Inference of gene regulatory networks for *Mus musculus* by incorporating network motifs from yeast.

Holger Weishaupt

Master's dissertation

University of Skövde

02.05.2007

Inference of gene regulatory networks for *Mus musculus* by incorporating network motifs from yeast.

Holger Weishaupt

Submitted by Holger Weishaupt to the University of Skövde as dissertation towards the degree of Master by examination and dissertation in the School of Humanities and Informatics.

02.05.2007

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

_____

Holger Weishaupt

Table of content

1. Abstract ……4

2. Background ……4
2.1 Previous work on modelling regulatory relationships ……5
2.2 Previous work on inferring regulatory relationships ……8
2.3 Objectives ……11

3 Methods ……12
3.1 How to incorporate information ……12
3.2 How to use scores in order to propose regulatory relationships ……18
3.3 Examination of Permutations ……19
3.4 Course of action ……22

4 Results ……26
4.1 GO functional identity ……26
4.2 Sequence similarity ……30
4.3 Inferring complete networks ……31

5 Conclusion ……32
5.1 Summary of conclusions ……32
5.2 Detailed conclusion ……33
5.3 Problems ……35
5.4 Implications and improving the method ……37

6 References ……40

1. Abstract

In recent time particular interest has been drawn to the inference of gene regulatory networks from microarray gene expression data. But despite major improvements with data based methods, the network reconstruction from expression data alone still presents a computationally complex (NP-hard) problem. In this work it is incorporated additional information – regulatory motifs from yeast, when inferring a gene regulatory network for mouse genes. It was put forward the hypothesis that regulatory patterns analogous to these motifs are present in the set of mouse genes and can be identified by comparing yeast and mouse genes in terms of sequence similarity or Gene Ontology (The Gene Ontology Consortium 2000) annotations.
In order to examine this hypothesis, small permutations of genes with high similarity to such yeast gene regulatory motifs were first tested against simple data-driven regulatory networks by means of consistency with the expression data. And secondly, using the best scored interactions provided by these permutations it were then inferred networks for the whole set of mouse genes.
The results showed that individual permutations of genes with a high similarity to a given yeast motif did not perform better than low scored motifs and that complete networks, which were inferred from regulatory interactions provided by permutations, did also neither show any noticeable improvement over the corresponding data-driven network nor a high consistency with the expression data at all.
It was therefore found that the hypothesis failed, i.e. neither the use of sequence similarity nor searching for identical functional annotations between mouse and yeast genes allowed to identify sets of genes that showed a high consistency with the expression data or would have allowed for an improved gene regulatory network inference.

2. Background

"The ultimate goal of bioinformatics, genomics, proteomics and system biology is to be able to model a living system" (IIT Research Institute 2003), i.e. to capture all the biological mechanisms and processes that compose the system. Now in the postgenomic era that many genomes have been sequenced, the genes that represent the basic components of such systems have mostly been identified and the next step would be to understand how these genes relate to the biological processes of the corresponding system by investigating the functions of all the gene-products (Kustra et al. 2006).
Many important biological processes (e.g., cellular differentiation during development, aging, disease aetiology etc.), however, are very unlikely controlled by a single gene instead by the underlying complex regulatory interactions between thousands of genes (Li et al. 2006, Silvescu and Honavar 1997). Accordingly the task of modeling the biological processes of a living system requires the identification of the corresponding genetic relationships. As a consequene whole systems of gene interactions are increasingly studied (Mazurie et al. 2005, Baitaluk et al. 2006) to gain insights into biological mechanisms of functional dependencies / regulatory relationships and to understand how genes work together in both healthy and abnormal cells.
In this rapidly evolving research field particular attention has been drawn to the reconstruction of gene regulatory networks for model organisms (Herrgard et al. 2003; Banerjee and Zhang 2002; Stormo and Tan 2002; Wyrick and Young 2002), which aim to capture and depict the interrelated regulatory patterns among genes (Noman and Iba 2005). Development of high-throughput experimental techniques such as location analysis (Ren et al. 2000; Iyer et al. 2001; Lee et al. 2002a) and genome wide expression profiling (DeRisi et al. 1997; Eisen et al. 1998; Hughes et al. 2000) has recently allowed to explore

these dependencies of genetic regulation and provide meaningful data as a promising base for rapid reconstruction of underlying networks (Herrgard et al. 2003). Since then the inference of gene regulatory networks, in particular from gene expression data obtained by DNA microarray, has become one of the major tasks in the field of bioinformatics (Mimura and Iba 2001, Iba and Ando 2001, Maki et al. 2001).
DNA microarrays, as one of the most accessible genome-wide methods for expression profiling (Kustra et al. 2006), are used to produce large sets of measurements representing the gene expression levels (mRNA levels) of most, if not all, of the genes of an organism simultaneously (Rougemont and Hingamp 2003, Silvescu and Honavar 1997). As the technology advances, microarray experiments are becoming less expensive (Pan 2002) and are routinely performed to examine gene expression (Schena et al. 1995, Sherlock 2001) not only allowing for an increasing number of experiments but consequently also resulting in a gathering of more and more data that is abundantly available from the Gene Expression Omnibus (GEO) and various Websites (Radivoyevitch 2005). Pathway specific analyses of gene-gene correlations across these datasets, however, remain relatively unexplored, though they could be informative (Radivoyevitch 2005). This fact can be attributed to two problems:

1. Modelling regulatory relationships: Gene regulatory networks are complex biological systems, which are dynamic and highly nonlinear in nature and comprise of many interacting components (Noman and Iba 2005). And because of poor understanding of these biological components, their dependencies, interaction and nature of regulation grounded on molecular level, it is difficult to model these complex mechanisms mathematically (Noman and Iba 2005; Sakamoto and Iba 2001).
2. Inferring regulatory relationships: Inference of gene regulatory network models from expression data has shown to be a computationally complex problem (Oliveira et al. 2007), because of a huge search space of possible network topologies and the lack of appropriate evaluation methods able to initialise or optimise models in a way that would allow to satisfactory resolve the search space, as explained below.

2.1. Previous work on modelling regulatory relationships

In past years a great variety of representations has been put forward as potential frameworks to model the complex nature of regulatory networks and has also been used and examined in combination with gene regulatory network inference from gene expression data. (Herrgard et al. 2003).

The most common models to be named in this context include:

- abstract Boolean models (Hakamada et al. 2001; Akutsu et al. 2000b),
- statistical models (Hakamada et al. 2001; Akutsu et al. 2000b) such as Bayesian networks (Herrgard et al. 2003; Hartemink et al. 2001; Pe'er et al. 2001)
- quantitative models (Hakamada et al. 2001; Akutsu et al. 2000a) such as S-Systems (Spieth et al. 2004; Noman and Iba 2005),
- linear, causal models (Herrgard et al. 2003; Yeung et al. 2002; Tegner et al. 2003) such as path diagrams,
- combinatorial / hybrid models (Hakamada et al. 2001; Trey et al. 2001; Ideker et al. 2000).

Each of these models exhibits, depending on the nature of its representation, different advantages and limitations with respect to the capability of depicting regulatory relationships in a meaningful and close-to-real way. Due to specific structural and

parameter requirements, i.e. the type and amount of data needed to construct such a model, the chosen representation furthermore also affects the process of inferring the respective model from expression data.

Boolean networks: A Boolean network can be represented as a directed graph and is defined by a set of nodes whose states are determined by other nodes trough a list of Boolean functions (Wosik 2004). When utilized as a genetic network model, these nodes correspond to genes, whereas gene expression levels are discretized into binary states (1 and 0); a gene is either ON or OFF (van Someren et al. 2002). The Boolean functions represent the nature of regulatory interactions between genes, relating the expression state of each gene to the expression states of some other genes, using logical rules, e.g. AND, OR and NOT (Wosik 2004).
Boolean networks can be reconstructed with only small computational effort (Spieth et al. 2004), but their major disadvantage is the restriction to binary states, leaving the representation rather discrete and causing loss of information when translating continuous-value expression levels to binary expression levels.

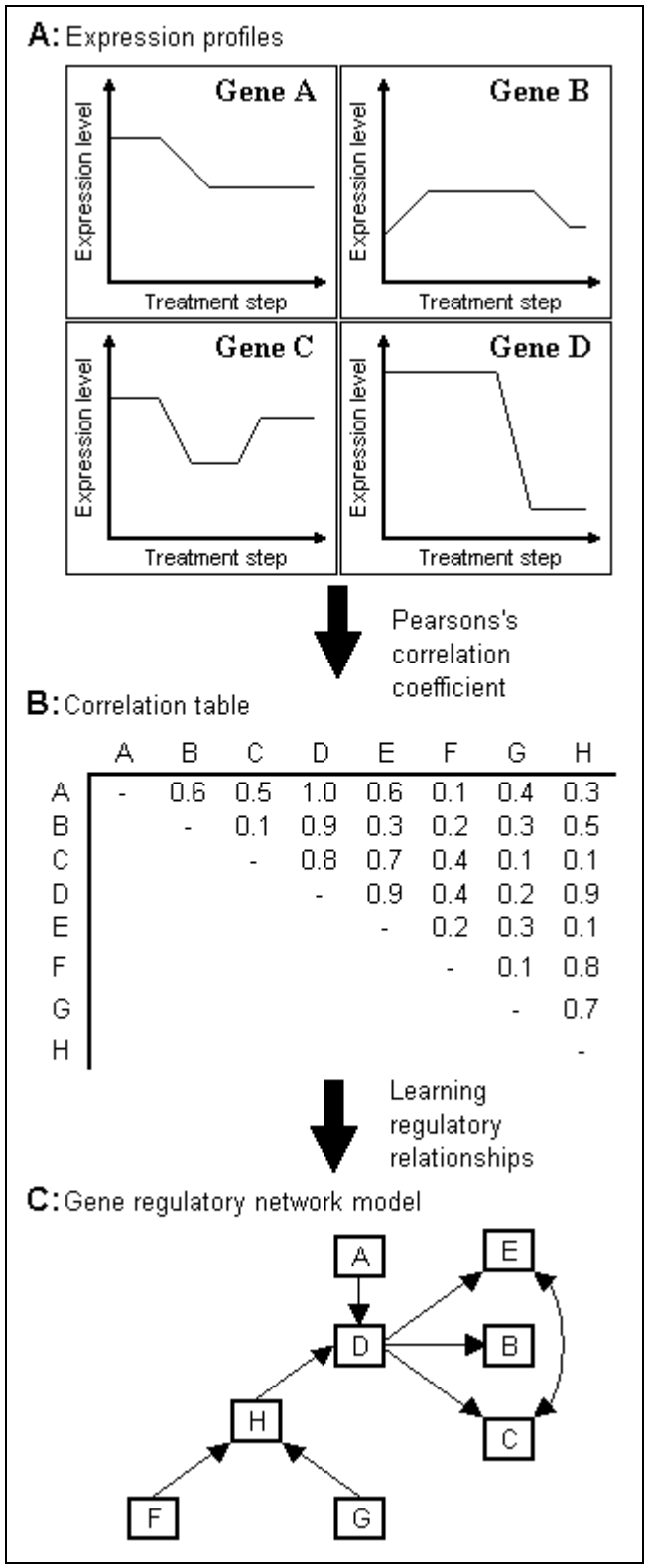
Bayesian networks: A Bayesian network is a representation of a joint probability distribution over a set of random variables and consists of two parts (Dojer et al. 2006):
1. A directed acyclic graph whose nodes correspond to random variables and edges indicate conditional dependence relations.
2. A family of conditional distributions for each variable, given its parents in the graph.
When utilized as a model for genetic regulatory networks the nodes represent genes and their expression levels, the edges illustrate interaction between genes and conditional distributions describe these interactions (Dojer et al. 2006).
The problems with Bayesian networks are as follows (Dojer et al. 2006):

- when only relying on expression levels, the origin and the target of an interaction become indistinguishable, meaning that several networks with the same undirected graph structure but different directions of some edges may represent the same distribution.
- Probabilities are discretized into binary values (true and false) disregarding potentially useful information.
- Bayesian networks do not allow for cyclic networks, and are therefore lacking the power of describing feedback loops of genetic regulation.
- the dynamics of a gene regulatory system is not taken into account. (Dojer et al. 2006).

S-Systems: S-Systems are a type of power-law formalism and can be described by a set of nonlinear differential equations, which are determining the structure of a gene regulatory network by describing expression inducing and degrading dependencies between genes. (Spieth et al. 2004)
Among the familiar models for describing biochemical networks S-system is rich enough to reasonably capture the nonlinearity of genetic regulation (Noman and Iba 2005; Savageau 1991), but such non-linear methods face a severe disadvantage of having many

more model parameters to be inferred from expression data and accounting for rather high computatial costs for network reconstruction (Spieth et al. 2004).

Path diagrams: The path diagram is a model representing causation and relation between independent, intermediary, and dependent variables (including error terms, which account for changes in a variable that cannot be explained by the influence of prior variables alone) (Garson 2006). Independent or also called exogenous variables in a path model are those with no explicit causes (no arrows going to them, other than the measurement error term and double-headed arrows, which are connecting and indicating a correlation between two exogenous variables) (Garson 2006). Endogenous variables, then, are those, which do have incoming arrows (Garson 2006). Those variables include intervening causal variables and dependents.
Intervening endogenous variables have both incoming and outgoing causal arrows in the path diagram. The dependent variables have only incoming arrows (Garson 2006).
Single arrows indicate causation between exogenous or intermediary variables and the dependent(s) and the path weights accompanying these arrows are standardized regression coefficients showing the direct effect of a variable on the one it is linked to (Garson 2006).
As a potential form of an abstract model for a gene regulatory network, path diagrams are able to depict genetic regulatory relationships by representing the direction and strength of causal effects in the expression levels of different genes.
A causal effect for two genes *A* and *B* ($A \rightarrow B$) in this case is assumed, if a change in the expression of gene *A* directly also results in a change of the expression of gene *B*, all other factors



	A	B	C	D	E	F	G	H
A	-	0.6	0.5	1.0	0.6	0.1	0.4	0.3
B		-	0.1	0.9	0.3	0.2	0.3	0.5
C			-	0.8	0.7	0.4	0.1	0.1
D				-	0.9	0.4	0.2	0.9
E					-	0.2	0.3	0.1
F						-	0.1	0.8
G							-	0.7
H								-

Figure 1.1. Steps towards inferring a gene regulatory network from expression data. From the expression profiles **(A)**, which describe the changes in the expression levels of genes during microarray experiments, it is estimated the correlation coefficients **(B)** between each pair of genes by Pearson's correlation coefficient. Common network inference methods then attempt to learn regulatory relationships from this data in order to infer the gene regulatory network **(C)** underlying the data-set.

kept constant.
The disadvantage of path diagrams however is their restriction to linear dependencies, which do not allow to capture the real complex and non-linear nature of gene regulatory networks.

2.2. Previous work on inferring regulatory relationships.

Given a model of gene interactions, the problem of gene network inference is equivalent to learning the structural and functional parameters from the time series or steady state of the microarray data, which are representing the gene expression kinetics, i.e. the network architecture is reverse engineered from its activity profiles (Noman and Iba 2005).
For data-based approaches a basic step in extracting such regulatory relationships from DNA microarray data involves the detection of similar expression patterns among genes in microarray data, compare figure 1.1 A. Similarity in expression profiles commonly point at dependencies between and highlight genes, which interact jointly in order to achieve a specific function within the organism (as a reaction on the specific treatment used in the microarray experiment) (Yona et al. 2006). Identification of similar expression patters thus provides crucial information aiding the process of linking genes and hypothesizing causations for the generation of a model.
The similarities in expression profiles from raw microarray data are commonly examined, e.g. by e.g. Pearson's correlation coefficient (Mudelsee 2003), in order to determine gene correlations, meaning the strength of linear relationships between genes, compare figure 1.1 B.
The above-mentioned complexity of the inference problem arises however during the attempt to interpret correlation data in means of learning regulatory relationships, compare figure 1.1 C. The major problem in particular is the fact that correlation data, obtained from e.g. Pearson's correlation, indeed emphasizes interactions between genes, but does not directly provide any further information about the underlying network that account for these dependencies. More precisely one can neither learn the direction of a regulatory causation nor distinguish direct from indirect regulation, since there could be different reasons, why a set of genes show high correlation values to each other, compare figure 1.2.
Due to this uncertainty with respect to the nature of a gene-gene regulation, i.e. direct or indirect relationship and direction of the regulation, data-driven model inference has to deal with two problems:
1. Initializing models by focusing on the strength of correlations, e.g. defining a threshold level and suggesting direct regulatory causations only between genes who's correlation value exceed this level, commonly results in rather erroneous approximations to possible solutions.

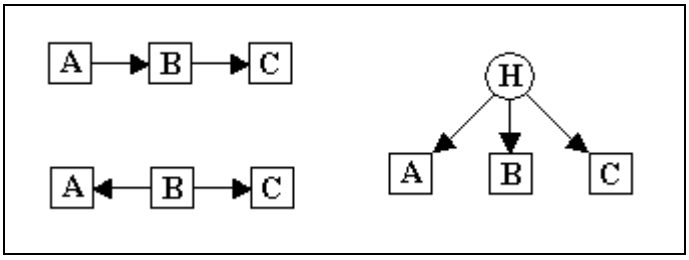


Figure 1.2: Three different graphs representing high correlations between the genes A, B, C with H as non-observable or hidden gene (Markowetz 2005).

2. Once a model has been initialized, its accuracy can be checked against the dataset, but lacking the ability of interpreting the correlation data in a sufficient way, it is almost impossible to exactly identify and eliminate incorrect relationships that cause faulty models. Hence, optimizations to a model have usually to be done in a systematical fashion, i.e. testing one model after the other.

In general, caused by the inability to initialize or optimize a model properly, the network reconstruction task is simply hampered by the enormous number of potential regulatory network structures that are generated and must be searched in order to identify the structure that is most consistent with the data sets (Herrgard et al. 2003).

A number of techniques, e.g. genetic algorithms (Iba and Ando 2001) or simulated annealing (Wang et al. 2004), have been utilized for this problem (Kyoda et al. 2000), each representing a different approach to dealing with and resolving the huge search space of possible solutions.

Genetic algorithms: Genetic algorithms are considered global search heuristics and as one class of evolutionary algorithms they are based on principles of biological evolution (Wong and Wong 1996). More precisely the method commonly starts with a random initial population of individuals, each describing a possible solution to a problem, that is modified and processed trough generations by mutations, recombination (crossover) and selection until a convergence is reached (Wong and Wong 1996; Schwarzbach and Börner 2001).

Genetic algorithms present a case of random search that is however, by following such stochastic evolutionary strategies, directed enough to resolve rather big search spaces.

Simulated annealing: Simulated annealing is another example of a global search heuristics applied to the inference and optimisation problem of gene regulatory network models. The basis of this method is given by its analogy to thermodynamics, i.e. the consideration of the fact that slow cooling of a metal allows its atoms to arrange in a state of minimum energy, while fast cooling inhibits the proper arrangement and leaves the metal in an undesirable (local) minimum of energy (Schwarzbach and Börner 2001; Eberl 1995).

The three fundamental components to simulate this concept are as follows (Eberl 1995):

- Possible solutions to a problem can be referred to as the state of the physical system.
- An error term for each possible solution describes the energy of the respective state.
- Temperature in the method is simulated by some control-parameter.

At higher temperatures even big increases of energy are allowed, helping the system to jump out of local minima, and with decreasing temperature only smaller increases or decreases of error are allowed, forcing the system to the bottom of a minimum (Black 2004).

But, despite significant progress in these data-driven reconstruction methods, the combinatorial expansion in the number of potential network structures still presents a major challenge for network reconstruction (Herrgard et al. 2003). To infer networks efficiently while avoiding numerous local solutions, it can therefore be reasoned that human intervention such as employing heuristics and adding sufficient search restriction or data seems to be necessary (Iba and Mimura 2002).

This work describes research on such an alternate network inference method, basing on the examination of regulatory relationships from a more well studied species in order to reconstruct a gene regulatory network model for a less understood organism. The expression data, for which a gene regulatory network was to be inferred in this project, is obtained by a DNA microarray experiment on genes of knock-out mouse (*Mus musculus).* The microarray chip contained 22690 probe sets and the experiment covered control and 6 case mice subjects.

Despite the fact that mouse (*M. musculus*) is one of the most studied organisms, besides e.g. fruit fly (*Drosophila melanogaster*), *Caenorhabditis elegans, Escherichia coli*, Zea mays and yeast (*Saccharomyces cerevisiae*), large parts of its gene regulation remain to be

determined. Years of experimental and computational molecular biology have however elucidated the gene regulatory networks of *E. coli* and *S. cerevisiae* (de Hoon and Vitkup 2005). Lee et al. published the dependencies between most of the transcriptional regulators and corresponding genes among the genome in *S. cerevisiae* (Lee et al. 2002a). Just as maps of metabolic networks describe the potential pathways that may be used by a cell to accomplish metabolic processes, the overall network of regulator-gene interactions that was discovered by Lee et al. describes potential pathways yeast cells can use to regulate global gene expression programs (Lee et al. 2002a).

The genetic regulatory relationships revealed in this network might just prove helpful for the reconstruction of network models of yet unknown species. This assumption can be motivated by the fundamental theory of evolutionism, according to which all species evolved from a few or maybe only a single common ancestor by means of natural selection, i.e. any pair of two species is believed to have a specific ancestral form or species from which both species then diverged differently during evolution, i.e. by adopting mutations.

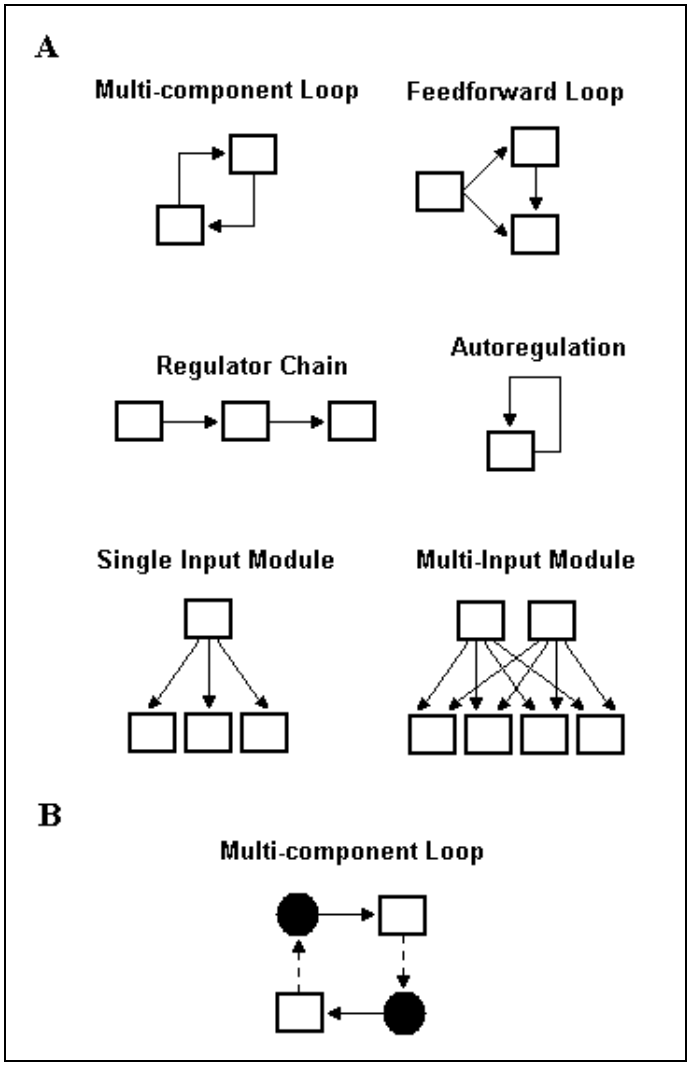


Figure 1.3. Network motifs identified among the regulatory map of yeast by Lee et al. (2002a). **A**: Rectangles represent genes and arrows indicate that the regulator encoded by the respective gene is binding to the promoter region of another gene (Lee et al. 2002a). **B:** The multicomponent loop in more detail. Genes are again represented by rectangles, black circles are regulators. Dashed lines illustrate that a regulator is encoded by the corresponding gene and a solid line means that the regulator binds to the promoter region of the according gene.

Examination of gene and protein evolution has however revealed that mutations are most likely not adopted and established with the same rates everywhere (University of Chicago Medical Center. 2005). When for example looking at proteins, mutations to secondary structure elements, which can be linked to function and structure of the protein, are only accepted at a slow rate, leaving those regions highly conserved over time, while parts of the protein not relevant for the function accept mutations more easily. So mutations are only adopted if they still allow the corresponding gene to perform (or improve its function), since otherwise the loss of its function or the resulting breakdown of complete regulatory cascades might result in an individual that is not viable, and the causing mutation could not be inherited (University of Chicago Medical Center. 2005).

As this is assumed for the evolution of proteins it behaves the same way for the whole genome. More important parts of the genome, encoding for crucial functions and behavior that are maintaining the viability of the organism, would show lower mutation rates than genes with less crucial information.
Accordingly, biological features could be conserved and nearly unmutated even for far diverged species, while other features differ even for closely related organisms. In other words, although yet unclear to what degree the knowledge based on well-studied organisms could reflect biological networks occurring in nature as a whole (de Hoon and Vitkup 2005), the mouse genome can be expected to hold genes with similar functions interacting to perform similar processes like some regulatory patterns discovered in yeast.
When however compared to the rather close relation between e.g. *M. musculus* and *Rattus norvegicus,* even tough both eukaryotes, *M. musculus* appears to be quite unrelated to *S. cerevisia.*
Finding on the other hand both organisms to share whole regulatory pathways, like the one modeled by the regulatory map presented by Lee et al., might therefore be quite unlikely, hence such large networks seem to be too large and complex to be conserved in total. So extracting information from the network discovered by Lee et al. might consequently more be a process of finding a medium, i.e. smaller fragments of regulatory interaction of genes that still are conserved, but on the other hand are also big enough to hold some significant information like typical arrangements of genes with respect to regulation that could be fundamental for yeast as well as for mouse.
Just such patterns might be represented by the set of transcriptional regulatory network motifs, the simplest units of network architecture, which Lee et al. identified among the regulator-gene interaction map of *S. cerevisiae* (Lee et al. 2002a), compare figure 1.3.

2.3. Objectives

The inference approach introduced and evaluated in this work is based on the hypothesis that incorporation of knowledge about these motifs allows to propose patterns or single instances of regulatory interactions between mouse genes, which show a high consistency with the expression data and thus allow to improve the inference of a whole gene regulatory network for the given dataset of mouse genes.
The experiment conducted to examine this hypothesis can be separated into four major steps:

1. Estimating a measure of similarity between selected mouse and yeast genes (using sequence similarity and comparison of functional annotations).
2. Basing on the similarity scores as approximative indications for similar biological roles / regulatory interactions of genes, permutations of high scoring mouse genes were generated in order to represent regulatory patterns potentially analogous to the corresponding reference motif from yeast.
3. Permutations were tested for their consistency with the real correlation data in order to evaluate the two similarity measures (sequence similarity and comparison of functional annotations) with respect to their capability of proposing patterns of genes with meaningful regulatory information.
4. Regulatory relationships were extracted from permutations and used as framework, when initializing a gene regulatory network model from the set of mouse genes.
5. The resulting models were examined against the performance of a purely data-based inference approach.

3. Methods

3.1. How to incorporate information

In case the motifs discovered by Lee et al. are not only basic patterns found in the genome of S. cerevisiae but can also be expected to be fundamental components of the mouse regulatory network, what kind of approach is suitable to detect instances of those network architectures among the given list of mouse genes?
Besides the knowledge about the general structure of those patterns, discovered from the yeast genome, Lee et al. also provides lists of yeast genes that were found to interact and building up instances of such network motifs (Lee et al. 2002b). Figure 1.4 represents two example sets of genes, one found to be arranged in a pattern described as a regulator chain and one building up a single input module, and how these array-representations translate to actual network topologies.

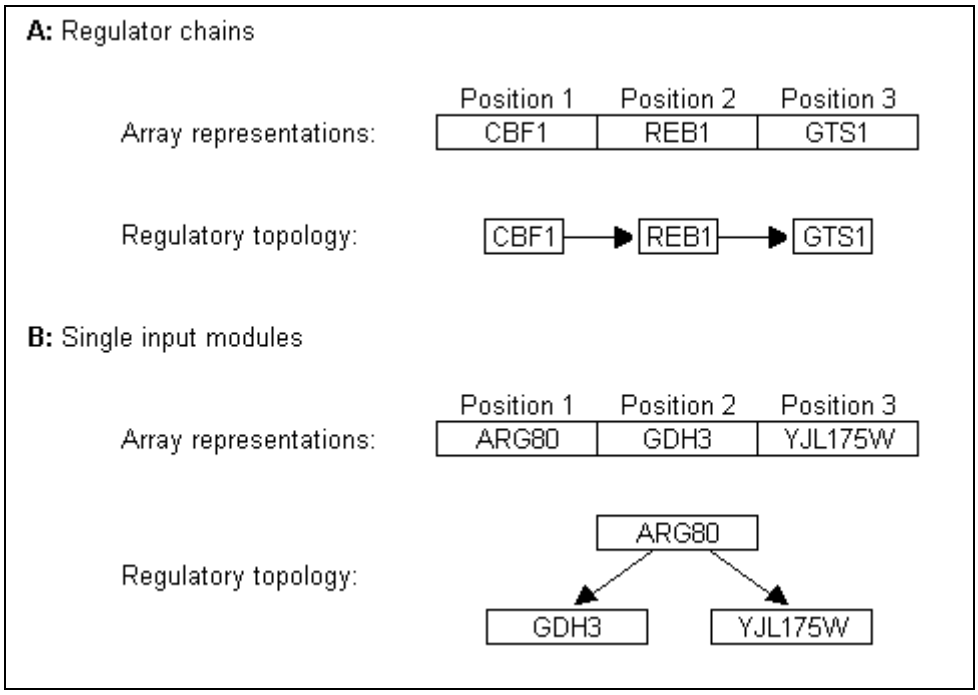


Figure 1.4. Example sets of genes found to be arranged in specific regulatory patterns by Lee et al. (2002b). When talking about positions in a motif it is referred to the array or list representation given here to describe an instance of a given motif class. **A:** A list representation of a set of genes describing a regulator chain motif and the corresponding regulatory topology translated from this list. **B:** A list representation of an instance of a single input module together with the translation to the corresponding topological representation.

The different topologies of regulatory interactions, as represented in figure 1.3, are further referred to as motif classes, whereas specific sets of genes that were found to interact in a corresponding pattern, compare figure 1.4, are referred to as motifs or instances of a motif-class, i.e. particular representations of given motif classes.
The two motif classes presented in figure 1.4, are the ones that were used later during the experiment. The complete list of Single Input modules (SI) covered 1597 yeast genes that build up to 89 instances of such motifs from size 2 up to 212 positions. The list of Regulator Chains (RC) contains 72 yeast genes that were arranged in 188 patterns of size 2 up to 10 positions.

Patterns of regulatory interaction as described by the general structure of such motifs, i.e. groups of genes interacting in the displayed topology, might occur in a dataset quite frequently, but the identification of such patterns is restricted to sets of mouse genes that are detectable because they show some similarity to one of the instances given for the respective motif topology. Meaning it is not searched for general topological structures among the dataset, but for conserved functional/regulatory patterns between the two species, whereas the instances provided by Lee et al. (2002b) serve as templates or references. In the following such a configuration of mouse genes believed to be a somehow conserved representation of a given yeast motif in the mouse data is referred to as an analogue motif in the set of mouse genes.
In order to identify regulatory patterns analogous/similar to those motif-instances among the set of mouse genes, one approach would be to look for some characteristics that are revealed by the yeast genes of a motif instance and search for such characteristics in the mouse genes used in the experiment, i.e. if the yeast genes that are building a motif have specific properties, an analogous motif in the mouse genome could be expected to have a set of genes that shows similar properties. When searching for conserved patterns, in the ideal case these genes are even orthologs to the compared yeast genes, whereas orthologs in this case are regarded genes from two species that show highly similar or even identical functions and have diverged from one common ancestral form by speciation.

In practice this means that each mouse gene provided in the data set needs to be compared to each yeast gene, which holds a position in a motif, to estimate how similar those two genes are in terms of the characteristics in question. Similarity in this case allows for assumptions or approximations about how probable it would be to find a specific mouse gene taking the same position in an analogous motif (if existing) among the set of mouse genes, like the corresponding yeast gene does in the yeast genome.
To ensure comparability, e.g. to distinguish if a mouse gene for instance would fit better into a motif at position *A* or at position *B*, the measurement of similarity needs to be consistent for each comparison of a mouse and a yeast gene.

The characteristic or property of a gene, which can be seen as the best indication for its biological role and interactions to other genes, is its function. Accordingly it seems reasonable to base the comparison and scoring of genes on some criteria describing their functions. Functional annotations provided by the Gene Ontology (GO) (The Gene Ontology Consortium 2000) represent such descriptions of functions and were therefore chosen as basis for matching mouse and yeast genes.

Gene Ontology: The Gene Ontology Consortium provides a structured standard vocabulary for describing the function of gene products and the Gene Ontology (GO) itself is divided into three orthogonal ontologies, biological process, molecular function, and cellular component, each represented as a directed acyclic graph (DAG) with nodes corresponding to functional terms and relationships between terms illustrated as edges (The Gene Ontology Consortium 2000, Schlicker et al. 2006).
These functional or so called GO terms allow for coherent annotation of gene products and thus a basis for new methods that rely on comparisons of those products regarding their molecular function, biological role or cellular component (Schlicker et al. 2006).

Two genes with the same GO annotations or GO annotations that refer to closely related functional terms therefore can be expected to have a similar function, i.e. the more identical or highly related annotations both genes have in common, the more similar should their overall function be, and they consequently can also be assumed to participate

in similar mechanisms when they belong to the same organism or to hold analogous roles in biological regulations when present in different organisms.
Scoring by use of GO annotations is thus done by comparing annotations of a mouse gene to the annotations of a yeast gene that is found at a specific position in a motif instance, and the level of similarity that can be assigned this way then corresponds to the score of the mouse gene to hold exactly this position or represent the same node in such a network structure.

In this work, two different approaches were proposed to incorporate GO annotation terms and compare any pair of mouse and yeast gene by means of functional similarity. The first method can be seen as a direct comparison between two genes, by identifying identical functional annotations among them. In the second method the functional similarity of two genes is estimated by determining similarities and relations between annotations. The implementation of the corresponding semantic similarity measure however revealed to be too time-consuming even when generating a substitution table to provide similarities between annotations. Consequently only the first method was used:

GO functional identity: Since each GO annotation represents a specific property (Biological process (BP), Molecular function (MF), Cellular component (CC)) the percentage of shared annotations for a pair of mouse and yeast genes can be directly linked to the number of shared functions/characteristics between those genes. So the number of equal annotations between two genes might further be seen as a direct indication of how probable it would be to find the mouse gene having a quite similar biological role and hence taking the position of the corresponding yeast gene in an analogous motif among the mouse genome.
The percentage of shared annotations between a mouse and a yeast gene is calculated separately for each ontology (BP, MF, CC) and the score for the particular gene pair is then determined as the average percentage of these three single percentages. In detail the percentage of shared annotations for each ontology is estimated as the number of identical annotations between the mouse and yeast gene, divided by the number of unique annotations from both genes in total. E.g. if there are 3 annotations for the mouse gene and 5 annotations for the yeast gene and 2 of these annotations are identical, then the total number of unique permutations would be 6 and the percentage of shared annotations would be about 33 % (2 devided by 6).
Because the measured percentages are absolute values of functional identity (0% = no shared annotations, 100% = exactly the same annotations), high values of course increase not only the confidence about similar functions, but might also point at similar biological roles between genes and provide thus a possible basis for the assumption of genetic interactions of an unknown gene (by assuming the particular mouse gene to interact with mouse genes similar to the yeast genes that are regulated by the one yeast gene, which obtained the highest score compared to the particular mouse gene). Gene pairs with only a few shared annotations, especially for common annotations, might on the other hand occur quite frequently and are most likely not desirable for further examination, i.e. not significant enough to allow for a conclusion or might even be misleading.

Sequence Similarity: Using only the GO identity approach as it is described above brings two problems: Firstly it has to be considered, that GO does not provide annotations for all genes, meaning that particular genes can just be excluded from this analysis. Secondly and more important, the direct GO identity approach is restricted to the comparison of identical annotations, i.e. it can only propose two genes to be similar in function, when they show identical functional annotations. But because genes between mouse and yeast

can be expected to be diverged, genes with similar biological roles or even homologs can be expected to exhibit similar and related functions rather than still only identical ones. The GO identity approach alone without the GO similarity approach is therefore quite restricted and likely to miss important information. Consequently another approach was proposed, a measure of sequence similarity that allows to assign an alternative score for a combination of mouse and yeast gene according to the similarity of their protein sequences. Sequence similarity and similarity in protein function cannot always be directly linked to each other, meaning that on the one hand a low sequence similarity not always indicates that functions are far distant from each other and on the other hand a high similarity does not automatically imply a high similarity between the corresponding functions. However, when looking at the evolutionary process of diverging proteins it can be said that most secondary structures and in particular parts forming the active centre remain highly conserved over time. The reason for this observation is given by the fact that mutations to secondary structure elements affect the overall fold of the protein in a stronger way, than mutations in loop regions. Since small changes to the structure, especillay the active centre, could cause the protein to loose its functionality, mutations to secondary structure elements are therefore hardly adopted. Sequences of related proteins are thus closer related in regions of similar secondary structures and high similarity based on highly conserved regions therefore also results in better chances for matching protein functions.

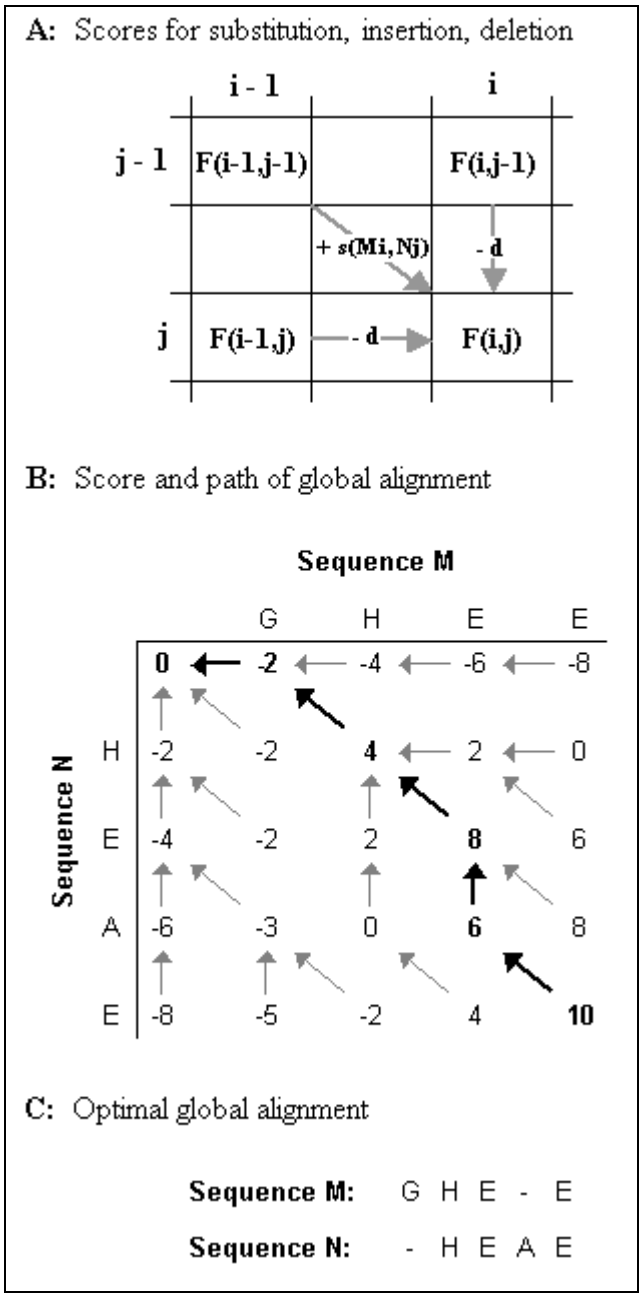


Figure 1.5. Gobal alignment by the Needleman-Wunsch algorithm. **A:** Scores for substitution (aligning aminoacids M_i with N_j), insertion (aligning M_i with a gap) and deletion (aligning N_j with a gap). i and j are positions in sequence M and N, respectivly, and $F(j,i)$ describes the score of matching residues i and j. d is the gapcost and $s(M_i,N_j)$ describes the score for substitution of i with j. **B:** The lower right corner shows the score of the optimal global alignment (= 10) and highlighted arrows show the decision made to derive this score. **C:** The optimal global alignment translated from trace backing trough the decisions shown in B.

As an alternative and compared to the GO similarity approach, which was not used because of an inadequate implementation and has to be replaced by this technique, sequence similarity still represents a more imprecise and vague method. Techniques like BLAST (Altschul et al. 1997), which represent rather fast methods capable of estimating a measure of sequence similarity between two sequences were therefore avoided, because they don't

guarantee to find the best alignment between two sequences and thus yield an even more approximative score and indication for similar functions.
A Smith-Waterman algorithm on the other hand determines the best scoring similar segments between sequences. Local alignments of such regions point at conserved structural elements, which are very important for the fold and structure of a protein, but do not directly (without further examination) provide an overall score for the similarity of two sequences.
In order to estimate a measure of sequence similarity between the proteins encoded by each pair of mouse and yeast genes, it was therefore chosen to implement a Needleman-Wunsch algorithm (Needleman and Wunsch 1970). The substitution of amino acids was scored according to Dayhoff's PAM250 matrix (Dayhoff et al. 1978, Durbin et al. 1999) and the gap cost was set to –2, compare figure 1.5 A.
The score for the global optimal alignment between two proteins, compare figure 1.5 B, is then used as score for matching the corresponding genes. The alignment itself was not relevant and thus not produced for the work.
In particular when compared to the GO functional identity (1), sequence similarity seems to be a more approximating and vague method, whereas the measured similarities do not simultaneously allow for as clear results as in the identity approach.
The reasons for this circumstance or in other words the three main problems arising with this technique are as follows:

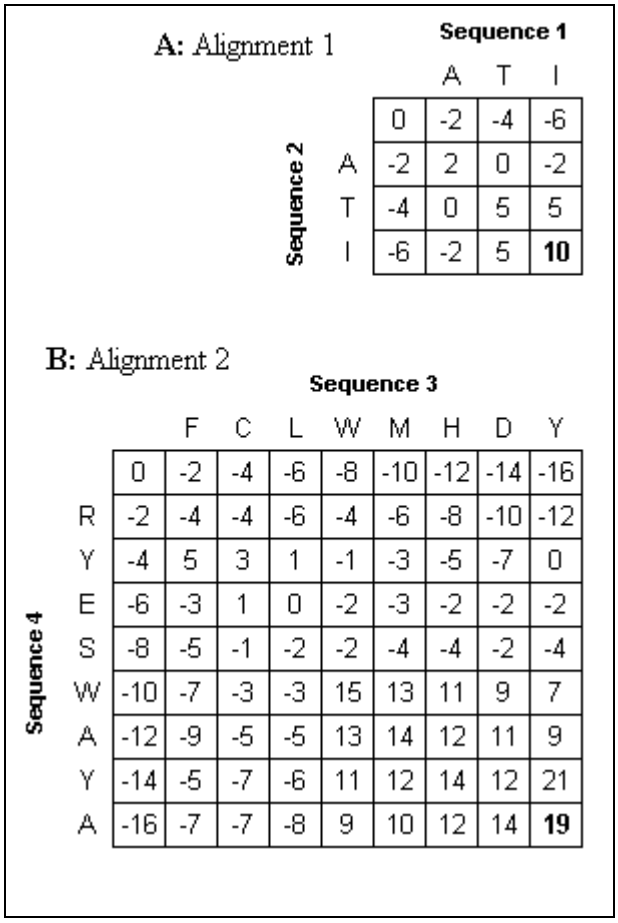


Figure 1.6. Sequence alignment scores in dependence of sequence length. A: Alignment of two small identical sequences and a global score of 10. B: Alignment of bigger, but rather unrelated sequences and a **global** score of 19.

1. As mentioned above, sequence similarity between two proteins is no absolute guarantee for the functions of both proteins to be similar as well, although higher similarities allow for more confident assumptions.
2. Even when disregarding this fact, and instead assuming that a higher sequence similarity always also indicates a higher similarity with respect to function, the corresponding scores would just allow for relative comparisons between the matching proteins (or respectively genes encoding the proteins), i.e. if the proteins encoded by two mouse genes (M_1 and M_2) are compared to the protein encoded by a yeast gene (Y) and the similarity of protein M_1 is higher, it is an indication that M_1 would, relative to M_2, be a better match for the motif position of Y. On the other hand, the calculated alignment scores do however not reflect any absolute percental similarity, meaning that even the mouse gene with the highest similarity to the corresponding yeast gene might be far distant to possess the same function, and

without further analysis of each alignment there is no way of inferring any absolute measure of similarity from sequence similarity.

3. Scores of sequence alignments also depend on the length of protein sequences as well as on their amino acid content, i.e. while a low score might represent a quite high sequence similarity if produced by an alignment of two short sequences, a lower similarity between two long sequences in contrast might although result in a significantly higher value, compare figure 1.6.

As consequence from 2. and 3. the results of this approach might tell that gene M_1 is more likely performing a similar biological role like the yeast gene Y_1 than M_2, but without further analysis these scores do neither state, with which probability it can actually be assumed that M_1 really has some significant similarity in its role and function to gene Y_1, nor do they allow to compare the similarities of different pairs of genes, e.g. thesimilarity of M_1 & Y_1 against the similarity of M_2 & Y_2.

For visualization, the results from both approaches would look like shown as example in figure 1.7:

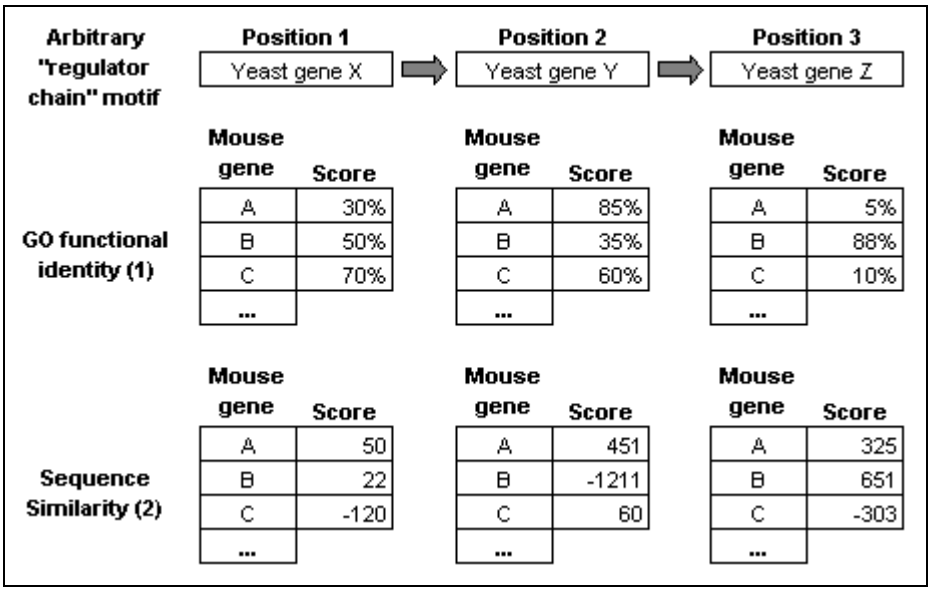


Figure 1.7. Illustration of examplescores generated by GO functional identity (1) and sequence similarity and (2). The theoretical scores outline the general appearance of results for both methods, when matching a set of three mouse genes *A*, *B*, *C* against the yeast genes *X*, *Y*, *Z* that combine to the presented motif.

3.2. How to use scores in order to propose regulatory relationships

On the supposition that a regulatory pattern, analogous to a the specific motif from the yeast genome, is present in the set of mouse genes, the major task as mentioned above is to find combinations/configurations of mouse genes that are most likely building up this pattern or show similar regulatory relationships.
With a measure of similarity as assigned by one of the above mentioned methods it is now given a measure that allows for approximate assumptions on how similar a mouse and yeast gene are with respect to their biological role. Based on these similarity measures it is now also possible to evaluate the similarity between a specific combination of mouse genes and a whole motif of yeast genes.

So the logical next step towards identifying potential regulatory patterns would be to combine and assess mouse genes as possible configurations with which a given motif could occur in the dataset.
A configuration in this case is described by any order of genes, where each gene represents one particular position within the given motif. And the fact that each gene is assigned a measure of similarity depending on it's position, taking the average of these scores also allows to determine how similar the whole configuration is compared to the original motif, compare figure 1.8.
In the project these configurations of genes were generated and examined as permutations, i.e. considering a pattern length of *n* positions and a pool containing *m* mouse genes, it were systematically build unique orderings of *n* mouse genes, compare figure 1.8., whereas an ordering is only allowed to contain a single gene once.

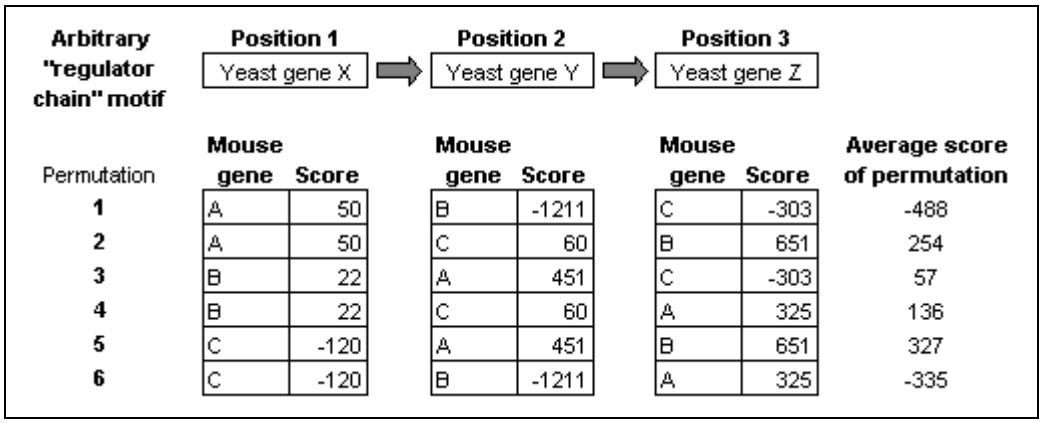


Figure 1.8. Illustration of construction and scoring of permutations. The figure shows possible orderings of length 3, generated from a pool of 3 mouse genes (*A*,*B*,*C*). As possible analogues to a given yeast motif, each gene possesses a score to be in a specific position, i.e. the score for mouse gene *A* to be in position 1 equals the similarity (as calculated by one of the similarity measures) between mouse gene *A* and yeast gene *X*, since *X* describes the properties desirable to have in position 1 of this motif. Each ordering is scored with respect to itsoverall similarity to the properties of the template motif by calculating the average of the scores that each gene gains for holding its specific position in the ordering.

These permutations illustrate every thinkable configuration of mouse genes for the given pattern, meaning if the specific regulatory pattern of the motif is present in the correlation table, then it is represented by one of the permutations, built this way.

3.3 Examination of Permutations

As described above, the problem with data-driven network inference is its restriction to estimate genetic causations purely from expression data, a process that is hampered by the uncertainty to identify the correct interaction from the enormous number of possibilities.
Motifs as network elements from the regulator map of yeast already depict real genetic relationships / causations, e.g. the arbitrary motif depicted in figure 1.8. contains two genetic interactions: 1. Yeast gene *X* induces the expression of Yeast gene *Y* and 2. Yeast gene *Y* induces the expression of Yeast gene *Z*.
In the ideal case, when generating permutations for each motif from all motif classes, all the causations in these motifs are considered and matched to the set of mouse genes by either functional or sequence similarity. And assuming that gene regulation is to some degree shared between yeast and mouse (and can be caught by functional or sequence similarity), the potential of permutations is given by the fact that they provide translations

of these causations for the set of mouse genes, i.e. they represent a scored list of possible causations for the set of mouse genes.
In general, such a list of permutations then allows for different approaches to extract or propose potential regulatory interactions, like including whole permutations or selecting particular promising causations from this list, when constructing gene regulatory networks. In order to produce any meaningful data, any approach for extracting information from these permutations, or even from any comparison between a mouse and a yeast gene, however requires the scoring to be proportional to real associated genetic regulatory relationships, rather than only indicating sequence or functional similarity. In other words a permutation with a higher score must represent a relationship more likely to be true than the relationships from a lower scored permutation, which can only be achieved if the used similarity measure catches relevant properties that are related to the regulation behavior of a specific gene.
So before implementing an approach for inferring a complete network from such permutations or genetic relationships proposed from any functional or sequence similarity, it seems advisable to examine the similarity measures with respect to their capability of relating to genetic regulatory behavior.
With a proper scoring method and a set of genes possessing the regulation patterns searched for, the highest scored permutations should be expected to reveal some proportion of these patterns, while low scored permutations do not.
Based on this theory, the evaluation of the presented similarity measures is done by comparing differently scored sets of permutations against each other and against data-based inferred networks by means of their consistency with the correlation table, which was obtained from real expression data.
If there are patterns present in the dataset that are analogous to the motifs, for which these permutations are generated, and the similarity measure is scoring the permutations because of related regulatory behavior, then the comparison of differently scored permutations and data-driven networks should yield results as follows:

- Higher scored permutations should also show a higher consistency with the expression data compared to permutations with lower scores. Such findings would indicate that the scoring is able to point out better assumptions for regulatory relationships.
- When generating initial models by a simple data-driven approach for the genes that are contained in the highest scored permutations, the data-driven networks as quite erroneous should show a lower consistency to the correlation table when compared to the permutations itself. These findings would indicate that the

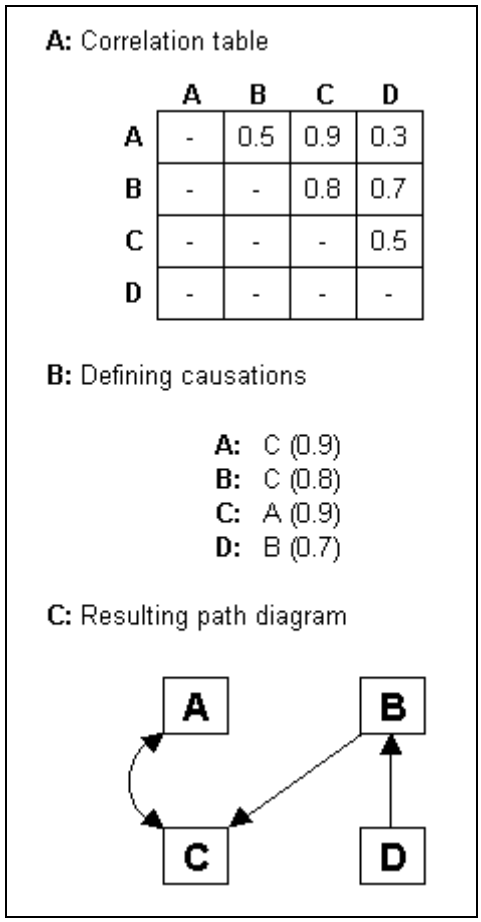


Figure 1.9. Example for the used procedure to infer gene regulatory networks from correlation tables. **A:** A correlation table containing 4 genes (*A*,*B*,*C*,*D*). **B:** For each gene it is set one causation to the particular gene, which exhibits the highest correlation value to it. **C:** The path diagram combined from the selected genetic causations.

permutations contain regulatory information useful for the reconstruction of the network for the given dataset.
- As a consequence low scored permutations should show no improvement over data-driven networks for the same set of genes.

If on the other hand higher scored permutations do not show an improvement over lower scored permutations or data-driven networks in terms of consistency with the expression data it could imply that the dataset does not contain a corresponding regulatory pattern or that under the assumption that such a pattern is present, the chosen similarity measures are incapable of detecting them.

The model representation chosen to infer data-driven networks to and to map out the particular permutations is the path diagram as it is described above, compare also figure 1.9 and figure 1.10.
The topology for such a network description of a permutation or a set of permutations can simply be copied from the respective permutations by selecting the genes as nodes and placing a directed edge between two genes whenever there is a causation for those genes in the permutation.

The topology of purely data-driven models for a given set of genes was inferred from the correlation table by an algorithm invented by Laurio et al. (2007): For each gene only one directed edge was placed, connecting the selected gene with the gene that showed the highest correlation value to it, compare figure 1.9.
Basing on the same algorithm by Laurio et al. (2007), weights of both models (derived from permutations and data-based) are then initialized randomly and optimized by a stochastic optimization algorithm to fit the model to the correlation table.
To optimize the weights as well as afterwards evaluate the final accuracy of the model, a method is needed to determine the correlation values that are produced by the model. (The error accuracy of the model is calculated as the difference between correlation values produced by the network and real measured values from the correlation table – mean squared error)

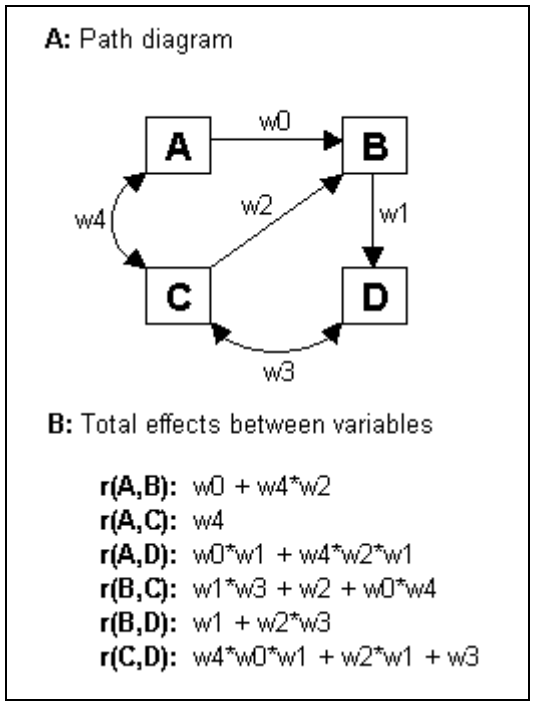


Figure 1.10. Example for application of path analysis. **A:** Simple path diagram containing 4 genes (*A*,*B*,*C*,*D*), 3 causations and 2 correlations. Strength of relationships are described by 5 weights (*w0*,*w1*,*w2*,*w3*,*w4*). **B:** Total effects between pairs of genes following the rules of path analysis.

The particular method applicable for evaluating linear path models and chosen for this purpose is path analysis.

Path analysis: Path analysis, is an extension of multiple regression and its purpose is to provide estimates of hypothesised causal connections between sets of variables linked together in a path model (Webley and Lea 1997). Since the path model illustrates direct effects and the strength of any of these effects is given by regression coefficients, the value of effect for any compound path, i.e. the indirect effect between two variables,

which are linked together by a set of intermediate variables, can also be calculated as the product of the coefficients connecting all the variables in the path (Garson 2006).
According to Sewall Wright (1934), the principal rules for inferring relationships from a path diagram were described: (Lessem 2002)

"Any correlation between variables in a network of sequential relations can be analyzed into contributions from all the paths (direct or through common factors) by which the two variables are connected, such that the value of each contribution is the product of the coefficients pertaining to the elementary paths. If residual correlations are present (represented by bidirectional arrows) one (but never more than one) of the coefficients thus multiplied together to give the contribution of the connecting path, may be a correlation coefficient. The others are all path coefficients." (Lessem 2002)

In other words, the expected correlation or total effect that one variable has on another in a path model can be derived by summing the effects of all compound paths, which connect both variables. The extraction of any potential compound path in this case has to be consistent with the following conditions. It is allowed to: (Lessem 2002)

- Trace backward along an arrow and then forward, or simply forwards from one variable to the other but never forward and then back
- Pass through each variable only once in each chain of paths
- Trace through at most one two-way arrow in each chain of paths

Based on these rules the error for a network in this thesis was determined as proposed by Laurio et al (2007). It is first estimated the correlation value for each gene-gene-combination that would be produced by the path diagram, compare figure 1.10, and comparing these values back to the original correlation table it is then possible to estimate the distance between the network induced causation strength for a gene pair to the one given in the correlation table:

$$\text{Distance}_{G_{ij}} = \text{PDr}(G_i, G_j) - \text{CTr}(G_i, G_j)$$

With $CTr(G_i,G_j)$ describing the correlation value between gene i and gene j that is obtained from the correlation table and $PDr(G_i,G_j)$ representing the correlation value between gene i and j produced by the path diagram. The error for each gene is then calculated by multiplying all of its distances to other genes:

$$\text{Error}_{G_i} = \prod_{j=1}^{i-1} \text{Distance}_{G_{ij}} * \prod_{j=i+1}^{n} \text{Distance}_{G_{ij}}$$

With n as total number of genes in the path diagram.
The error of the whole model is then defined as the sum of the errors of its genes:

$$\text{Error}_{\text{Network}} = \sum_{j=1}^{n} \text{Error}_{G_j}$$

3.4. Course of action

Following the above-mentioned principles, the steps performed in the project were as follows:

1. Two example motif classes were selected, which then served for evaluation of the similarity measures as well as representatives for the whole list of motifs in order to determine the presence and traceability of regulatory patterns with either approach, i.e. if this method produces an appropriate amount of data that would justify its use for network inference. The two motif classes chosen were "Single-Input Modules" (SI) and the list of "Regulator Cascades" (RC) with the following motivations: Motif instances that describe rather small patterns with only one or two genes, as found for the Autoregulation + Multi-component Loop, were not desirable for examination, because potential matches for these motifs might occur more likely by random chance than when using motifs of bigger size. The two motif classes additionally only showed a few instances (13 instances in total – in contrast, SI and RC showed 89 and 188 instances, respectively) and thus lacked an appropriate number of templates, i.e. focussing only on these 13 motif instances it is most likely not possible to identify an adequate amount of patterns among the set of mouse genes to allow for any network inference. From the other motif classes it was chosen to use the Single-Input Modules, because the corresponding motifinstances are also part of the Multi-Input Modules. So the examination of the Single-Input Modules also allows for a first assessment of both motif classes, i.e. if the Single-Input Modules do not provide useful data, the Multi-Input Modules will most likely fail as well. And the Regulator Chains were picked because they represent a very simple pattern of regulatory interactions, and can therefore be expected to occur and hence be found more often than feed-forward loops.
2. First, a small dataset containing 185 mouse genes was used, and after evaluation of the first results it was then decided to switch to a larger dataset of 1189 mouse genes. The 185 genes of the first set were selected from 22690 probe sets by a gene knock-out experiment with a maximum P-value of 0.05 in expression change between a set of 6 control and 6 case mice subjects, a minimum fold-change of 1.2 and a minimum intensity level of 75. The second dataset consisting of 1189 mouse genes was filtered from 22690 probe sets by a transgenetic experiment searching for significant expression changes (maximum P-value of 0.05) between 6 control subjects and 6 case subjects.
   Annotations for these mouse genes and the yeast genes contained in the two motif lists were downloaded from (The Gene Ontology Consortium 2000) and using the GO functional identity approach mentioned above a substitution table was created, stating the functional identity for each pair of mouse and yeast gene. The numbers of different annotations downloaded for the mouse genes were as follows: 738 for BP, 202 for CC and 544 for MF. For yeast the numbers of different annotations were: 727 for BP, 291 for CC and 563 for MF.
   Using this substitution table, the mouse genes were then combined to each position in each motif instance, and the number of occurrences for any gene matching a motif position with a specific score was counted. As a first assessment of the dataset, checking the occurrences of specific similarities between the motif instances and the mouse genes allows for first conclusions if the data set even contains any promising genes for such a regulatory pattern.
   These results allowed the calculation of the number of possible permutations that could be generated for each motif class (SI and RC) when considering different threshold levels of minimum shared GO annotations. More precisely, when

allowing only genes above a specific percentage of shared annotation to a yeast gene to be considered as a match for the corresponding motif position, it was estimated the number of motifs with at least one gene matched to each position and the total number of possible permutations from these motifs.
Considering the possible permutations for the particular threshold levels three sets of permutations were then chosen for list of SI's as well as three sets for the list of RC's. The number of selected permutations was based on the number of best scored permutations for the respective motif class, e.g. the best permutations for SI were found with scores between 65 and 70%. These 23 permutations were chosen and then evaluated against two equal sized sets of lower scored permutations of the same motif class in order to determine if the better scored permutations are more consistent with the correlation data (provide more real regulatory relationships) compared to low scored permutations.
The sets of permutations were chosen as follows:
SI: 23 Permutations including 24 genes above 65% shared GO annotations (the total number of permutations found for that threshold) and average score of 75.66%, 23 Permutations including 15 genes between 20 and 25% shared GO annotations and an average score of 22.22%, 23 permutations including 21 genes between 0 and 5% shared GO annotations and an average score of 3.84%.
RC: 42 Permutations including 12 genes above 51% shared GO annotations (the total number of permutations found for that threshold) and average score of 55.34%, 23 Permutations including 23 genes between 20 and 25% shared GO annotations and an average score of 20.78%, 23 permutations 21 including genes between 0 and 5% shared GO annotations and an average score of 3.53%.
In order to evaluate the used scoring method (functional identity by GO annotations) a path diagram was generated from each permutation and compared to a data-driven network, that was inferred for the same genes by using the approach mentioned above.

3. The sequence similarity measure was applied and evaluated on only 17 regulator cascade motifs each four positions long, compare step 4.
   Sequences were downloaded from (Ensembl 2007) for proteins encoded by mouse genes and from (Hong et al. 2007) for proteins encoded by yeast genes and a substitution table was created by using the sequence similarity method as described above.
   From the results four sets of permutations were generated, chosen from different intervals of scores:
   1. A set containing the 3 best permutations for each motif, covering 38 genes with an average sequence similarity of 654.75.
   2. A set containing 3 permutations obtained for each motif with medium high scores, covering 37 genes with an average sequence similarity of 317.36.
   3. A set containing 3 permutations from each motif with medium low scores, covering 41 genes with an average sequence similarity of –214.43.
   4. A set with the 3 lowest scored permutations from each motif, covering 33 genes with an average sequence similarity of –5673.54.

   As in step 2, the sets of permutations were translated to path diagrams and compared to each other as well as to data-driven networks for the respective sets of genes.
4. Finally the permutations were used to generate networks for the whole data set of 1189 mouse genes.
   For this purpose, permutations scored by shared functional annotations were first generated, as well as permutations scored by sequence similarity.

For the first case all permutations for SI and RC were created that incorporated genes with scores higher than 30%.
Since sequence similarity does not provide an absolute measure of similarity, using this measure does not as easily allow to set a specific threshold level to filter out low scored genes and reduce the number of permutations that need to be generated.
So, when using this similarity measure for building permutations from the given set of genes, on the one hand it would be desirable to incorporate all genes in order to not lose any valuable information, and on the other hand findings for bigger motifs can be expected to be more significant, while smaller motifs might show high scores without representing any real regulatory pattern in the mouse data set and could be simply random matches.
In other words it would be worthwhile to focus on larger motifs, because inferred regulatory patterns from these motifs might be more significant, but the number of possible permutations grows exponentially for each additional position. When e.g. looking at a motif with 10 positions, even the usage of only the 6 best genes for each position might in the worst case imply that more than 60 million permutations need to be generated.
Since the restriction to the 6 best genes might already be too selective and disregard potential information, but generation of permutations from this set simultaneously is yet too time-consuming to incorporate even more genes, a motif size has to be chosen as a medium, which is still big enough to yield significant results, but is on the other hand small enough to avoid high computational times, when building permutations on more appropriately sized gene sets.
So as a compromise between large motifs and large sets of genes for the generation of permutations it was chosen to limit construction of permutations on 17 motifs with 4 positions and the 30 best genes for every position.
In order to keep the results of step 3 related to the results obtained from this complete model inference, step 3 was also performed on these 17 motifs.
Since most of the bigger motifs within the list of regulator cascades consist of smaller RC motifs, the evaluation of smaller motifs, e.g. 4 positions as done in this work, should be sufficient to gain a fair insight into the usefulness of RC motifs in combination with sequence similarity in general, i.e. if the smaller motifs do not produce meaningful data, than the bigger ones will most probably fail, too.
For the inference of a complete network model these permutations were used as follows:
When picking an arbitrary mouse gene, the permutation with the highest score that contains this gene should describe the best suggestion for the causations of the gene in the data set. By iterating trough all the mouse genes of the dataset and using this method to select the causation(s) from the best scored permutation that contains the respective gene, a list of gene-gene interactions is acquired, which can be combined to construct a potential network for the set of genes.
Following this procedure a complete network was then created from the permutations scored by sequence similarity and one network from the permutations scored by shared functional annotations. Some genes within this sets of permutations, e.g. the last elements of regulator chains, have only incomming arrows, but do not provide any outgoing causations. The relationships for those genes were obtained from the given correlation data according to the method used to infer purely data-based networks, compare figure 1.9.
To examine the performance of both networks a data-driven network was also generated for the whole set of genes.

The evaluation was then done by estimating the consistency of each network with the real measured correlation data, i.e. by determining the correlations produced by the network and comparing them to the expected values. These error terms are an indication for the performance of each network/approach and are compared to each other.

4. Results

4.1. GO functional identity

The first used dataset contained 185 mouse genes, from which annotations could be provided for 165 genes and for the second dataset could be downloaded annotations for 1011 from 1189 genes. After creating substitution tables for each set by calculating the percentage of shared GO annotations for each needed mouse and yeast gene, the mouse genes were matched to the list of SI and RC.
Table 1.11 shows first the distribution of scores in these substitution tables, i.e. the number of gene-pairs found within different ranges of percentage of shared annotations, and secondly the distribution of scores of matches between mouse genes and motif-positions separated into the same ranges of scores.

Table 1.11: Distribution of scores when matching mouse genes to yeast genes and motif-positions of RC and SI motifs.

		% Identical annotations	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
165 genes	RC	Mouse vs. yeast	590	1170	250	568	43	12	0	0	0	0
		Mouse vs. motifs	8668	15502	3542	6534	605	258	0	0	0	0
	SI	Mouse vs. yeast	11434	11892	962	4578	164	239	858	224	0	0
		Mouse vs. motifs	11765	12325	1047	4821	179	243	870	225	0	0
1011 genes	RC	Mouse vs. yeast	3067	7771	1357	3272	216	87	11	2	2	0
		Mouse vs. motifs	40158	102822	20699	38336	2272	1235	182	35	15	0
	SI	Mouse vs. yeast	74088	84413	8271	27037	2533	2838	3475	241	691	0
		Mouse vs. motifs	76120	87544	8770	28344	2643	2906	3520	344	691	0

The table outlines the result of calculating functional identities between mouse and yeast genes from SI and RC (**Mouse vs. yeast**) and matching mouse genes to motif-positions from SI and RC (**Mouse vs. motifs**). Different intervals of scores are shown (percentage identical annotations) and the results for each interval present: **Mouse vs. yeast:** How many times pairs of mouse and yeast genes were found with the given score and **Mouse vs. Motifs:** how many times combinations of mouse genes and motif-positions were found with a score in the range of the given interval.

As to be expected, the results show that most occurrences of mouse genes matching a motif position from SI as well as RC are found with low percentages of shared annotations. With increasing functional identity, the occurrences of genes matching a motif position reduced. The distribution of matches between RC and SI showed only a noticeably large difference, when looking at functional identities above 50%. While there were no matching genes found above that percentage of shared annotations for the list of RC and the small dataset and hardly any matches for the list of RC and the large dataset, the list of SI still showed quite a few matches up to a functional identity of 80% for the small set and up to 90% for the large set.

Table 1.12 then shows different tested thresholds and the number of complete permutations that could be generated from the above shown results, if only matches are considered that show a percentage of shared annotations above each respective threshold level.

Table 1.12. Number of complete permutations for RC and SI motifs that could be generated when restricting to genes with scores above specific threshold levels.

	threshold (% identical annotations)		30	35	40	45	50	51	52	55	60	65	70
Possible Permutations	165 genes	RC	> 100000000	370	97	0	0	0	0	0	0	0	0
		SI	> 1000000	0	0	0	0	0	0	0	0	0	0
	1011 genes	RC	> 100000000	603070	30846	2647	466	42	8	1	0	0	0
		SI	> 100000000	47	43	36	36	32	32	32	30	23	0

Permutations for a given threshold can only be generated if a motif is found with at least one discrete mouse gene matched to each position with a score above the chosen threshold, i.e. to generate a permutation for an arbitrary motif at a threshold of 50%, there has to be a mouse gene for each position that has at least 50% of the annotations of the yeast gene within that position.

Although more instances could be found of a mouse gene matching any position in the list of SI motifs, these matches did not allow for the construction of a single motif above a threshold of 35 % shared annotations for the small dataset and for the large dataset they provided only very few permutations that however showed average sores of up to 65%.
In contrast, the list of regulator cascades got far fewer occurrences of genes matching a motif position with higher scores for both datasets. But they nevertheless showed even a few possible permutations for the small dataset when choosing a threshold of 40% and noticeably more possible permutations when choosing a threshold of 30%. When looking at the possible permutations for the large dataset, it becomes even clearer that the list of RC generally allowed for more permutations, for this set of genes even providing permutations up to a score of 55%.
All permutations that were found for the list of RC and generated on genes above a score of 35% consisted only of 3 positions. For the list of SI and using genes above 30% the permutations had only 2 positions.

Table 1.13. provides a graphical illustration of the results presented in tables 1.11 and 1.12 respectively comparing the findings for the small dataset against the results for the large dataset.

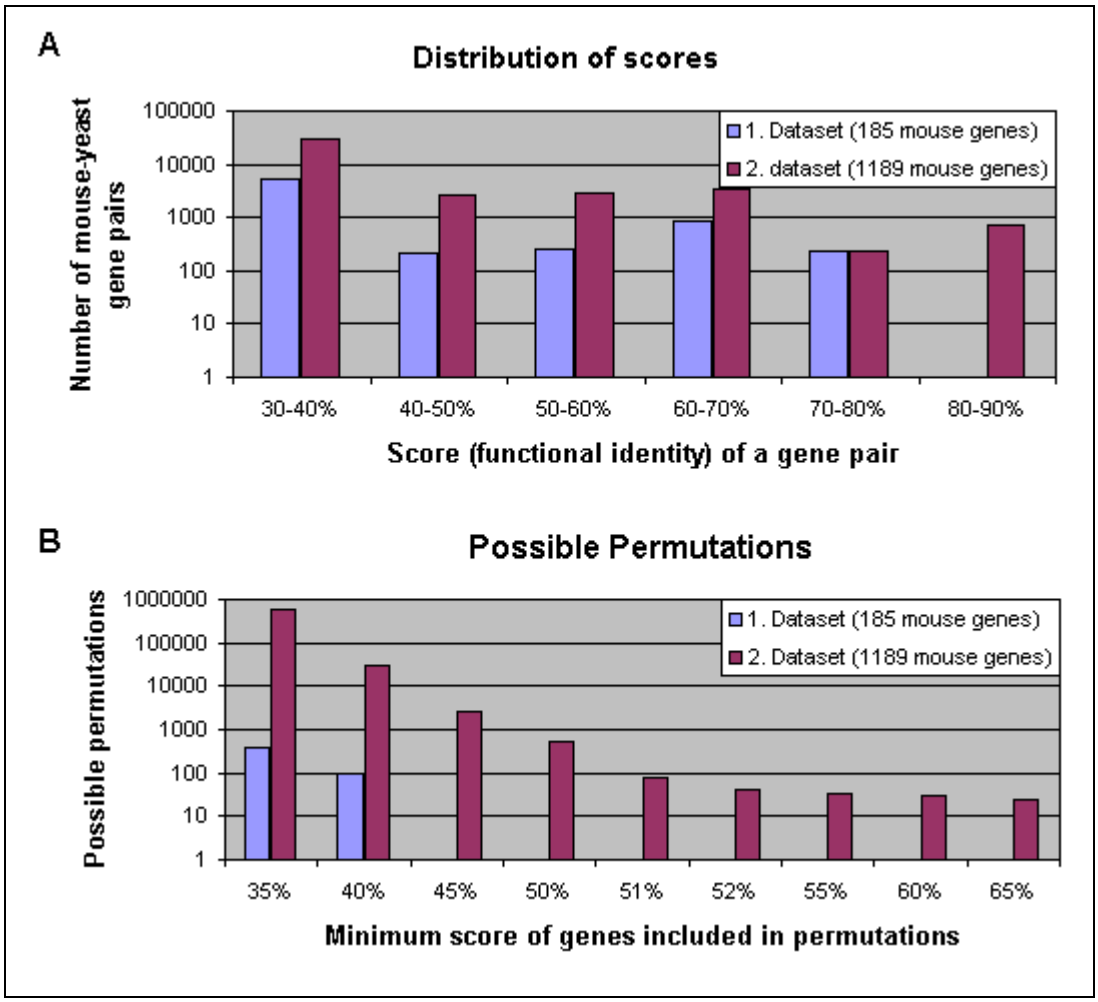


Figure 1.13. Prior results for the first data set (185 mouse genes) and second data set (1189 mouse genes) compared to each other. **A:** Distribution of scores (percentage shared GO annotations) when matching the mouse genes from both datasets against the yeast genes included in the list of SI and RC motifs. The graph shows how many pairs of mouse and yeast gene had, which scores. **B:** Possible permutations of mouse genes for any motif from the RC and SI list, when only considering mouse genes for a motif position, if they matched to it with a score above the displayed threshold levels.

Since the small dataset produced only very few and very low scored permutations, only the results for the large dataset, with 1011 mouse genes, were used.
For examination of the GO functional identity approach as scoring method to filter out real genetic regulatory relationships one set of most promising permutations had to be selected for each motif class. Basing on the above shown results it was chosen to pick the 42 permutations from RC, which were generated from genes with scores above 51% functional identity to the respective motif position, and the 23 motifs from SI built up from genes with a score above 65% functional identity.
The set of SI permutations was first evaluated by comparing it against two equally sized sets of SI permutations (23 permutations each) that were generated from genes with different ranges of scores (0-5% and 20-25%). All three sets were translated to network models and for the genes contained in each set of permutations an additional data-driven network was generated according to the method described above. After initialising and optimising the weight of each network, the errors of every model was evaluated with respect to consistency against the expression data and are presented in table 1.14.

Table 1.14. Small networks inferred from sets of permutations on SI that were scored according to functional identity.

Set nr.	Average score	Nr. genes included	Error of model	
			Permutations	Data-driven network
1	75.66%	24	32.56	69.85
2	22.22%	21	34.79	42.81
3	3.84%	21	30.95	47.03

Each set contains 23 permutations from matching the large data set (1011 mouse genes) against the list of SI. Set 1: 23 permutations built from genes with scores above 65%; Set 2: 23 permutations built from genes with scores between 20% and 25%; Set 3: 23 permutations built from genes with scores between 0 and 5%. The table presents the error of networks translated from these permutations and the error of data-driven networks inferred for the genes contained in the particular set of permutations.

The results shown in table 1.14 revealed two findings:

1. The networks that were generated from permutations showed in all cases a higher consistency to the correlation table (lower error) when compared to the respective data-driven network generated from the same set of genes. The difference between the permutation and data-driven network was however far higher for the first data set compared to the two other sets of permutations.
2. When comparing the permutations against each other by considering the error terms in relation to the number of included genes and the improvement over the data-driven networks, the consistency with the correlation table did not show a significant change between any of those three permutations.

The procedure, as applied on the SI permutations, was then repeated for the selected set of RC permutations. Two different sets consisting each of 42 permutations were chosen from the list of RC permutations for comparison. Error terms of network models created from these sets and errors of the respective data-driven networks are given in table 1.15:

Table 1.15. Small networks inferred from sets of permutations on RC that were scored according to functional identity.

Set nr.	Average score	Nr. genes included	Error of model	
			Permutations	Data-driven network
1	54.51%	21	29.53	49.63
2	20.78%	23	42.07	60.72
3	3.53%	21	30.45	46.79

Each set contains 42 permutations from matching the large data set (1011 mouse genes) against the list of RC. Set 1: 42 permutations built from genes with scores above 51%; Set 2: 42 permutations built from genes with scores between 20% and 25%; Set 3: 42 permutations built from genes with scores between 0 and 5%. The table presents the error of networks translated from these permutations and the error of data-driven networks inferred for the genes contained in the particular set of permutations.

Similarly to the results presented in table 1.14, networks from the three sets of RC permutations yield overall lower error values compared to each of the corresponding data-driven networks constructed from the respective set of genes. Relating to the number of genes and the improvement over data-driven networks the second and third set of permutations did not show a significant difference in their consistency with the correlation table. The first set of permutations however showed a quite low error value, but so did also the corresponding data-driven network.

4.2. Sequence similarity

The sequence similarity approach as described above was performed on 17 motifs with 4 positions each from the list of RC. From the 34 yeast genes contained in these motifs, protein sequences were downloaded for 33 and protein sequences could be found for 1182 from 1189 mouse genes.
By calculating the sequence similarity between each mouse protein and yeast protein, a substitution table was created and using these scores, mouse genes were matched to the positions of these 17 motifs. From these matches four different sets of permutations were created. A set with the 3 best scored permutations from each motif, a set with 3 medium-high scored permutations from each motif, a set with 3 medium-low scored permutations from each motif and the 3 lowest-scored permutations from each motif. These sets were translated to gene regulatory networks and data-driven networks were generated for their genes. The error values of these networks are given in table 1.16.

Table 1.16. Small networks inferred from sets of permutations on 17 RC motifs each with 4 positions that were scored according to sequence similarity.

Set nr.	Average score	Nr. genes included	Error of model	
			Permutations	Data-driven network
1	654.75	38	142.96	166.90
2	317.36	37	116.48	138.26
3	-214.43	41	142.22	174.43
4	-5673.54	33	93.79	118.84

Each set contains 51 permutations from matching the large data set (1011 mouse genes) against 17 motifs from the list of RC. Set 1: 3 best scored permutations for each of the 17 motifs; Set 2: 3 medium high scored permutations for each of the 17 motifs. Set 3: 3 medium low scored permutations for each of the 17 motifs. Set 4: 3 lowest scored permutations for each of the 17 motifs. The table presents the error of networks translated from these sets of permutations and the error of data-driven networks inferred for the genes contained in the particular set of permutations.

Similarly to the results presented in table 1.14 and table 1.15, all networks presented in table 1.16 showed, independently of the average score of its permutations, a constant improvement over the corresponding data-driven networks. But although every network that was translated from a differently scored set of permutations performed better than the corresponding data-driven network, the performance between these permutations itself did not show any differences or gradation.

4.3. Inferring complete networks

Finally three networks were inferred for the whole dataset of 1011 mouse genes. The first network was generated following the same principles with which the data-driven networks in the prior steps were constructed. The second and third networks were generated by incorporating causations from permutations, as described above. The permutations for the second model are built on SI and RC and scored according to shared GO annotations. The permutations for the third model were constructed for the 17 RC motifs with four positions and scored by means of sequence similarity.
Weights for all three networks were again initialized randomly and optimized and the models afterwards evaluated against the correlation table. The results are presented in table 1.17.

Table 1.17. Complete networks inferred for the large data set (1011 mouse genes)

Network	Source of causations	Error of model
1	Data-based - Correlation table	154166.27
2	Permutations - Functional identity	153972.38
3	Permutations - Sequence similarity	154054.83

The table shows the accuracy of three different gene regulatory networks, which were inferred for the complete data set (1011 mouse genes). Network 1: Data-based network inferred from the correlation table alone. Network 2: Model inferred from permutations on the list of SI and RC that were scored by the used functional identity method. Network 3: Model inferred from permutations on 17 RC motifs with each 4 position and scores determined by the presented sequence similarity measure.

The results given in table 1.17 simply indicate that all the three networks generated for the set of 1011 genes were equally erroneous and extracting causations neither from permutations scored by GO identity nor from permutations scored by sequence similarity allowed to infer a network that performed better than the purely data-based network.

5. Conclusions

5.1. Summary of conclusions

The small number and low average score of permutations found by the GO functional identity approach first indicated that this similarity measure was not really able to detect pairs of yeast and mouse genes that were as similar as desirable to confidently expect them to be involved in analogous regulatory interactions and as numerous as needed to allow for the inference of a gene regulatory network. The small size of these permutations further accounted for the possibility that they represent matches, which occurred by random chance and mostly without displaying any regulatory relationships. The results from the following experiments helped to fortify these first assumptions. Small networks as translation of permutations did show a general improvement over data-driven networks, but the fact that this improvement was independent from the score of the permutations and that the best scored permutations performed equally well compared to very low scored permutations, gives reason to the assumption that high scored permutations do not provide any meaningful regulatory interactions. Considering then the results of the final test, the inference of a gene regulatory network from these permutations, which was as erroneous as the data-driven network, it can be confirmed that these permutations do not or only insufficiently provide useful regulatory causations. With the sequence similarity approach also producing high scored permutations that do not perform better than the lowest scored permutations and an equally erroneous network for the whole dataset, it follows that the hypothesis and the presented network inference approach failed, i.e. the given similarity

measures were incapable to identify noticeable groups of genes with regulatory relationships analogous to the given network motifs.

5.2. Detailed conclusion

When implementing a network inference method that proposes topological structures of genetic interaction by comparison to regulatory motifs of other species, the following major prerequisites can be specified, which need to be satisfied in order to achieve success:

1. The set of genes that the network has to be generated for must contain a sufficient number of regulatory patterns consisting of genes that display similar properties to the genes of a reference motif, i.e. to find patterns analogous to a yeast motif, such patterns need to be present in the dataset and in order to reconstruct a network for the whole dataset there need to be enough patterns to provide an appropriate framework for the topology of the network.
2. The similarity measure must then be capable of detecting such pairs of genes, which are involved in analogous regulatory interactions, i.e. pairs of genes proposed to exhibit similar regulatory behaviors need to be proposed because they show similar regulatory behaviors, or in other words: a similarity score has to indicate not only similarity in the asked characteristics of the genes but also similarity in their regulatory interactions.
3. Once the similarity between the genes of the dataset and the genes of the reference motifs is known, this knowledge has to be examined properly in order to reconstruct regulatory patterns and also a complete network.

Confirming the absolute presence or absence of specific regulatory patterns in the dataset might rather not be directly possible, meaning that statements about the amount of promising data in the dataset generally depend on the nature of the used similarity measure to detect the genes. Accordingly the results presented in table 1.11, table 1.12 and table 1.13, provide the first insight into the nature of the dataset on the basis of the used GO identity measure. The two most obvious findings are first the fact that matching the dataset to the list of SI motifs yielded generally more highly scored instances of any mouse gene matching a position in one of these motifs, i.e. more pairs of mouse and yeast genes were found that appeared to be similar to each other. In contrast the list of RC motifs however allowed generally for more permutations. This circumstance can be attributed to the fact that the list of RC motifs consists of 72 different yeast genes, while the SI motifs contain 1597 different yeast genes and provide therefore far more possibilities to find a gene similar to any mouse gene. But the list of SI motifs also contains motifs consisting on average of far more positions (up to 213) than the RC motifs, which range from 2 to 10 positions. Thus SI motifs need generally more high scored genes in order to find a complete permutation of high scored genes, while complete sets for the smaller RC motifs only need a few high scored genes in order to build a high scored permutation.
The small data set (185 mouse genes) did show a few genes similar (50%-80% functional identity) to genes from the motifs, but the fact that these matches did not allow for any overall high scored permutation gives reason to the assumption that the data set does not cover enough genes to find potential meaningful regulatory patterns analogous to any motif. The dataset was therefore increased to 1189 genes, increasing also the number of similar mouse-yeast gene pairs, and allowing for permutations with sets of genes more similar to the respective motifs.
Considering however the number and average scores of these permutations, it still can be said that:

1. Even the highest permutations detected might consist of genes not similar enough (55% functional identity for permutations on RC and 65% functional identity for permutations on SI) to their matched yeast genes to exhibit the same regulatory role, leaving the permutation without real regulatory relationships.
2. And even if these highest scored permutations describe real regulatory relationships, there are probably too few permutations to allow for a network construction of the whole data set.

Additionally, the higher scored permutations that are listed in table 1.12 only consisted of the minimum number of positions (2 or 3 for any permutation on RC motifs and 2 for any permutation on SI motifs).
Matches for a motif that consists of only two or three positions can generally be considered less significant findings than matches for bigger motifs, since combinations of genes with high scores can occur as a match for a small motif more easily by random chance, while bigger motifs require a larger number of potential genes before they allow for a high scoring permutation.
The observation that only minimum sized permutations are found, related to their scores and number, leads to the first assessment that the dataset still might be too small to contain analogous patterns or those that are not detected by the used GO functional identity approach.
The next step then was to evaluate the findings in order to estimate their meaningfulness and the performance of the similarity measures used to produce data, i.e. an assessment if pairs of genes proposed highly similar by GO functional identity or sequence similarity are likely to also display a regulation similar to the template pair of yeast genes.
Comparing table 1.14, 1.15 and 1.16, it can be found that all the three results show the same pattern: independently of the used measure to calculate a similarity between mouse and yeast genes, creating networks from permutations always reduced the error of the network compared to the corresponding data-driven network. The difference between the data-driven network and the permutations appears to be constant and unrelated to the average score of the permutations.
Apart from this general improvement it can however not be found a trend in the data that would allow for the conclusion that higher scored permutations show a significantly higher consistency with the correlation data when compared to lower scored permutations.
Since very low scored permutations do on the contrary perform about as equally good as the high scored ones, it can be reasoned that permutations are not sorted according to real regulatory relationships, i.e. the high scored permutations do not hold significantly more real regulatory information than low scored permutations.
Considering however the nature of the results and the overall good performance of permutations over data-driven networks it becomes difficult to interpret the data and draw clear conclusions from them.
The only indication for a possible finding of some useful / real regulatory relationships is therefore given by the fact that the best scored permutations for the list of SI, compare table 1.14, showed a far higher difference to the corresponding data-driven network than the 2 sets of lower scored permutations. The results are however rather unclear and, as already said, the other results do not fortify this finding.
The results presented in table 1.17 in contrast are clearly stating that the extraction of single causations neither from permutations scored by sequence similarity nor from permutations scored by percentage of identical GO annotations allowed to significantly reduce the error of the network that was inferred for 1011 of the 1189 mouse genes in the whole dataset.
Combining these findings (the failure to produce a complete network from these permutations that is more consistent with the expression data than the corresponding data-

driven networks as well as the fact that sets of high-scored permutations did not seem to perform better than sets of low-scored permutations when translated to gene regulatory networks) fortifies the prior assumption, that either the dataset might be too small to exhibit mouse genes with identical functions and interactions like the yeast genes in the given motifs, or such genes are not detected by directly comparing functional annotations between yeast and mouse genes.
Either way this leads to the conclusion that the highly scored permutations, which were selected and compared to lower scored motifs by means of consistency to expression data, do not perform better than lower scored permutations probably because of the fact that they simply do not contain enough, if at all, real regulatory relationships. Or in other words the used similarity measures in combination with the generation of scored permutations did not manage to propose genetic regulatory relationships, which would have allowed for a gene regulatory network inference approach alternate to and solving the computational complexity given by purely data-driven inference methods.
From the given results it is however not directly distinguishable if better scored permutations lack such close to real regulatory relationships, because these were as such not present in the data set or could not be detected by a measure as used when checking for identical functional annotations or sequence similarity between genes.
Permutations scored by sequence similarity and permutations scored by functional identity gave however quite similar results for comparison of different scored permutations as well as for the inference of a complete network. This implies that both methods would have had missed the patterns of genes with similar properties as well as similar regulation behavior if those are contained in the data set. With two methods not pointing out such patterns within the data set, it would generally be reasonable to account this to the fact that the required genes, which are building up these patterns, are in fact not or only sparsely present in the data set.

5.3. Problems

The approach presented in this work does however show some severe disadvantages and different problems were found.

Evolutionary distance between mouse and yeast: The first thing to consider is the evolutionary distance and biological difference between the two used organisms. The evolutionary distance gives reason to the assumption that many gene regulations are already far too diverged to share enough detectable similar characteristics. Further on, the cells of mouse as a multicellular organism show different specializations compared to the single cell organism yeast. This means that mouse in general contains far more genes than yeast, and implies further that the cell selected for the microarray experiment might well contain many specialized regulatory interactions (e.g. when looking at muscle-cells) that cannot be found in yeast.

Accordingly the two used methods (checking for identical GO functional annotations and calculating sequence similarity) might be insufficient when applied to score a pair of genes from two different species like yeast and mouse according to their similarity in terms of biological role.

GO functional identity: Considering the evolutionary distance between mouse und yeast it is quite assumable to expect a pair constisting of a yeast and a mouse gene, both still involved in analogous biological processes and regulatory interactions, being diverged

during evolution, i.e. the mouse gene and yeast gene show slightly different functions although both are still involved in analogous regulatory interactions.
Scoring pairs of genes by identifying identical functional annotations between them as done in the project might thus be too specific, i.e. disregarding closely related functions although they might be informative indicators for similar biological roles.

Sequence Similarity: Sequence similarity on the other hand is a more vague method, since proteins with a high sequence similarity are not necessarily related in function as well. Accordingly scores calculated by this approach only give approximations for similar functions between genes and as mentioned above this approach does not produce absolute values, meaning it would normally require further analysis to draw conclusions from the calculated similarity. But further analyzing each created alignment is a very time-consuming process especially when generating alignments for thousands of proteins, which would make the measure almost unfeasible for this task.
So due to the nature of those two methods the possibility has to be admitted that patterns analogous to any motif are present in the dataset more frequently but are not or not in sufficient number catched by either method.

No analogous motifs in the dataset: If on the other hand the similarity measures are regarded sufficient or at least indicating presence or absence of promising data in the data set, another problem arises when considering the amount of data found in relation to the size of the data set.
For the first data set, consisting of 185 genes, the functional identities for the mouse genes matched to the SI and RC motifs indicated that this set contained a few genes with some degree of shared functions compared to some yeast genes in one of the selected motifs. But considering the few permutations that could be generated from these matches, the data set already appeared quite insufficient and lacking potential analogous regulatory patterns. Evaluation of the further results revealed that enlarging the data set to 1189 mouse genes did indeed increase the number of genes with similar functions, but did still not allow to produce a proper number of permutations, which are built up of a number of overall highly scored genes and also contain meaningful regulatory relationships.
The problem arising with this finding is the following: Under the supposition that analogues for motifs from the other three motif classes are likely to be rare in the dataset, a further increase of the data set might help to detect sets of genes which show very high functional identities to some motifs and thus also meaningful regulatory information, but from the known findings it can be assumed that such patterns would then only cover a fractional amount of the genes contained in the whole data set. Hence increasing the dataset in order to find such patterns will result in significantly increased computational times for similarity measures as well, but will on the other hand most likely only provide a small number of genetic interactions. So these interactions found for very large sets of genes are then insufficient to directly infer the network underlying the given dataset or guiding a data-based approach to reconstruct the network, since such a large dataset stands in direct conflict with the computational complexity of purely data-based network inference.

Problems with permutations: Further complications result from the use of permutations. Permutations as complete patterns might on the one hand, when highly similar to the genes of a motif, represent more significant findings than partly matched motifs, which could occur more easily by random chance. But focusing only on complete permutations might just as well disregard useful information, e.g. when regulatory patterns analogous to a yeast motif are only partly present in the data set.
But the major disadvantage with permutations is the fact that their generation is a rather time-consuming procedure as explained above. When only generating a few permutations as done in this work to evaluate the scoring methods, or focusing on a few smaller motifs, permutations are useful because they provide a direct representation of potential regulatory patterns among mouse genes, which can be compared to each other.
When however dealing with larger motifs or the need to create a huge number of permutations in order to exhaust as much data and information as possible, as done when generating networks for the whole dataset, they turn out to be too time-consuming.

Consistency of permutations with correlation table: The evaluation of these permutations turned out to be likewise problematic. The method used to estimate the performance of networks, i.e. comparison against data-driven networks, might be seen as appropriate when dealing with complete networks inferred for the whole dataset. But as obvious from the results, evaluating smaller sets of only a few permutations with respect to meaningfulness or presence of regulatory relationships can produce rather unclear data when done by estimating their consistencies with expression data or by comparison against the performance of data-driven networks.

Data-Driven networks: Permutations are measured against the performance of data-driven networks, which are generated according to the method described above. The problem however is the fact, that there is no indication of how well the method works to propose relationships for the data-driven networks. The comparison of permutations and data-driven networks thus lacks an absolut explanatory power for the performance of any network. A random generated network would be needed in order to get a reference for first estimating, how well the correlation-based network inference and the corresponding data-driven networks are.
The fact that all permutations performed better than the corresponding data-driven networks might further also be caused be the use of correlation data from knockout experiments, which might produce erroneous results, if a deleted gene is selected for the construction of a data-driven network.

5.4. Implications and improving the method

All together it can be said that the presented approach failed to produce data that could be interpreted as close approximations to real regulatory interactions between the given mouse genes or would have allowed to reconstruct a network less erroneous than the data-driven one, hence the hypothesis can also be concluded as failed. According to theabove given reasons it remains rather unclear if the failure has to be attributed to the lack of analogous motifs in the dataset or the use of the presented similarity measure. Considering however the problems mentioned above, different modifications can be proposed that would help to improve a procedure similar to the one given in this work.

Using more closely related species: With the availability of motifs from yeast, as one of the most studied organisms with respect to gene regulatory networks, it stood to reason to use these for first examinations of the presented approach. In addition, if the use of two

species as far diverged as mouse and yeast would have allowed for a proper network inference, it would have meant that the approach is likely to work even better for more closely related species. Since the results were however rather unclear and did not allow for the reconstruction of well performing networks further work should focus on more closely related species and also consider differences between single cell and multi cell organisms as mentioned above.

Semanitic Similarity: As already mentioned in the methods section, it was first planned to implement a measure for semantic similarity between GO annotations instead of the sequence similarity and GO functional identity, but the implementation turned out to be too time-consuming to be used and it was switched to the used similarity methods. A semantic similarity method would however be more appropriate for this task, since it is, by means of comparing semantic and/or structural characteristics of annotations, capable of detecting pairs of genes that have related annotations and thus related functions. And due to the evolutionary distance between mouse and yeast homologues might rather exhibit similar functions than being totally conserved and still possess identical functions.
If a semantic similarity measure does not provide useful data then it can be concluded more confidently that the data set does not contain regulatory patterns analogous to the given yeast motifs.
And in order to let these results then become more meaningful instead of just representative approximations for the presence of analogous motifs, it appears also reasonable not to restrict to only a few motif classes / instances but to incorporate maybe even all motifs found for the yeast regulatory map.

When not using a measure of semantic similarity and instead utilizing sequence similarity it might, with respect to comptutational time and comparability of results, appear advisable to use BLAST (Altschul et al. 1997), since it is faster than e.g. the Needleman-Wunsch algorithm (Needleman and Wunsch 1970) and also provides E-values as a measure of relative similarity, which allows for more confident conlusions about signicifant findings compared to the results obtained from the Needleman-Wunsch implementation.

Using on the other hand GO annotations alone might cause problems, if the selected genes are not provided with all important annotations, e.g. some genes are only registered with CC annotations or MF annotations. If genes are not annotated completely it becomes hard to compare those genes (although they might have identical functions) and even harder to draw any useful conlusions from such a test. It therefore also appears worth-while to incorporate additional information, e.g. other transcription factor databases.

Evaluation of scoring and regulatory relationships: Testing regulatory patterns, which are proposed from genes found to be promising by a semantic similarity measure, against the correlation table might produce more clear findings, if the scoring performed better than the scoring methods presented in this thesis. But evaluating small patterns or permutations by their consistency with the expression data and against other patterns or data-driven networks might just as well yield as imprecise and unclear results as seen in this work.
An alternative approach to assess if genes are sorted according to real gene regulation would be to compare the best scored permutations against known genetic regulatory patterns from the mouse genome, and determine to what degree the calculated relationships are true.

Inference of gene regulatory networks however mostly deals with or aims at the exploration and reconstruction of yet unknown regulatory relationships from expression data, which means that there is not always a known network to confirm the results.
But more importantly, even if such known maps are available for comparison, in order to find significant matches, meaning permutations that show more than only one or two correct causations, which could occur just by random chance, it might be necessary to include many if not all genes from the mouse genome when generating permutations. When focusing on a few specific motifs as done for the sequence similarity approach and a small set of genes to find those, it might rather be the case that analogous patterns of regulation are if at all only partly present, as already indicated by the findings stated above.

Proposing regulatory relationships / network topology: Scored permutations provide a very clear representation of potential regulatory relationships, i.e. they directly illustrate possible analogous motifs and such topologies can also be directly assigned an average similarity to the reference motif. So permutations lend themselves for displaying small numbers of potential analogue motifs, which then can be evaluated.
For the purpose of proposing gene-gene interactions in order to reconstruct a network for whole datasets, the number of possible and potentially needed permutations especially for large sized motifs would however result in computational times too large for the approach to be suitable for this task.
But absolute scores of similarity as produced by semantic similarity of GO functional identity measures do also directly point out, which motif and position show the best match for a mouse gene and how probable it is to find the gene taking another motif position.
So proposing gene regulatory relationships for the dataset can also be done in different ways alternative to generating permutations:

1. Complete patterns: When choosing an arbitrary mouse gene, the motif is selected, which shows the position with the highest similarity to the gene. The gene is considered to be taking this position in an analogous motif and the pattern is then completed by picking a gene for each other position to give the pattern the best score. Repeating this procedure over the dataset, best pattern is created for each gene. A complete network can then be created by selecting between or combining these pattern.
2. Substitution tables for single gene-gene interactions: E.g. when working with threshold levels of similarity in order to allow a mouse gene to be a match for a motif position, it can be counted how often a mouse gene A is found in any motif position where it is regulating gene B. Counting these numbers for every pair of mouse genes, a table is created that lists how often one gene was found to be assumed to be regulating another gene. The higher these numbers, the more likely it is to find the corresponding genes to be truly performing the corrsponding regulation. When not working with threshold levels, a configuration could simply be assumed a configuration in a motif in which gene A is in a position of regulating gene B. The scores of each gene beeing matched to its position is multiplied, indicating how probable it would be to find these two genes interacting with each other in exactly the chosen configuration. When combining these scores for every possible configuration in all motifs, another substitution table can be created, indicating how probable it would be to find gene A regulating gene B in general.
   Networks can be constructed from these substitution tables by e.g. picking the most probable genetic interactions for each gene and combining them to an overall topology of interactions for the whole dataset.

6. References:

Akutsu, T., Miyano, S., and Kuhara, S. (2000a) Inferring qualitative relations in genetic networks and metabolicpat pathways, *Bioinformatics*, 16: pp. 727–734

Akutsu, T., Miyano, S., and Kuhara, S. (2000b) Algorithm for identifying Boolean network and related biological networks based on matrix multiplication and fingerprint function, *Journal of Computational Biology*, 7: pp. 331–343

Altschul, S. F., Madden, T.L., Schäffer, A.A., Zhang, J. Zhang, Z., Miller, W., Lipman, D.J.. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*., 25: pp. 3389-3402.

Baitaluk, M., Qian, X., Godbole, S., Raval, A., Ray, A., Gupta, A. 2006. PathSys: integrating molecular interaction graphs for systems biology. BMC Bioinformatics 2006, 7: p. 55

Banerjee, N. and Zhang, M.Q. (2002) Functional genomics as applied to mapping transcription regulatory networks. *Curr. Opin. Microbiol*., 5: pp. 313–317.

Black, P.E. (2004) "simulated annealing", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology. 17 December 2004. Available from: <http://www.nist.gov/dads/HTML/simulatedAnnealing.html> [10 March 2007]

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) A model of evolutionary change in proteins. In Dayhoff, M. O., ed., *Atlas of Protein Sequence and Structure*, volume 5, supplement 3. National Biomedical Research Foundation, Washington D.C. pp. 345-352

de Hoon, M., Vitkup, D. 2005. Comparative systems biology of the sporulation initiation networks in prokaryotes. *Conference on Research in Computational Molecular Biology* (*RECOMB 2005*): pp. 62-70.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* 278: pp. 680–686.

Dojer, N., Gambin, A., Mizera, A., Wilczyński, B., Tiuryn, J. (2006) Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics* 2006, 7: p. 249

Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998) *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press. pp. 42-43.

Eberl, W. (1995) Simulated Annealing [online] available from <http://www.eberl.net/chaos/Skript/node50.html> [10 March 2007]

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.,* 95: pp. 14863-14868.

Ensembl (2007) Ensembl Genome Browser [online] available from <http://www.ensembl.org/> [10 March 2007]

Garson, G.D. (2006) Statistics Solutions: Path Analysis [online] available from <http://www.statisticssolutions.com/Path_Analysis.html> [30.04.2007]

Hakamada, K., Hanai, T., Honda, H., Kobayashi, T. (2001) Identifying Genetic Network Using Experimental Time Series Data by Boolean Algorithm, *Genome Informatics*, 12: pp. 272-273.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.,* 7: pp. 422–433.

Herrgard, M.J., Covert, M.W., and Palsson, B.O. (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res*., 13: pp. 2423-2434.

Hong, E.L., Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Livstone MS, Nash R, Oughtred R, Park J, Skrzypek M, Starr B, Andrada R, Binkley G, Dong Q, Hitz BC, Miyasato S, Schroeder M, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, and Cherry JM. (2007) Saccharomyces Genome Database [online] available from <http://www.yeastgenome.org/> [10 March 2007]

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102: pp. 109–126.

Iba, H. and Ando, S. (2001) Inference of Gene Regulatory Model by Genetic Algorithms. *Proc. of 3rd International Symposium on Adaptive Systems*, pp. 15-22.

Iba, H. and Mimura, A. (2002) "Inference of a Gene Regulatory Network by means of Interactive Evolutionary Computing" *Information Sciences*, vol.145, no.3-4, pp.225-236.

Ideker, T.E., Thorsson, V., and Karp, R.M. (2000) Discovery of regulatory interactions through perturbation: Inference and experimental design. *Pac. Symp. Biocomput.* 292: pp. 305–316.

IIT Research Institute, Reznik G. (2003) ITRI Newsletter Fall 2003 [online] available from <http://www.iitri.org/Newsletters/NL-Fall-03-1.shtml> [30.04.2007]

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409: 533–538.

Kustra, R., Shioda, R., Zhu, M. (2006) A factor analysis model for functional genomics. *BMC Bioinformatics 2006*, 7: 216

Kyoda, K. M., Morohashi, M., Onami, S. and Kitano, H. (2000) A gene network inference method from continuous-value gene expression data of wild-type and mutants. *Genome Inform. Ser. Workshop Genome Inform*. 11: 196-204.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002a) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799–804.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002b) Transcriptional Regulatory Network [online] available from <http://web.wi.mit.edu/young/regulator_network/> [30.04.2007]

Lessem, J. (2002) Tracing Rules for Standardized Variables [online] available from <*http://ibgwww.colorado.edu/twins2002/cdrom/HTML/BOOK/node78.htm*>[30.04.2007]

Li, X., Rao, S., Jiang, W., Li, C., Xiao, Y., Guo, Z., Zhang, Q., Wang, L., Du, L., Li, J., Li, L., Zhang, T., Wang. Q. K. (2006) Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics. 2006*, pp. 7: 26.

Laurio, K., Svensson, T., Jirstrand, M., Nilsson, P., Gamalielsson, J., Olsson, B. (2007) Evolutionary Search for Improved Path Diagrams. E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 114–121.

Maki, Y. *et al.* (2001) Development of a system for the inference of large scale genetic networks. *Proc. Pacific Symp. on Biocomputing '01*, *World Scientific*, 6: pp. 446–448.

Markowetz, F. (2005) Probabilistic Graphical Models for Cellular Pathways, *IPM workshop 2005*.

Mazurie, A., Bottani, S., Vergassola, M. (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biology*, 6: R35.

Mimura, A., Iba, H. (2001) Interactive Evolution of a Gene Regulatory Network Model. *Genome Informatics,* 12: 278-279.

Mudelsee, M. (2003) Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series. *Mathematical Geology,* 35: 651–665.

Needleman, S.B. and Wunsch, C.D. J. (1970) A general method applicable to the search of similarities in the amino acid sequences of two protein. *Mol. Biol*. 48: pp. 443-453

Noman, N. and Iba. H. (2005) Inference of gene regulatory networks using s-system and differential evolution. *Genetic and Evolutionary Computation Conference* (*GECCO 2005*): pp. 439-446.

Oliveira, A.L., Freitas, A.T., Sá-Correia, I. (2007) Bioinformatics: a new approach for the challenges of molecular biology. *A Portrait of State-of-the-Art Research at the Technical University*, (M.S. Pereira, ed.), Springer, Dordrecht, pp. 295-309

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics 2002*, Vol. 18 no. 4, pp. 546-554.

Pe'er, D., Regev, A., Elidan, G., and Friedman, N. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17: 215–224.

Radivoyevitch, T. (2005) Folate system correlations in DNA microarray data. *BMC Cancer*, 5: 95.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* 290: 2306–2309.

Rougemont, J., Hingamp, P. (2003) DNA microarray data and contextual analysis of correlation graphs. *BMC Bioinformatics*, 4: 15.

Sakamoto, E. and Iba, H. (2001) Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Proceedings of the 2001 Congress on Evolutionary Computation CEC2001*, pp. 720−726.

Savageau, M. A. (1991) 20 years of s-systems. *Canonical Nonlinear Modeling. S-systems Approach to Understand Complexity*, pp. 1−44.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 1995*, 270: 467–470.

Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics, 7: p. 302.*

Schwarzbach, C. and Börner, R.U. (2001) Genetische Algorithmen und Simulated Annealing: Nichtlineare Optimierung am Beispiel der Widerstandsgeoelektrik. *Protokoll über das 19. Kolloquium „Elektromagnetische Tiefenforschung", pp. 168-174.*

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., Cherry, J.M. (2001) The Stanford Microarray Database. *Nucleic Acids Research*, Vol. 29, no. 1, pp. 152-155.

Silvescu, A., Honavar V. (1997) Temporal boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems,* Vol. 13, no. 1, p. 54.

Spieth, C., Streichert, F., Speer, N. and Zell, A. (2004) A memetic inference method for gene regulatory networks based on s-systems. In: *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2004)*, pp. 152–157.

Stormo, G.D. and Tan, K. (2002) Mining genome databases to identify and understand new gene regulatory systems. *Curr. Opin. Microbiol.* 5: pp. 149–153.

Tegner, J., Yeung, M.K., Hasty, J., and Collins, J.J. (2003) Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci.,* 100: pp. 5944–5949.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat Genet 2000, 25:* pp. *25*–29.

Trey, I.,Vesteinn, T., Jeffrey, A.R., Rowan, C., Jeremy, B., Jimmy, K.E., Roger, C., David, R.G., Ruedi, A., and Leroy, H. (2001) Integrated genomicand protein analyses of a systematically perturbed metabolicnet work, *Science*, 292: 929–934.

University of Chicago Medical Center. (2005) University of Chicago study overturns conventional theory in evolution [online] available from <http://www.uchospitals.edu/news/2005/20050607-kaks.html> [30.04.2007]

van Someren EP, Wessels LF, Backer E, Reinders MJ. (2002) Genetic network modeling. *Pharmacogenomics*. 3(4): pp. 507-25.

Wang, T., Touchman, JW., Xue, G.L. (2004) Applying two-level simulated annealing on Bayesian structure learning to infer genetic networks. *Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB'04)*; August 16-19; Stanford University. pp. 647-648.

Webley, P., Lea, S. (1997) Principles of Path Analysis [online] available from <http://www.people.ex.ac.uk/SEGLea/multvar2/pathanal.html> [30.04.2007]

Wong, T. and Wong H. (1996) Genetic Algorithms [online] available from <http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html> [30.04.2007]

Wosik, E. (2004) Boolean Networks [online] available from <http://cnx.org/content/m12394/latest/> [10 March 2007]

Wright, S. (1934) The method of path coefficients. *Annals of Mathematical Statistics*, 5: 161-215.

Wyrick, J.J. and Young, R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.,* 12: 130–136.

Yeung, M.K., Tegner, J., and Collins, J.J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci.,* 99: 6163–6168.

Yona, G., Dirks, W., Rahman, S., Lin, D.M. (2006) Effective similarity measures for expression profiles. *Bioinformatics* 22(13): 1616-1622.