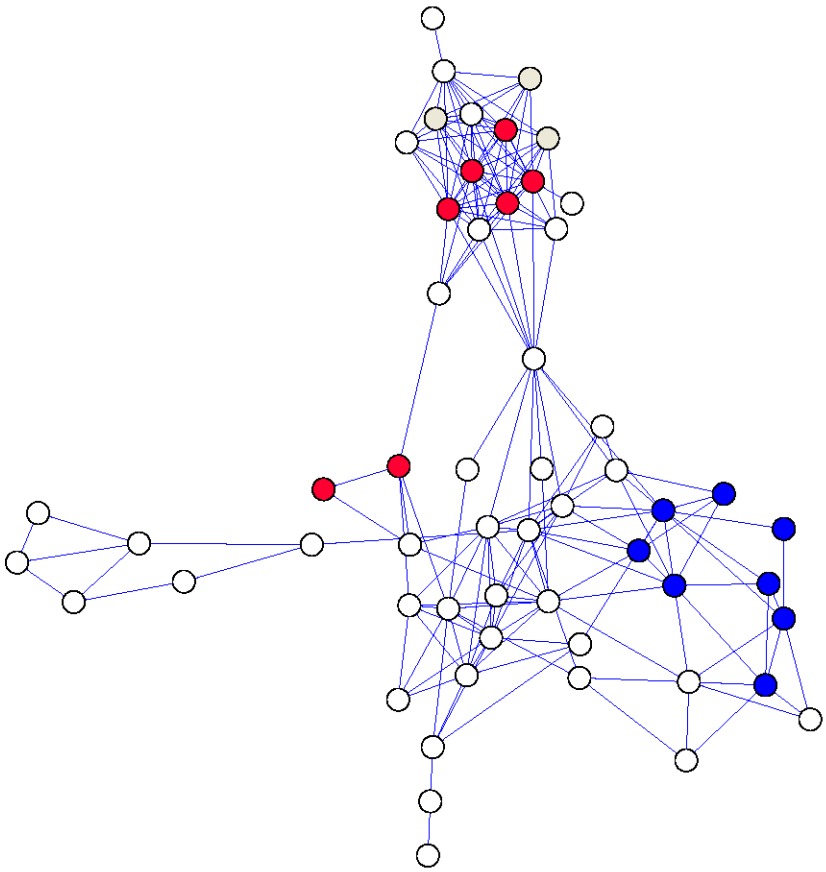
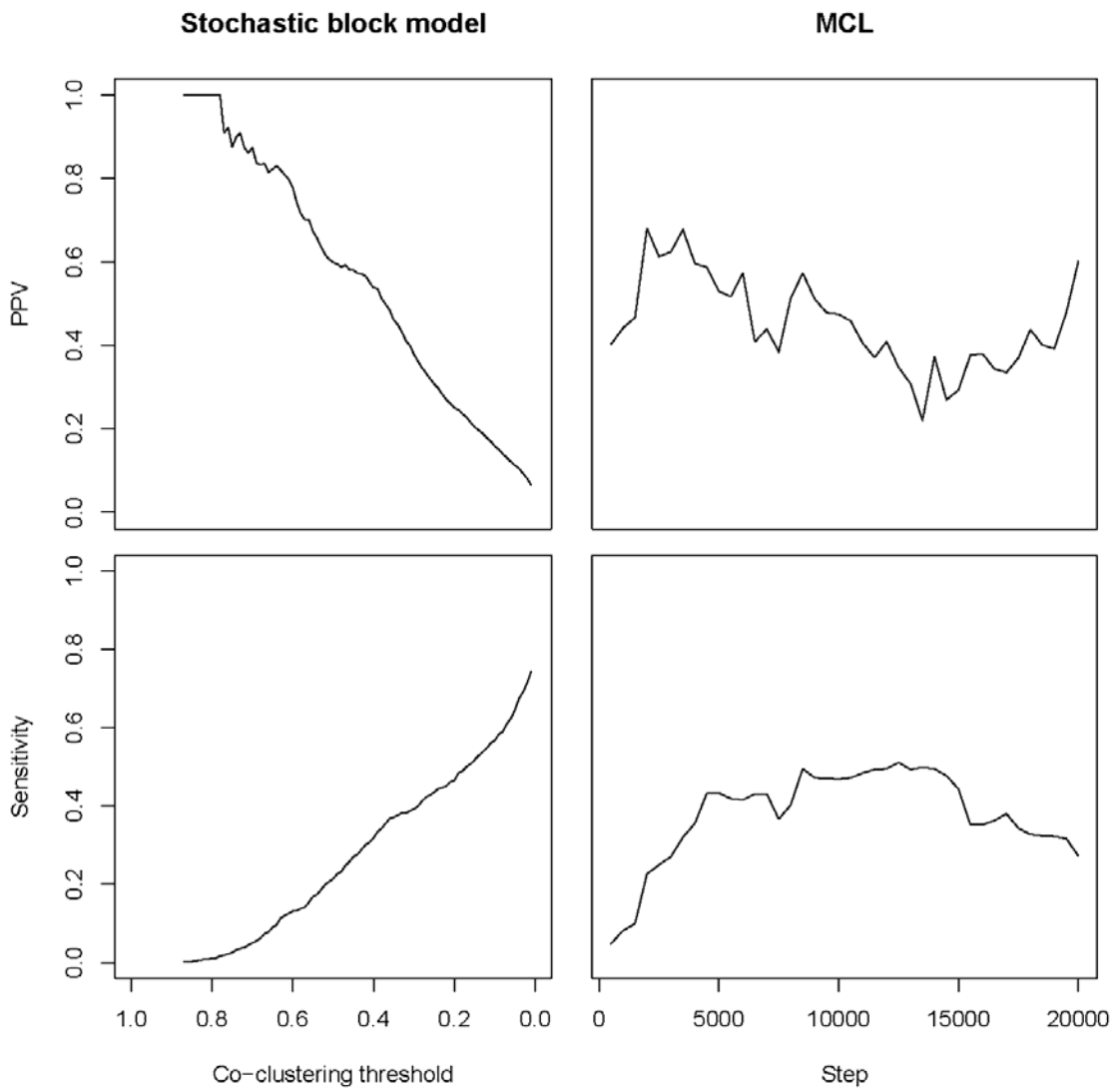


Figure 3.6: Comparison of stochastic block model to MCL clustering for different threshold graphs.

The top row shows the PPV and the bottom row the sensitivity. The x-axis represents the cutoff applied to the co-clustering rates for the stochastic block model (first column), and the number of edges in the threshold graph to be clustered for the MCL method (second column).



3.6 References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
2. Hartuv E, Schmitt AO, Lange J, Meier-Ewert S, Lehrach H, Shamir R: **An algorithm for clustering cDNA fingerprints**. *Genomics* 2000, **66**(3):249-256.
3. Sharan R, Maron-Katz A, Shamir R: **CLICK and EXPANDER: a system for clustering and visualizing gene expression data**. *Bioinformatics* 2003, **19**(14):1787-1799.
4. Rougemont J, Hingamp P: **DNA microarray data and contextual analysis of correlation graphs**. *BMC Bioinformatics* 2003, **4**:-.
5. Luce RD: **Individual Choice Behaviour**. New York: Wiley; 1959.
6. Stern H: **Models for distributions on permutations**. *J Am Stat Assoc* 1990, **85**(410):558-564.
7. Plackett RL: **Analysis of permutations**. *J Roy Stat Soc C-App* 1975, **24**(2):193-202.
8. Snijders TAB, Nowicki K: **Estimation and prediction for stochastic blockmodels for graphs with latent block structure**. *J Classif* 1997, **14**(1):75-100.
9. Nowicki K, Snijders TAB: **Estimation and prediction for stochastic blockstructures**. *J Am Stat Assoc* 2001, **96**(455):1077-1087.
10. Marden JI: **Analyzing and Modeling Rank Data**. London: Chapman & Hall; 1995.
11. Jasra A, Holmes CC, Stephens DA: **Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling**. *Stat Sci* 2005, **20**(1):50-67.
12. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al*: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis**. *Science* 1999, **285**(5429):901-906.
13. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH *et al*: **Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast**. *Cell* 2006, **126**(3):611-625.

14. Schluter C, Lam K, Brumm J, Wu B, Saunders M, Stevens T, Bryan J, Conibear E: **Global analysis of yeast endosomal transport identifies the Vps55/68 sorting complex**. *Mol Biol Cell* in press.
15. Van Dongen S: **Graph clustering by flow simulation**. *PhD Thesis.* Centre for Mathematics and Computer Science (CWI), University of Utrecht; 2000.
16. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics* 2006, **7**:488.
17. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: **A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8348-8353.
18. Thurstone LL: **A law of comparative judgment**. *Psychol Rev* 1927, **34**:273-286.

Chapter 4: Integrated clustering of yeast mutant phenotype profiles and protein-protein interaction data[3]

[3] A version of this chapter will be submitted for publication. Brumm J, Wasserman WW, Bryan J: Integrated clustering of yeast mutant phenotype profiles and protein-protein interaction data.

4.1 Background

Finding functional modules of genes, such as protein complexes and molecular pathways involved in intra-cellular transport or stress response, remains an important and challenging problem in genomics. In order to identify novel modules and unknown members of partially characterized modules, a gene profiling experiment can be performed to measure a gene-specific cellular response, such as gene expression of loss-of-function phenotype under specific conditions. Similarity in the observations for a set of experiments between two genes serves as an indirect measure of functional relationship. Although indirect, this similarity can reflect the *in vivo* membership of modules if appropriate treatments or conditions are assessed [1, 2]. To detect modules based on profile similarities, an unsupervised clustering algorithm is usually employed as sufficient and appropriate training data for a supervised algorithm is rare.

For the successful application of clustering methods, noise is a common limiting property of genomic biological data. Integration of data collected in independent assay formats is a promising strategy to overcome this challenge [3-6]. The integrated approaches offer the prospect of improved clustering as more information becomes available, an important consideration for researchers using high-throughput profiling assays and clustering as an initial screen to select promising candidate modules for expensive validation experiments.

In this paper, we present an application of the probabilistic clustering procedure derived earlier (Chapter 3) for the clustering of yeast mutant phenotype profiles aided by direct relational data in the form of protein-protein interactions. Such direct relational data derived through such techniques as the yeast-two-hybrid assay, is often sparse and reflects a capacity for interaction but does not specify the precise biological context (i.e. the physiological and environmental state) in which interaction occurs [7]. However, such direct measures offer a route to improved clustering of noisy gene profile data.

There are several methods that could be considered for integrated clustering of gene profiles and direct relational data. The active subnetwork approach [8, 9] identifies subnetworks in the direct relational data that show unusually high similarity in the indirect relational data. Alternatively, the distance measure (used to assess the similarity of gene profiles) underlying

the clustering can also be modified to integrate the data. In the shrinkage distance approach, the distance between a pair of genes observed in the direct relational data is reduced by a user-determined shrinkage factor. In the combined distance approach a graph-based distance measure constructed for the direct relational data is combined with the distances for the indirect data into a single distance score [10, 11]. Lastly, what we term the 'graph overlay method' is often presented in applied papers dealing with multiple networks that are combined into a single graph [3]. In this approach, both indirect and direct relational data are converted into a graph; these graphs are then overlaid and informally assessed.

The active subnetworks approach and the combined distance approach both emphasize the importance of the direct relational data and are therefore predicated on broad coverage of such data; that is, a high percentage of interactions between genes in the relevant modules need to be observed.. This coverage requirement limits the utility of these algorithms for applications in genomics research.

In contrast to these approaches, our emphasis is on clustering the indirect relational data, using the direct relational data as auxiliary information (avoiding the undesirable requirement for high coverage), leaving the shrinkage distance and overlay approaches as suitable competitors to our integrated clustering procedure. The shrinkage distance approach integrates the data before clustering, making it vulnerable to false positives in the direct relational data. The overlay approach, on the other hand, integrates the data after the clustering, relying heavily on suitable individual clusterings. Our integrated clustering approach uses the direct relational data in conjunction with the likelihood in the search for the best clustering solution, leading to superior performance in our application of the algorithm to the integrated clustering of gene profiles derived from yeast mutant phenotypes from [2] and direct relational data in the form of protein-protein interactions extracted from [12] as we show below.

Our method uses the stochastic block model for relational data clustering based on graph processes we introduced earlier (Chapter 3). Postulating an unobserved assignment of genes to blocks (clusters) and using a likelihood for the graph process given a candidate labeling, the algorithm delivers for each pair of genes an estimate of the co-clustering probability that measures how likely it is that genes are in the same module. Because our method uses a relational data structure, incorporating direct relational data in the form of a prior probability

distribution in the algorithm is straightforward (Chapter 3). Here we consider as a case-study the integration of protein-protein interactions into the clustering of yeast-mutant phenotype profiles. We show that in this case, the integrated clustering offers considerable improvements in the co-clustering rates of genes within known functional modules compared to the clustering ignoring the prior information, even if the protein-protein interactions contain a high proportion of false positives.

4.2 Results

4.2.1 Direct and indirect data reveal different structures

The CHS6 data set containing the gene profiles used in this study was introduced earlier (Chapter 3); it is a set of 329 gene profiles selected from [2]. The protein-protein interaction (PPI) data was extracted from [12] (using the extended set of interactions given there) and combined with interaction data in the SGD [13]. We selected the interactions matching the gene-pairs contained in the CHS6 data, giving 246 unique pairwise interactions.

The PPI data, depicted in Figure 4.1, reveals that several of the known modules also exhibit high within-module connectivity (all graph displays in this paper were obtained using Cytoscape [14]). The Class B/C, Garp, CCV and DNA complexes are visible as tightly connected subgraphs, as is a part of the AP1 module. The PPI data contains 162 interactions between genes that are annotated in different modules; many of these interactions are likely false positives. As customary in the analysis of PPI data to deal with false positive interactions, we clustered the PPI graph using the MCL graph clustering algorithm [15, 16] (in this paper we always use version 06-058 with default settings); our results below refer to the clusters derived from the PPI graph.

The clustered gene profiles, using the stochastic block model without prior information (with $M = 40$ and $r = 10$), also reveals some clusters corresponding to known modules (Figure 4.2; Table 4.1 gives the summary statistics). The Class B/C module appears fused with the Retromer and Escrt modules and the AP1 module appears as part of a larger cluster. There are also three clusters of size greater than 5 that have no clear functional association.

The two data sources reveal complementary, but differing structures (Table 4.1, Figure 4.2). The Escrt and Retromer modules are part of a larger cluster for the gene profile data, but appear as small individual clusters for the PPI data. The AP1 module is also represented by a tight and pure cluster in the PPI data but as part of a much larger cluster in the gene profile clustering. On the other hand, only two genes of the Garp module appear as a block in the gene profile clustering but this module appears as a part of a cluster of size 8 in the PPI data. The CCV cluster revealed by the PPI data was not observed in the available gene profile data. The clustered PPI data reveals many small clusters not found in the gene profile clusters, but the mixed annotations and small sizes makes it hard to judge which could be subsets of functional gene modules. The gene profile clustering reveals a few larger clusters with mixed and sparse annotations, possibly including genes in the cluster that are not part of the associated module.

4.2.2 Integration of prior information improves clustering

We evaluated the three strategies identified above of incorporating prior information: the shrinkage distance method, the graph overlay method and our stochastic block method. We evaluated two issues concurrently: 1) Does the incorporation of prior information improve the performance of the respective method, as compared to the method not using prior information and 2) Which methods perform best. As metrics for assessing the utility of prior information in the clustering of gene profiles and to compare the performance of different methods, we used the rate of true positive predictions, PPV (the number of true positive relationship predictions divided by the number of all predictions) and the sensitivity, the rate of true interactions recovered (the number of true positive predictions divided by the number of all available true interactions). To evaluate the stochastic block models, we applied a threshold to the estimated co-clustering probabilities to obtain binary relationship predictions (pairs of genes with an estimated co-cluster probability above the threshold are predicted to be in the same module). Note that these predicted relationships are not transitive.

The shrinkage distance method [10, 11] incorporates the PPI data in the distance measure by reducing the distance of pairs of genes which are *a priori* related by a pre-defined shrinkage factor (see Materials and methods). We use this shrinkage distance with varying shrinkage factors in conjunction with our stochastic block model; i.e. we use the shrinkage distance to

compute the pairwise distances for the gene profiles which are subsequently clustered using the stochastic block model (without using prior information).

To apply the graph overlay approach to our data, we first applied the stochastic block model (not using prior information) to the gene profile data. Thresholding the estimated co-clustering probabilities yields a graph of predicted relationships between genes. This graph is then overlaid with the PPI graph, and subsequently clustered with MCL to obtain predicted modules.

The integration of PPI improves all methods (Figure 4.3). The shrinkage distance method performs best at an intermediate choice of the shrinkage parameter, but it has a higher PPV than the clustering not using prior information only at thresholds corresponding to low sensitivity. The graph overlay method has a higher sensitivity than the PPI data clustered by itself. The stochastic block model using prior information has a higher PPV at all levels of sensitivity than the stochastic block model without prior information.

Overall, the stochastic block model using prior information performed best. It outperformed both the stochastic block model without prior information and the stochastic block model using the shrinkage distance uniformly, offering a higher PPV at each given sensitivity. The overlay method is the most competitive for the co-cluster graph derived at higher thresholds (corresponding to fewer edges) but the performance of this method deteriorates at a threshold of 0.3 for the estimated co-clustering probabilities. The union graph is a good structure for capturing complementary information, but it has no effective use for 'double' edges (pairs that are present in both co-cluster and prior graph). The stochastic block model, on the other hand, incorporates additional information by modifying the objective function, taking into account the presence of 'double edges', which could also be extended to incorporate more than one type of prior information.

The stochastic block model using the protein-protein interactions delivers high quality clusters, extending the PPI clusters and improving the clusters obtained without using prior information (Figure 4.4 and Table 4.1, compared to Figures 4.1 and 4.2). Compared to the clustering not using prior information, the Class B/C module separates from the Escrt and Retromer modules and the Garp cluster is new. The clusters with sparse annotations from the clustering without prior information are mostly retained. The stochastic block model with prior information

delivers the most suitable clustering for follow-up studies, striking a balance between the tight PPI clusters and the larger, but noisier gene profile clusters. It also eliminates many small clusters, leaving only few candidate modules compared to the PPI clustering.

The improvement of the estimated co-clustering probabilities within modules through the incorporation of PPI is shown in Figure 4.5A for the Garp module and in Figure 4.5B for the ClassB/C module (although the incorporation of prior information may not always lead to improved estimated co-clustering probabilities within a module; see the DNA module example in Figure 4.5C). The improvement in estimated co-cluster probability is not just for pairs related in the prior data, but also for other pairs within the module, showing the power of integrated clustering. The incorporation of prior data also broke the continuous spectrum of estimated co-cluster probabilities obtained when not using prior information into well-differentiated groups, corresponding to 'within' and 'between' module edges (Figures 4.5 A and B).

4.3 Discussion

The stochastic block method, using prior information, outperforms competitive methods and is able to identify relevant gene modules in this difficult data set better than the stochastic block model not using prior information. The estimated co-clustering probabilities it provides are a useful measure for the prioritization of follow-up experiments.

A class of algorithms not considered here are clustering methods that represent the gene profiles as vectors in a high-dimensional feature space [17, 18], extended to allow the incorporation of prior data [19]. Note however, that the feature-space methods cannot readily incorporate prior *relational* data [20].

The shrinkage and overlay methods could be used with any suitable clustering algorithm. However, using the same algorithm for the competitive methods as for our method allows us to isolate the impact of prior data on the results.

The overlay method is most competitive in this application but it only works well in sparse graphs. This problem will be exacerbated if more than one set of prior information is to be

incorporated. Future research will be directed towards the inclusion of such multiple priors (for example joint integration of functional categorization such as GO [21] and protein-protein interactions).

Clustering is now an essential tool in the arsenal of the applied researcher. As cellular systems are studied in more depth, more data and information becomes available. Data is typically noisy, so it is important to take advantage of this information to deliver quality predictions for follow-up studies. This paper has demonstrated the utility of the use of prior information in the stochastic block model in the important example of integrated clustering of protein-protein interactions and yeast mutant phenotype profiles.

4.4 Materials and methods

4.4.1 Data preprocessing

The CHS6 data (introduced in Chapter 3) is a subset of 329 genes measured under 82 conditions selected from [2] according to their mutant phenotype in an independent screen. The gene profile data was normalized by dividing by the standard deviation for each condition. The similarity of two gene profiles was computed as one minus the correlation of these two profiles.

Documented protein-protein interactions used as prior knowledge were downloaded from the yeast community database SGD on March 20, 2007. In pre-processing, redundancy was removed.

The ‘DNA module’ is a broad category of genes which participate in a variety of modules related to the integrity and production of DNA in yeast, such as DNA repair and DNA replication.

4.4.2 Stochastic block model

The model is introduced in detail in Chapter 3. For the Gibbs sampling procedure, parameters were set to 110 iterations, of which the initial 10 iterations were discarded as the burn-in period

for the sampler. The prior co-cluster probability was set to 0.99, if an edge was present in the PPI data (called α in Chapter 3), or 0.5 if no edge was present (called β in Chapter 3). Results were not very sensitive to changing α to other values close to 1.

4.4.3 Shrinkage distance method

The shrinkage distance method introduced in [11] computes a distance from gene profiles using a given distance function *d* and a set of known functional relationships between genes. The function d is then 'shrunk' by a factor *s* ($0 < s < 1$), if a pair of genes has a functional relationship, that is if x and y are two gene profiles, $d^*(x,y) = s * d(x,y)$ if the two corresponding genes are related, and $d^*(x,y) = d(x,y)$ if not. In [11], this method is applied to functional categories and used with a customized algorithm. We adapt the philosophy here and use observed protein-protein interaction as 'functional relationships'. We use the stochastic block model as the clustering algorithm.

4.5 Tables and figures

4.5.1 Tables

Table 4.1: Recovery of known gene modules in clusterings of protein-protein interaction data. Column A: Gene modules known to be relevant for the CHS6 data; number of constituent genes given in parentheses. Columns B – D: results of various approaches to identify biological modules in the CH6 data; reported as (number of genes from the given modules in cluster)/(number of genes in cluster). Column B: PPI data (Figure 4.1) clustered with MCL (version 06-058 with default settings). Column C: stochastic block model result, no prior information used. Column D: stochastic block model, using the PPI data as prior information. The MCL clustering revealed in addition two clusters of size 7, three clusters of size 5, 5 clusters of size 4, 11 clusters of size 3 and 26 clusters of size 2 (the complete set of clusterings for column C is given in Figure 4.2; for column D in Figure 4.4)

(A) Module	(B) PPI data processed with MCL	(C) Stochastic block model, no prior	(D) Stochastic block model with prior
Class B/C (9)	9/18	7/17	8/13
Escrt (9)	3/3	4/17	8/15
Garp (5)	5/8	2/2	4/6
AP1 (5)	4/6	5/17	5/16
Retromer (4)	3/4	3/17	4/15
DNA (8)	3/3	4/7	5/7
CCV (5)	4/6	1/17	1/2; 1/15

4.5.2 Figures

Figure 4.1: PPI network extracted from [12] for the genes contained in the CHS6 data. Selected gene modules are identified by color.

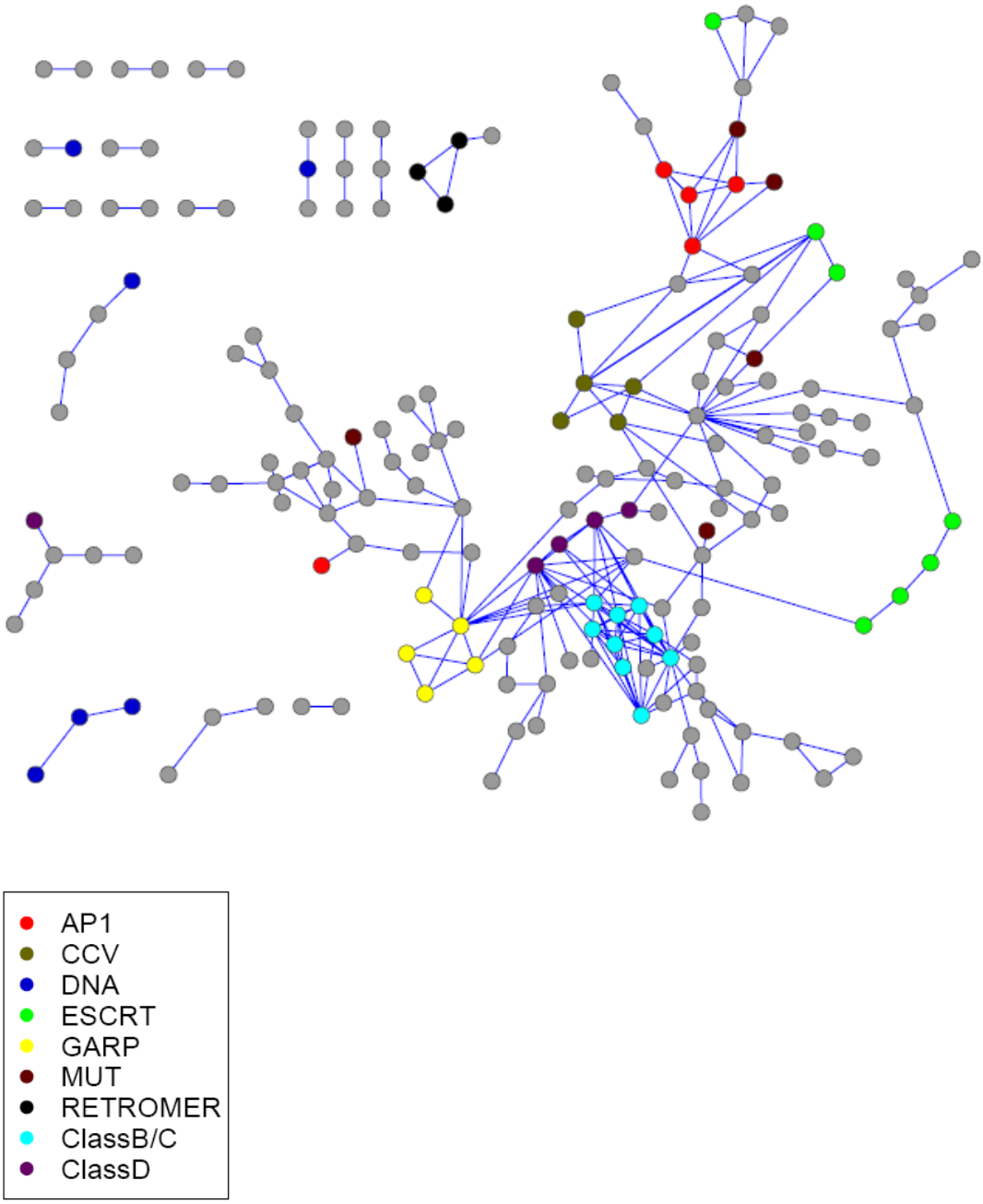


Figure 4.2: Stochastic block model clustering (M = 40, r = 10) of the gene profiles (not using prior information).

The graph displays edges where the estimated co-clustering probability is greater than 0.5.

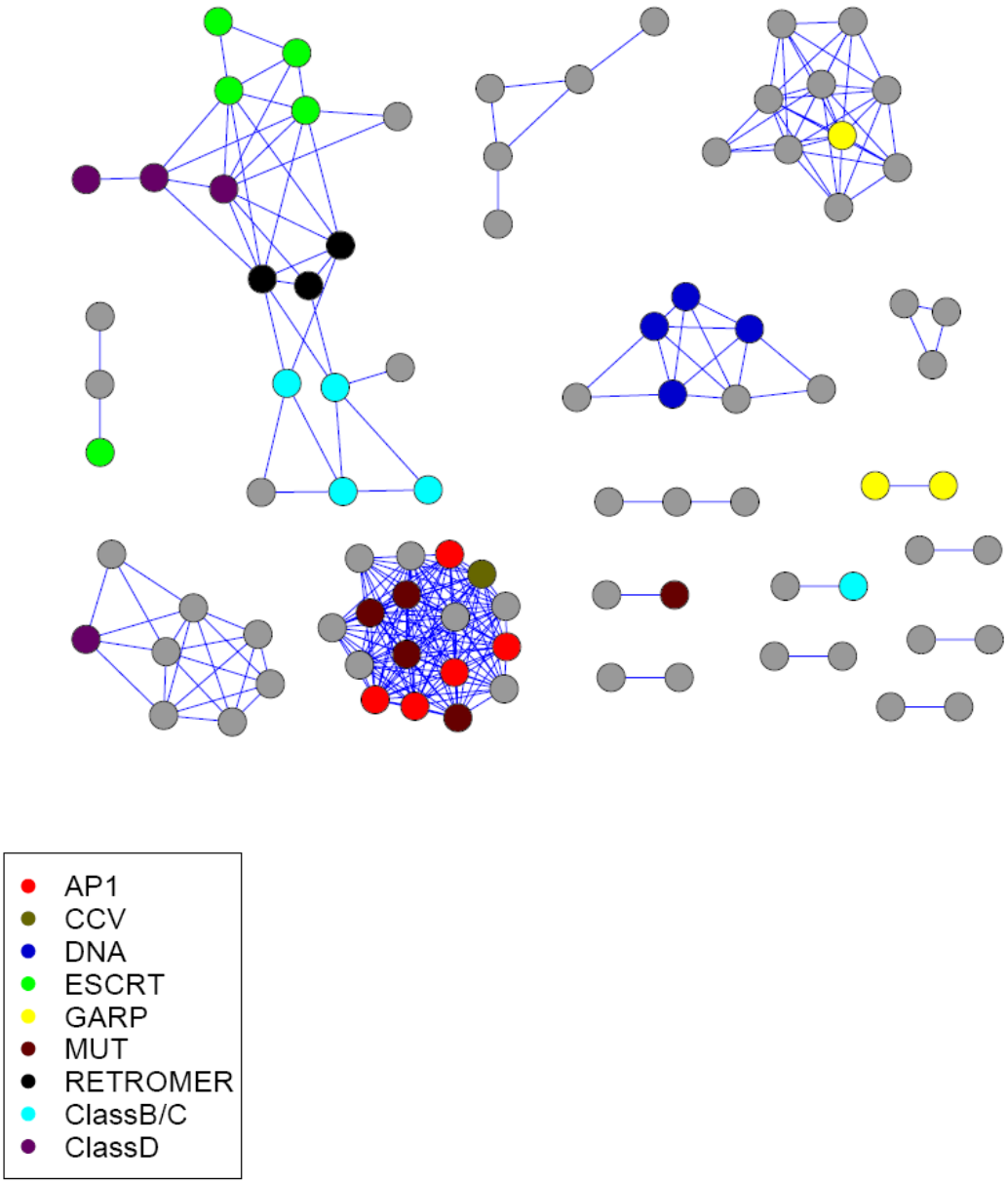


Figure 4.3: Comparison of performance of methods.

The PPV is plotted against the sensitivity; lines are obtained by decreasing the threshold for the co-cluster probability from stringent (threshold=1) to weak (threshold=0). Dots on the line corresponding to the stochastic block model using prior information indicate thresholds of 0.6, 0.5, 0.4 and 0.3.

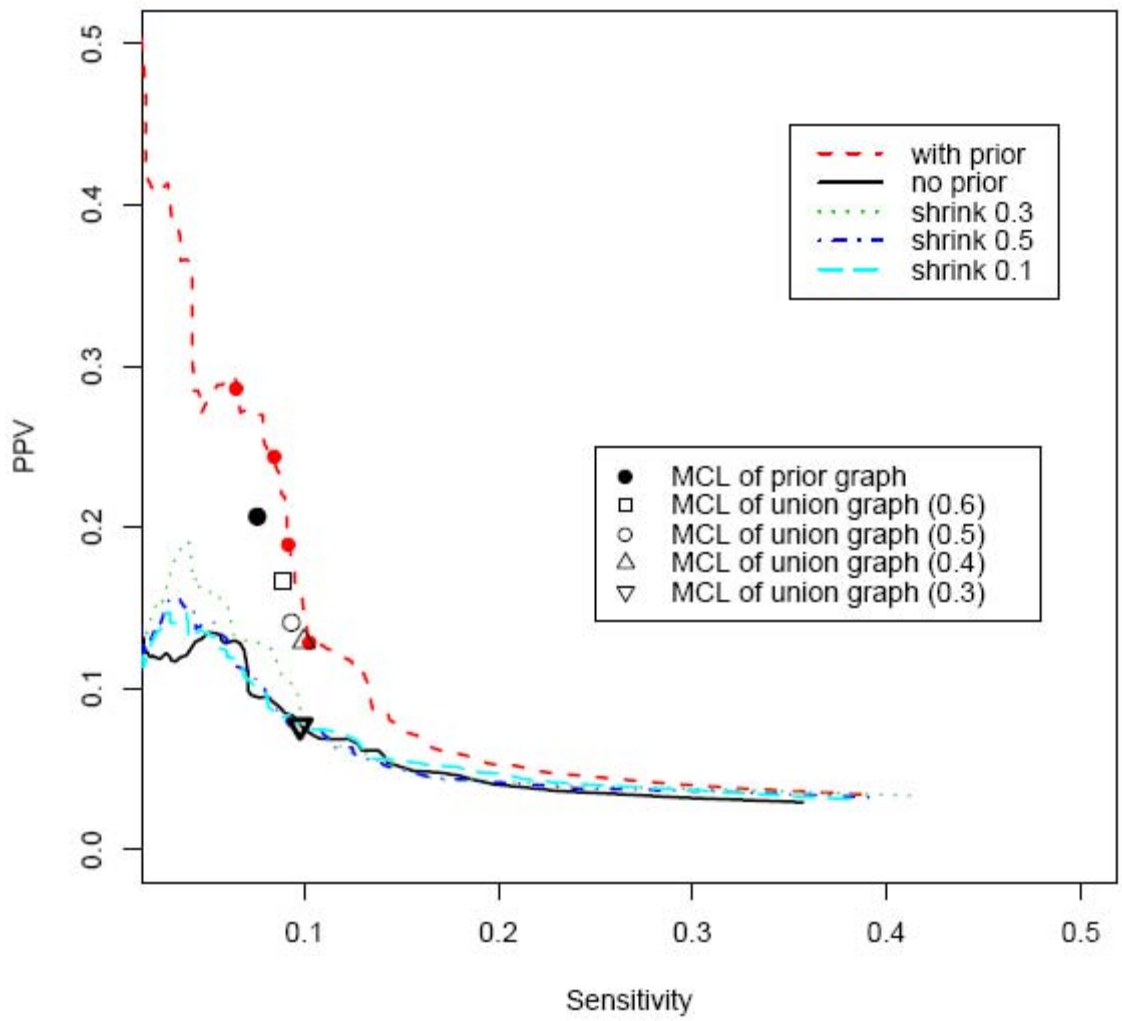


Figure 4.4: Integrated clustering of the gene profiles ($M = 40$, $r = 10$).

The graph displays edges where the estimated co-clustering probability is greater than 0.5.

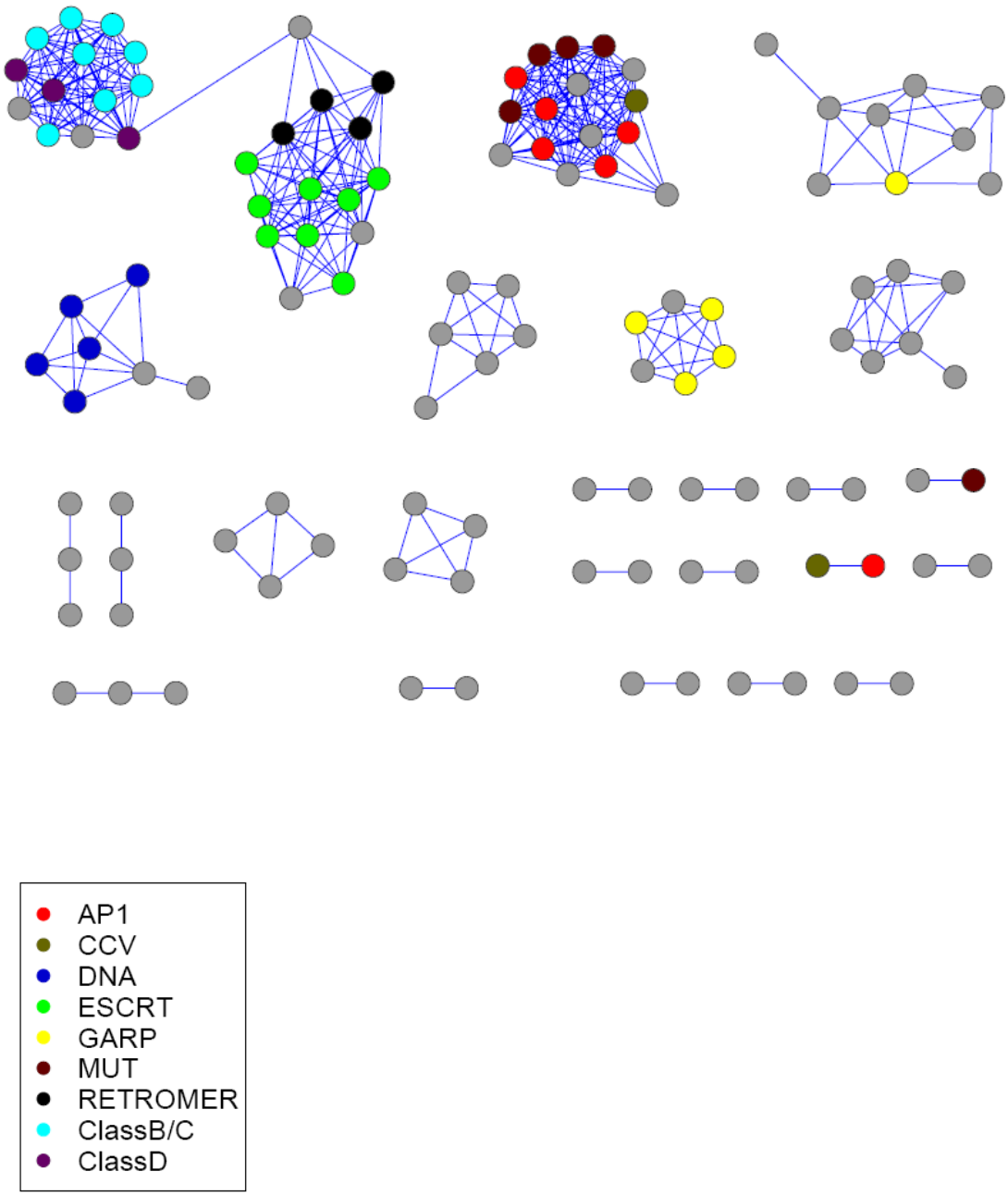
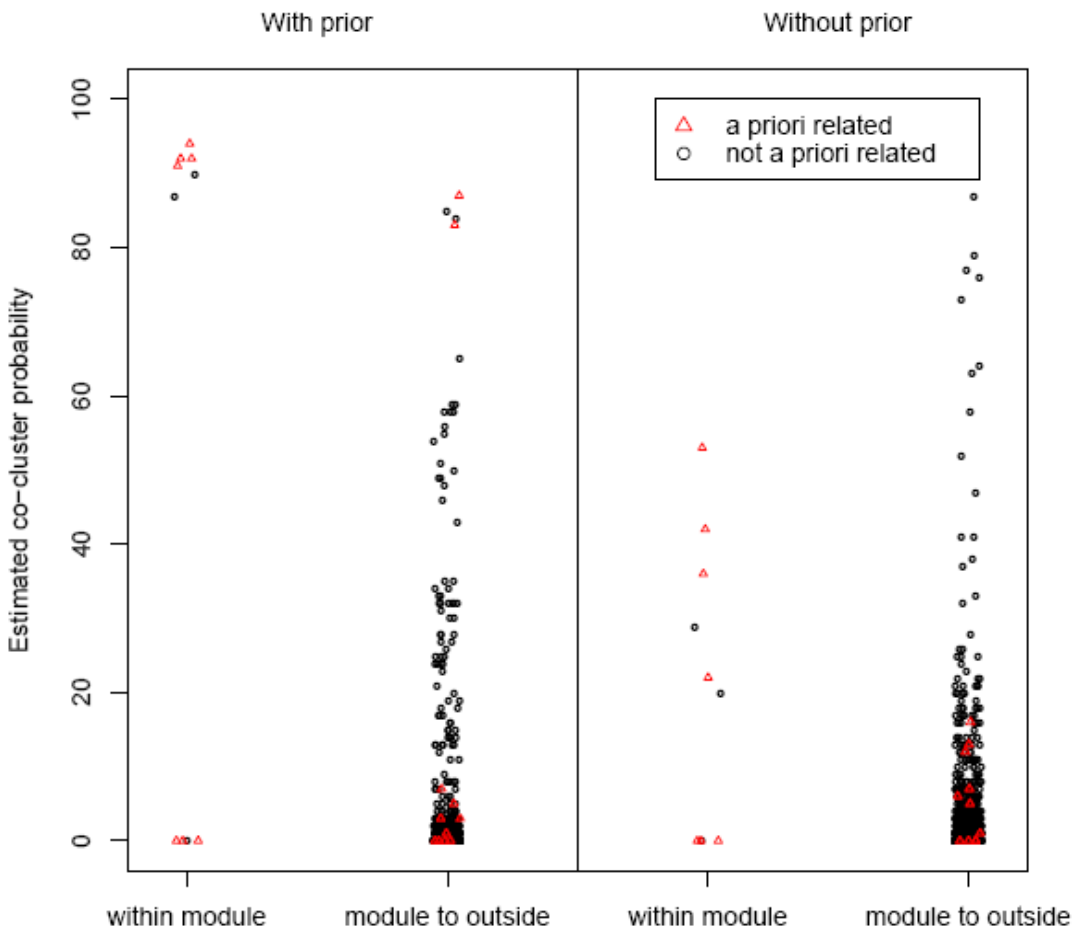
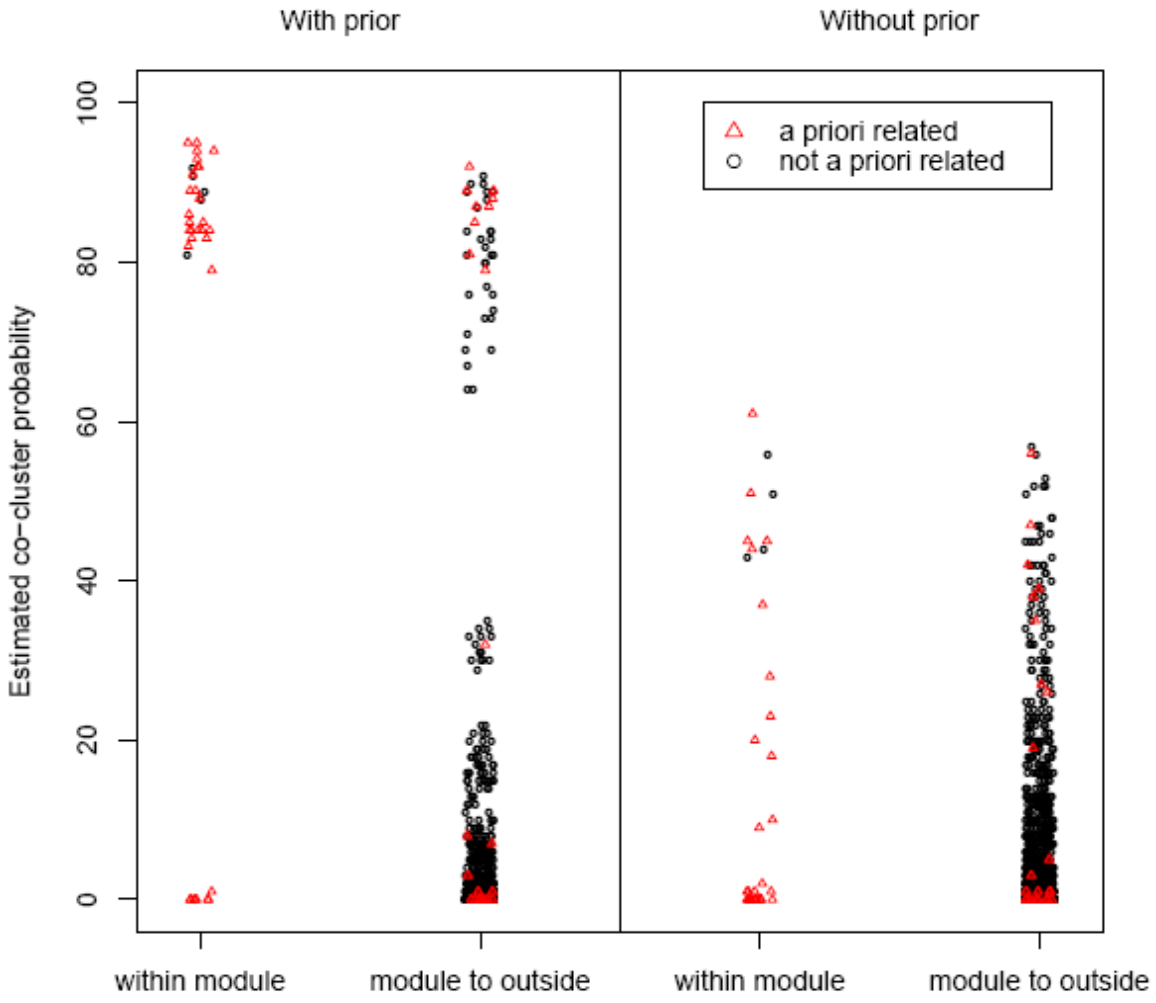


Figure 4.5: Impact of prior information on co-clustering probabilities (multiplied by 100). A shows the estimated co-clustering probabilities for the Garp module, B for the ClassB/C module and C for the DNA module. The probabilities are split by method (with and without prior) and by nature of relationship (within module, between module and outside). Relationships contained in the protein-protein interaction data are indicated by a red triangle. The plotted symbols are jittered along the x-axis to reduce overplotting.

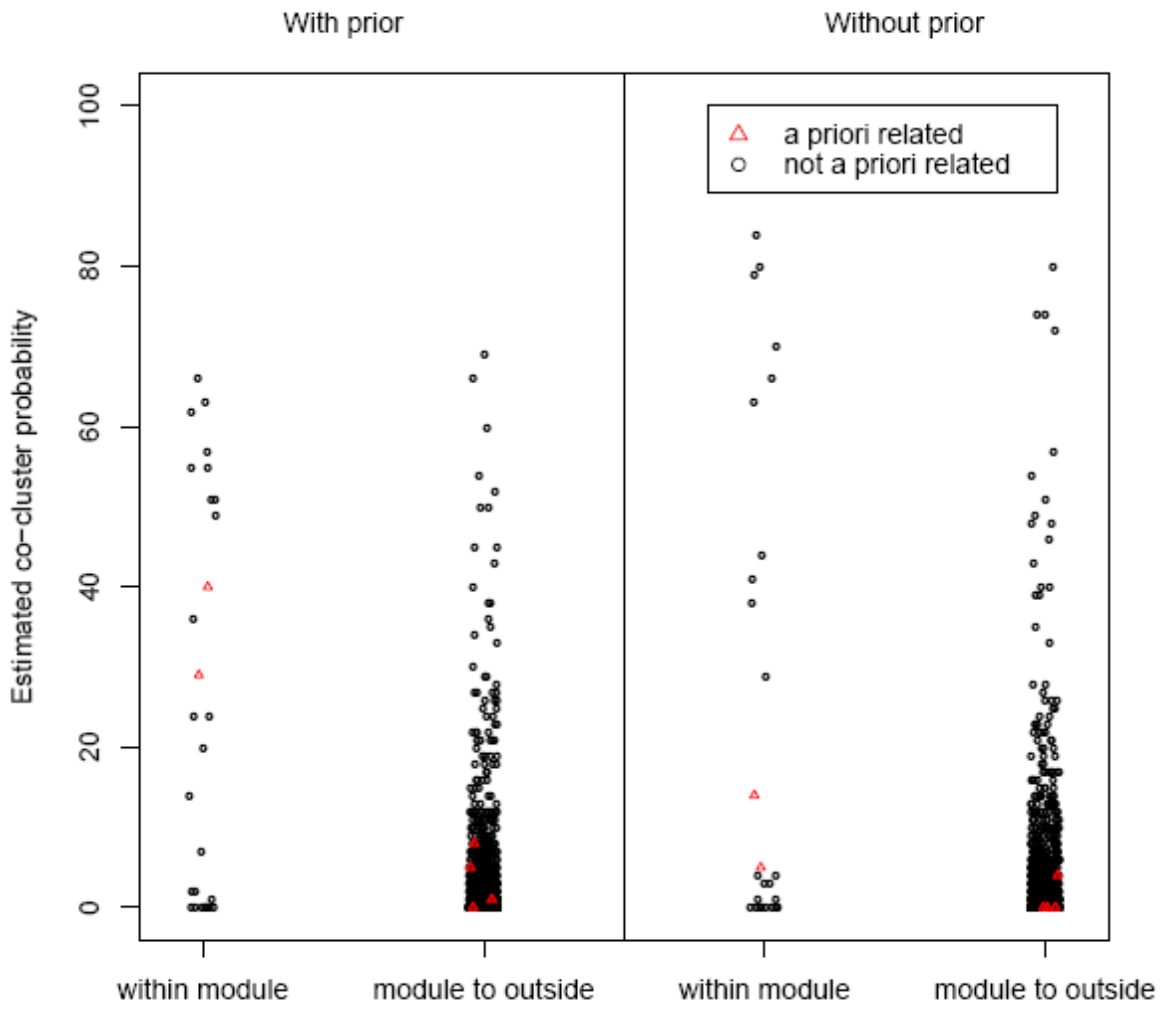
A



B


With prior
Without prior
a priori related
not a priori related
Estimated co−cluster probability
0
20
40
60
80
100
within module
module to outside
within module
module to outside

C


With prior
Without prior
a priori related
not a priori related
Estimated co−cluster probability
0
20
40
60
80
100
within module
module to outside
within module
module to outside

4.6 References

1. Lee W, Onge RPS, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G: **Genome-wide requirements for resistance to functionally distinct DNA-damaging agents**. *PLoS Genet* 2005, **1**(2):235-246.
2. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH *et al*: **Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast**. *Cell* 2006, **126**(3):611-625.
3. Ge H, Walhout AJM, Vidal M: **Integrating 'omic' information: a bridge between genomics and systems biology**. *Trends Genet* 2003, **19**(10):551-560.
4. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: **A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8348-8353.
5. Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng CC, Afshari CA, Brown JM, Ideker T, Van Houten B: **Integrating phenotypic and expression profiles to map arsenic-response networks**. *Genome Biol* 2004, **5**:R95.
6. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.
7. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11).
8. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction data**. *Bioinformatics* 2002, **18**(Suppl 1):S233-S240.
9. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data**. *BMC Syst Biol* 2007, **1**(1):8.
10. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data**. *Bioinformatics* 2002, **18**(S1):S145-S154.
11. Huang DS, Pan W: **Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data**. *Bioinformatics* 2006, **22**(10):1259-1268.

12. Kiemer L, Costa S, Ueffing M, Cesareni G: **WI-PHI: a weighted yeast interactome enriched for direct physical interactions**. *Proteomics* 2007, **7**(6):932-943.

13. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res* 1998, **26**(1):73-79.

14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

15. Van Dongen S: **Graph clustering by flow simulation**. *PhD Thesis.* Centre for Mathematics and Computer Science (CWI), University of Utrecht; 2000.

16. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics* 2006, **7**:488.

17. Fraley C, Raftery AE: **Model-based clustering, discriminant analysis, and density estimation**. *J Am Stat Assoc* 2002, **97**(458):611-631.

18. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**(10):977-987.

19. Pan W: **Incorporating gene functions as priors in model-based clustering of microarray gene expression data**. *Bioinformatics* 2006, **22**(7):795-801.

20. Law MHC, Topchy A, Jain AK: **Clustering with soft and group constraints**. *Lect Notes Comput Sc* 2004, **3138**:662-670.

21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene Ontology: tool for the unification of biology**. *Nat Genet* 2000, **25**(1):25-29.

Chapter 5: Discussion

5.1 Contributions of this thesis

Clustering methods are and will continue to be central to the interpretation of experimental data in genomics. There are an increasing number of assays that query relationships between genes, producing data for both direct and indirect interactions. Beyond the assays producing data analyzed in this thesis, there are high-throughput approaches in use for study of co-localization within cellular compartments, in-situ expression (providing two dimensional RNA or protein expression patterns from slices of an organ or organism), protein-protein interaction data from mass spectrometry, and more. In short, new approaches that provide a capacity to incorporate data of diverse form and character have a receptive and waiting audience.

Clustering methods will also remain important because gene modules can be different in different cell types. There is increasing evidence that regulatory systems in cells rely on combinatorial mechanisms. Subsets of a large group of proteins called transcription factors come together in various combinations to activate the expression of specific sets of genes, each of these sets being a gene module in our terminology. For example, there is now evidence that there are many types of neurons, and that the differences between these types are driven by combinatorial control of a set of transcription factors [1]. Deciphering these combinatorial mechanisms that are active in subtypes of cells will require more detailed experiments.

While standard clustering algorithms have been used successfully in many applications in genomics, output from standard clustering algorithms has to be interpreted with care in the presence of noise, which is prevalent in genomic data. Since these algorithms always produce a clustering, regardless of the nature of the input data, noise can force genes into the same cluster that are functionally unrelated.

This thesis introduces methods based on ranked relational data for the unsupervised discovery of functional modules of genes. Ranked relational data captures information in the data beyond binary pairings, while maintaining the universality and utility of relational data structures. Viewing ranked relational data as a ranking of items also allowed us to develop probabilistic clustering methods based on classical likelihoods developed for the analysis of horse races. Our probabilistic methods allow the determination of confidence scores for clusters and the

integration of prior knowledge about direct relationships (such as protein-protein interactions). Our approach is robust to noise in the data and provides reliable prediction of true blocks of small size.

5.2 Topics related to this thesis

5.2.1 Feature based clustering

Gene profiles, which are measurements of a phenotype or an expression collected under a number of conditions, can be represented as a feature vector in a vector space. There are many clustering algorithms, such as *k*-means and others, that use this representation [2], as well as probabilistic approaches that have been used in genomics [3, 4]. These algorithms have drawbacks, however. The number of conditions measured is often high (leading to a high-dimensional feature space) and the values measured may be categorical, making probabilistic approaches that rely on a distribution for the feature vectors less attractive. We have also seen in Chapter 4 that the integrated clustering of direct relational data and gene profiles may not be straightforward [5].

5.2.2 Feature selection and geneset selection

In experiments with many conditions, it is often of interest to distinguish conditions that allow one to identify a gene module from conditions that do not help in identification (often called "feature selection"). Feature selection is not easily accomplished for gene profile data, if the gene profiles are converted into relational data using a distance (or similarity) measure that summarizes across all conditions. To accommodate feature selection, avoiding the use of a distance measure and instead representing genes and conditions as nodes in a graph which is subsequently clustered may be a promising approach [6]. See also [7] for a more general application of such data structures.

Which set of genes is selected for clustering may have substantial influence on the clustering results. In the MISO method for example, the isolation p-values might depend on the nature of gene modules in the data. Since MISO depends on the ranking of the data, the presence of multiple gene modules with comparable within-module similarities will lead to later birth of the full modules as opposed to the situation where only one of these modules is present in the data. However, this dependence of the clustering result on the selection of genes is true for most clustering methods, since the identification of clusters and the discrimination of genes into clusters will depend on the relative density and isolation of potential clusters.

5.2.3 Model extensions

This thesis considers only simple stochastic block models, with homogenous 'strength' parameters for the within block and between blocks relationships. Extensions could be considered that relax these constraints, but would be useful only if a suitable estimator for the strength parameters could be found.

We also used prior information only in binary form. This restriction can be relaxed, at the expense of a moderate increase in computing time, if the database used records a reliability coefficient (such as WI-PHI [8]).

The greatest obstacle to the utility of the methods proposed is the computational complexity (discussed in Chapter 3). The stage-wise form of the likelihood is a precise model for the graph process, but implies a large computational burden. A potential avenue for improvement are approximations to the likelihood computations. The label update step for a given gene, for example, could potentially be based only on the first m edges hitting the node, with m being much smaller than the number of genes.

5.2.4 Probabilistic models and statistical inference

Our probabilistic procedure produces confidence scores for genes being in the same clusters, but we are hesitant to recommend cut-offs classifying gene-gene relationships as 'present' or 'absent' without establishing a model that appropriately reflects the data. Such a model could be used to evaluate such cut-offs and establish realistic performance estimates for future data, possibly even for statistical inference eliminating the need for expensive follow-up studies.

A model for relational data could be based on a parametric model for the observed similarities like the one suggested in [9]. However, modeling genomic data is ambitious and often regarded skeptically by biologists. The module clustering assumption used in Chapter 3 reduces complex cellular systems to relationships between genes. Clearly there are biological limits to our fundamental assumption that 'genes within a module will produce similar profiles'. In a living cell, the delineation of what genes are within a module and outside a module might be difficult. A related problem is the fact that genes within modules have differing functions, leading to heterogeneous gene profiles within a module.

5.3 Future research

A useful and methodologically close extension of our model would be to integrate multiple gene profile datasets. Experiments are often done in different labs and may even come from very different assays, making data standardization for subsequent integration difficult. Combining the ranked relationships may offer an effective way to circumvent this.

For this extension, however, deviations from the stochastic block model we consider may be necessary as a gene may belong to one module in the first dataset and to a different module in the second dataset. Also, to extend the method to gene expression datasets, hierarchical structures for the module parameters may be useful. When considering gene regulatory networks bipartite graph structures on the gene-gene relationships may also provide additional flexibility [6, 10, 11]. Incorporating the dynamics of gene modules into the model could also be

important in the analysis of time-course studies, as gene modules are often assembled transiently (for example, in the response to stress [12, 13]).

Other model specifications that capture the graph process evolution are also available and may provide advantages over the models considered in this thesis. Models that map relational data into a feature space are used in the social sciences [14]. It may also be possible to model the distances directly (by using exponential random variables, such as suggested in [9]). Such an approach could be more effective since it uses more information than the rank-based approach.

The analysis of high-dimensional data sets to uncover the functional organization of genes within a cell remains a challenging and important task. New assays, such as RNA interference [15], are constantly emerging to collect such data. These assays offer more and more insight into the complex networks of genes within cells; improving clustering methods ensures that scientists gain the most from these novel information sources.

5.4 References

1. Allan DW, Park D, St. Pierre SE, Taghert PH, Thor S: **Regulators acting in combinatorial codes also act independently in single differentiating neurons**. *Neuron* 2005, **45**(5):689-700.
2. Kaufman L, Rousseeuw PJ: **Finding Groups in Data: An Introduction to Cluster Analysis**. New York: Wiley; 1990.
3. Fraley C, Raftery AE: **Model-based clustering, discriminant analysis, and density estimation**. *J Am Stat Assoc* 2002, **97**(458):611-631.
4. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**(10):977-987.
5. Law MHC, Topchy A, Jain AK: **Clustering with soft and group constraints**. *Lect Notes Comput Sc* 2004, **3138**:662-670.
6. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data**. *Proc Natl Acad Sci U S A* 2004, **101**(9):2981-2986.
7. Zhong S, Ghosh J: **A unified framework for model-based clustering**. *J Mach Learn Res* 2004, **4**(6):1001-1037.
8. Kiemer L, Costa S, Ueffing M, Cesareni G: **WI-PHI: a weighted yeast interactome enriched for direct physical interactions**. *Proteomics* 2007, **7**(6):932-943.
9. Bock HH: **Probabilistic models in cluster analysis**. *Comput Stat Data An* 1996, **23**(1):5-28.
10. Scholtens D, Vidal M, Gentleman R: **Local modeling of global interactome networks**. *Bioinformatics* 2005, **21**(17):3548-3557.
11. Beyer A, Bandyopadhyay S, Ideker T: **Integrating physical and genetic maps: from genomes to interaction networks**. *Nat Rev Genet* 2007, **8**(9):699-710.
12. Hahn JS, Hu ZZ, Thiele DJ, Iyer VR: **Genome-wide analysis of the biology of stress responses through heat shock transcription factor**. *Mol Cell Biol* 2004, **24**(12):5249-5256.

13. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes**. *Mol Biol Cell* 2000, **11**(12):4241-4257.

14. Hoff PD, Raftery AE, Handcock MS: **Latent space approaches to social network analysis**. *J Am Stat Assoc* 2002, **97**(460):1090-1098.

15. Fire A, Xu SQ, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. *Nature* 1998, **391**(6669):806-811.