UNIVERSITY OF TARTU
Faculty of Biology and Geography
Institute of Molecular and Cell Biology

Hedi Peterson

The discovery and characterization of regulatory motifs in *Saccharomyces cerevisiae*

Master's Thesis

Supervisor: Jaak Vilo, PhD

Tartu 2006

"Computers are useless. They can only give you answers."

Pablo Picasso

Contents

List of Figures

List of Tables

Abbreviations

AD	Activation Domain
bp	base pair
CC	Cellular Component
ChIP-on-chip	Chromatin ImmunoPrecipitation on chip
DBD	DNA Binding Domain
DNA	DeoxyriboNucelic Acid
EP	Expression Profiler
GO	Gene Ontology
MF	Molecular Function
ORF	Open Reading Frame
PPI	Protein-protein interaction
RNA	RiboNucleic Acid
SGD	Saccharomyces Genome Database
TAP	Tandem Affinity Purification
TF	Transcription Factor
TFBS	Transcription Factor Binding Site

Chapter 1

Introduction

1.1 Background

The full genome of *Saccharomyces cerevisiae* was sequenced already ten years ago and the studies of genes and regulatory regions of baker's yeast has been ongoing for even longer (Goffeau *et al.* 1996; Liti & Louis 2005). The current number of *Saccharomyces cerevisiae* open reading frames (ORFs) in Saccharomyces Genome Database (SGD) is 6604, however only approximately 70% of them have gene names and a similar proportion have at least one Gene Ontology (GO) annotation attached to it (Hong *et al.* ; Ashburner *et al.* 2000; 2001; Harris *et al.* 2004). It means that we know very little about the remaining 30% of the ORFs. Of course some of the ORFs can be false positives and not actually code proteins, but there is still a lot to study about these nearly 2000 ORFs.

Even if a gene has a GO annotation, it does not mean that the function or biological process is known. Almost every fourth ORF described with some GO has the annotation '*molecular function unknown*', and approximately 5% are connected with the GO annotations '*biological process unknown*' or

'*cellular component unknown*'. These facts increase the number of ORFs that should and could be described up to almost 3000.

The most reliable sources of gene function come from *in vivo* and *in vitro* experiments. The bioinformatics approaches of analyzing data from chromatin-immunoprecipitation on chip (ChIP-on-chip) or gene expression, and the predictions of putative transcription factor binding sites (TFBS) have contributed to the overall knowledge as well.

The function predictions often begin with the comparison of similarly regulated genes. The regulatory regions have been studied extensively in *Saccharomyces cerevisiae* (Stormo 2000; Vilo *et al.* 2000; Tompa 2001; van Helden, André, & Collado-Vides 1998). The sequencing of several additional yeasts have made available the comparative genomics approach to study *Saccharomyces cerevisiae* more closer. The phylogenetic footprinting approach helps to find regulatory motifs that are conserved in different species and therefore increase the trustworthiness of the predicted motifs (Duret & Bucher 1997).

Previous studies have shown that using additional yeasts, it is possible to discover stronger regulatory motifs in *Saccharomyces cerevisiae* (Kellis *et al.* 2003; Dujon *et al.* 2004). A previous study also showed that genes interacting with each other, especially if they have co-expression, are often functionally related (Kemmeren *et al.* 2002).

The abundance of various data sources, bioinformatics methods and the growing computational power have made a good basis for starting large-scale analysis. With the ever-growing data relationships it is a huge challenge to be able to tie together data from Gene Ontology, protein-protein interactions and known transcription factor target genes. Connecting this multidimensional dataset will help to create new knowledge.

1.2 Objective

The aim of the thesis was to develop methods for large-scale pattern discovery and characterization of regulatory motifs in *Saccharomyces cerevisiae*. Usually the regulatory signals are studied for one specific dataset, for example similarly expressed genes or ChIP-on-chip data. In this project we aimed to detect and describe putative transcription factor binding sites from different large-scale datasets.

Three big datasets i.e. known targets of transcription factors, Gene Ontology annotations and protein-protein interactions were combined. Different combinations of the data gives several possibilities to predict new knowledge about functional annotations, possible regulatory motifs or target genes.

Chapter 2

Background

2.1 Biological background

The main emphasis of the current molecular biology is to understand the complex mechanisms acting inside and between the cells. The protein functions depend on the structure, the solution they are in, and the interactions they make with other proteins or macromolecules. Studies for understanding the possible function of a protein starts from comparing the sequence similarity to other previously known proteins, comparisons of homologous proteins from other species, pattern discovery and matching in regulatory regions, just to mention few starting points. To describe a protein from the scratch with *in vivo* or *in vitro* methods is time and money consuming. With the help of bioinformatics the possible functions can be predicted and therefore the search range can be reduced.

Lots of work has been done with more primitive organisms like bacteria, lower eukaryotes and plants. Huge amount of knowledge has been collected to large variety of databases describing gene functions, protein structures, regulatory sequences etc. One of the most widely used data source of biological

data is Gene Ontology. GO consortium builds and maintains a vocabulary of different kind of descriptors to relate genes with their functions or cellular localizations in regard of the actual species (Ashburner *et al.* 2000; 2001; Harris *et al.* 2004).

2.1.1 Gene Ontology

The Gene Ontology consists of three ontologies, that are described as a tree structured vocabulary (Ashburner *et al.* 2000). The three ontologies describe gene products with their associated biological processes (BP), molecular functions (MF) or cellular components (CC). The ontology trees are described as directed acyclic graphs, where each annotation can have one or more parent annotations as well as several children annotations. There can be more than one path from a leave up to the root, but there can be no path leading from a node to itself. The relationships between graph nodes are '*part_ of*', which states that the child is a structural part of its parent, or '*is_ a*' stating that a child is an instance of the parent. A small example of a molecular function subtree is shown in Figure 2.1.

The ontologies describe the following:

- **Molecular function** describes the activity of the gene at the molecular level, e.g. *nucleic acid binding GO:0003676.*
- **Biological process** is series of events accomplished by one or more ordered assemblies of molecular function, e.g. *biopolymer metabolism GO:0043283*
- **Cellular component** describes the location of the gene product in the cell, e.g. *membrane-bound organelle GO:0043227*

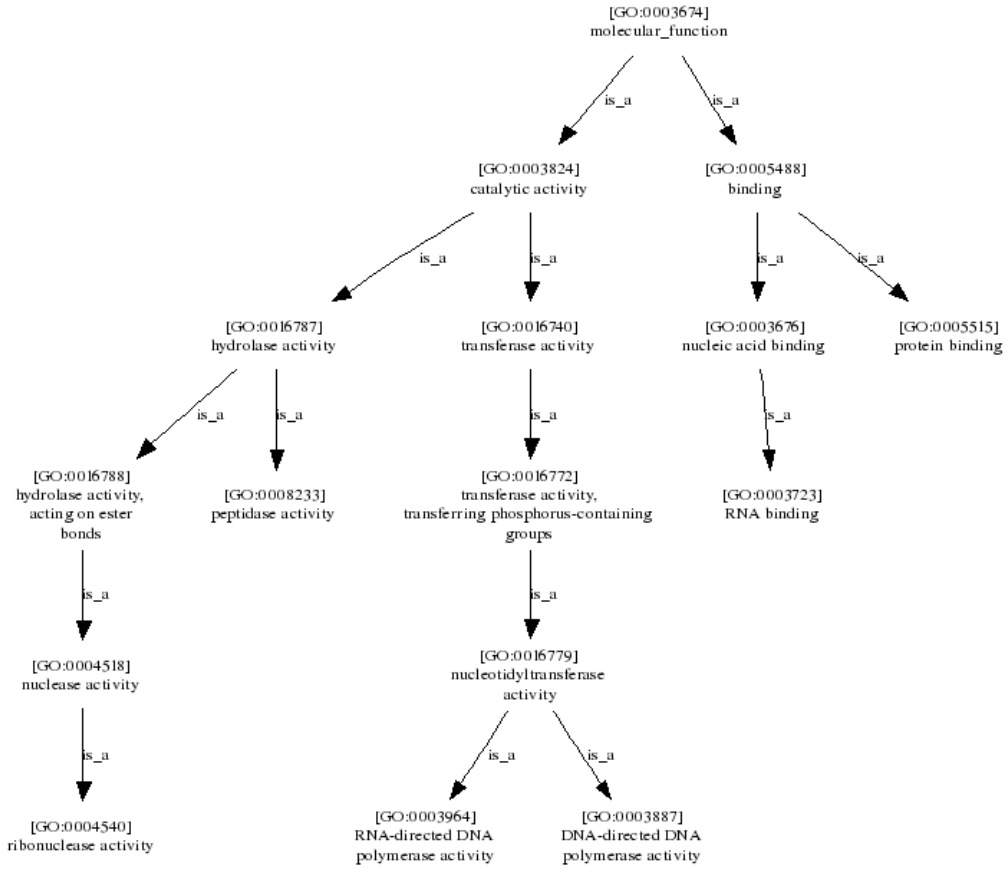

Figure 2.1: Example of a GO molecular function subtree. Here GO:0003887 is a leave and GO:0003674 is a root node. The parent of GO:0003887 is GO:0016779 and it holds genes from both of its children (GO:0003887 and GO:0003964). In this example only is_a relationships occur.

The knowledge about the function or cellular location of a gene can originate from various sources. For example data about transcription regulation can come from ChIP-on-chip experiments, DNA footprinting methods or from bioinformatics analyses. Each of the approaches have its pros and cons and ranked in GO by evidence codes. The evidence codes are the following, given by suggested reliability hierarchy:

- Inferred by Curator (IC)
- Traceable Author Statement (TAS)
- Inferred from Direct Assay (IDA)
- Inferred from Genetic Interaction (IGI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Physical Interaction (IPI)
- Inferred from Expression Pattern (IEP)
- Inferred from Sequence or Structural Similarity (ISS)
- Non-traceable Author statement (NAS)
- Inferred from Reviewed Computational Analysis (RCA)
- Inferred from Electronic Annotation (IEA)
- No biological Data available (ND)
- Not Recorded (NR)

2.1.2 Regulatory motifs

The large variety of proteins in an eukaryotic cell are transcribed by RNA polII polymerase. The RNA polII polymerase needs transcription factors to be bound to their specific binding sites in the regulatory regions to recognize the correct starting point for the transcription. Around 2% of the 6604 ORFs in *Saccharomyces cerevisiae* are so far described as transcription factors. For these approximately 150 TFs roughly a thousand binding motifs have been distinguished and described (Wingender *et al.* 2000; Dwight *et al.* 2002; Teixeira *et al.* 2006; Peterson 2004). Transcription factors can have a number of a bit different binding sites and some sites can be bound by various transcription factors. For example transcription factor Gcn4 binds both to TGATTCAT and TGACTA motifs. The later motif is also bound by Yap1. The multiple-valued relationships between the binding sites and TFs makes it complicate to always infer the correct relations between factors and target sites.

In *Saccharomyces cerevisiae* and other lower eukaryotes the distances between ORFs are relatively short and therefore the regulatory regions are short as well. Usually the regulatory region or upstream is considered to be 600 or more rarely 1000 base pair (bp) long. Most of the transcription factor binding sites (TFBSs) are 6-20 bp long and located in the close proximity of ORFs (Zhu & Zhang 1999; Qiu 2003; Vilo 2002). However, some TFBSs can be found near the 3' end of the ORF or even in the coding regions.

2.1.3 Protein-protein interactions

Most proteins need interactions with other proteins to be active in the cell. The structural units like proteasome or ribosome are big PPI complexes.

The members of such complexes are quite often regulated by the same transcription factors. For example the proteasome genes are regulated by Rpn4 (Xie & Varshavsky 2001). Some transcription factors, like Gcn4, have DNA binding domain (DBD) and activation domain (AD). The domains are used to combine the interactions between different transcription factors. Gcn4 binds to DNA with its DBD and the transcription is activated when another protein, for example MBF1, binds to GCN4's activation domain.

The protein-protein interactions have been studied in *Saccharomyces cerevisiae* for a long time. The most widely used methods for PPI detection are yeast two-hybrid and tandem affinity purification (TAP) methods (Fields 2005; Puig *et al.* 2001).

In yeast two-hybrid method the functionality of protein is studied using DNA binding and activity domains. A hybrid DBD binding to the DNA leads to gene expression of the reporter gene only when a AD hybrid protein binds to it. The method is mainly used to study DNA binding proteins. The method is reviewed by Fields 2005.

The tandem affinity purification technique uses the fusion of specific tag to the target protein, usually to the 3' end. The construct, after insertion into the host cell, is expressed from the regular promoter. The tagged protein forms a complex with its target proteins. The complexes are concentrated and fractionated on a denaturing gel. The purified complexes are further analyzed with the mass spectrometry to detect the interacting proteins. For a detailed overview of the method look at Puig *et al* 2001.

2.2 Bioinformatics approaches

2.2.1 Pattern Discovery

Pattern discovery is one of the widely used bioinformatics approaches for transcription regulation studies. Several tools have been made to search for the motifs that occur in the input sequences more frequently than in background or random sequences (Vilo 2002; van Helden, André, & Collado-Vides 1998; Brazma *et al.* 1998). The biological motivation behind these extensive searches is the fact that similar genes (either similarly expressed, similarly located in the cell or with similar functions) are often regulated by the same transcription factors.

2.2.2 Phylogenetic footprinting

The term *phylogenetic footprinting* was first introduced in 1988 by Tagle and is defined as a phylogenetic comparison, that reveals evolutionally conserved functional elements from homologous genes (Duret & Bucher 1997). The method is based on different mutation patterns, which can be found in DNA. Genomes change constantly in time, but the results of particular mutation depends on its phenotypic effect. Most of the mutations, whose outcome is negative to the host, will be removed by natural selection. Sequences that are highly conserved in time are probably functional (Duret & Bucher 1997).

There are two main problems for choosing the organisms/sequences to compare. The first one: if one chooses species which are very close in phylogenetic tree then highly conserved elements cannot be differentiated from the overall sequence. The evolutionary time has not been long enough to fix the mutations to DNA by natural selection (Duret & Bucher 1997; Cliften *et al.* 2003; Lenhard *et al.* 2003). The second problem: if one

chooses species which are too far away from each other in phylogenetic tree, then the species have been diverged too much to carry any highly conserved regions in their sequences or the species may already have different regulatory processes (Duret & Bucher 1997; Cliften *et al.* 2003). Substitutions in neutral positions in DNA happen with probability 0.5% for million years (Li, Luo, & Wu). From that we can conclude that sequences which diverged 300 million years back should have 30% of similarity if they are not under purifying selection. If there are highly conserved regions after such long time, then it refers to strong natural selection and to important functional element (Duret & Bucher 1997).

The evolutionary distance between widely compared and analyzed *Homo sapiens* and *Mus musculus* is approximately the same distance between *Saccharomyces cerevisiae* and *Saccharomyces bayanus*. The coding regions of the relative pairs are highly conserved and the non-coding regions tend to stay conserved as well. Adding more distant species to comparison we can expect that the non-functional regions are less conserved and functional regions are detectable more easily.

Chapter 3

Material and methods

3.1 Data

The *Saccharomyces cerevisiae* has been a model organism for molecular biology already a long time. The methods for exploring the baker's yeast have evolved simultaneously with the knowledge and growing datasets from wet-labs. The quickly increasing data sources need large-scale analysis pipelines.

The main research object of the project was budding yeast *Saccharomyces cerevisiae*. *Saccharomyces cerevisiae* was chosen because the main mechanisms and central transcription factors are well described. However there is still a lot ORFs not described and the complexity of gene regulation is not well known.

Additionally seven species were used for comparative pattern analysis. Three out of other seven species are closely related to *Saccharomyces cerevisiae* i.e. *sensu stricto* yeasts: *Saccharomyces bayanus*, *Saccharomyces paradoxus*, *Saccharomyces mikatae* (Kellis *et al.* 2003). In addition, phylogenetically more distant *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis* and *Yarrowia lipolytica* were chosen (Dujon *et al.* 2004). The *sensu*

stricto group we denote as Kellis dataset and the other four we refer as Pasteur data.

The wide variety of yeasts were used to cover species diverged at various timepoints in evolution and to make use of the idea that functionally active non-coding sequences tend to stay conserved during the evolution (Duret & Bucher 1997; Cliften *et al.* 2003). The evolutionary relationships between the studied yeasts are shown in Figure 3.1.

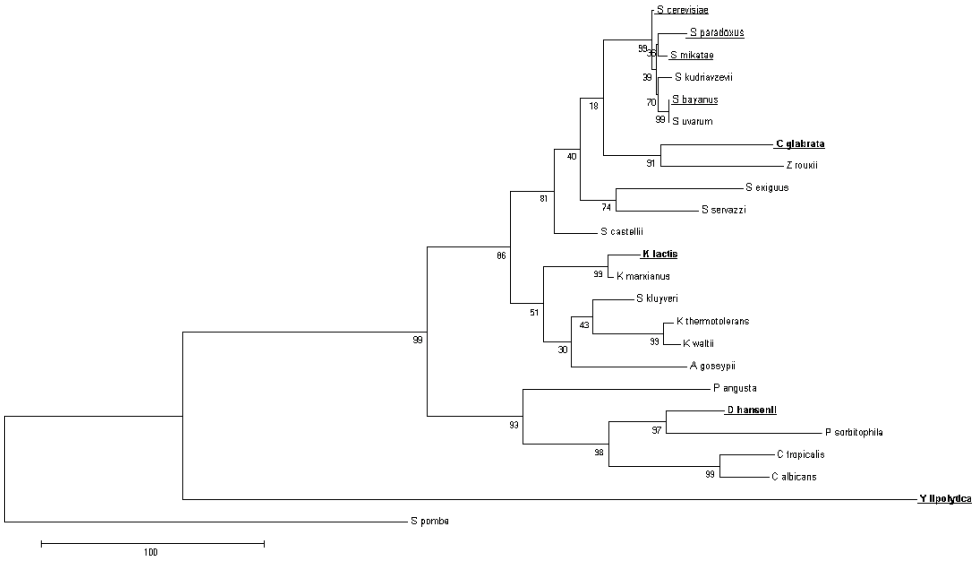

Figure 3.1: Phylogenetic tree of yeast species derived from 25S rDNA sequences. The species used in the current thesis are underlined. The picture originates from Dujon *et al* 2004.

3.1.1 Sequences

All the experiments were done using 600 base pair (bp) sequences immediately upstream from predicted or verified open reading frames (thereinafter upstream sequences or upstreams). The ORFs and upstream sequences

Table 3.1: Size and number of ORFs of yeasts

Species	Genome size (Mbp)	Chromosomes	ORFs*	Sequenced by
S.cerevisiae	12.2	16	6713	(Goffeau *et al.* 1996)
S.paradoxus	11.8	16	4788	(Kellis *et al.* 2003)
S.mikatae	12.1	16	4525	
S.bayanus	11.5	16	4492	
C.glabrata	12.3	13	5283	(Dujon *et al.* 2004)
K.lactis	10.6	6	5329	
D.hansenii	12.2	7	6906	
Y.lipolytica	20.5	6	6703	

* The estimated number of protein coding genes for Pasteur sequences.

for *Saccharomyces cerevisiae* were obtained from Saccharomyces Genome Database (SGD). The *sensu stricto* ORFs were predicted by (Kellis *et al.* 2003) and the four distant species were predicted by (Dujon *et al.* 2004).

For most of the *Saccharomyces cerevisiae* ORFs at least one ortholog has been predicted and about one third has 5 orthologs. The orthologs from Pasteur data have quality descriptions '*weakly similar to ORF*' or '*similar to ORF*' or '*highly similar to ORF*'. In current analysis all the predictions were treated as equal. The number of orthologs are given in the Table 3.2.

3.1.2 Previously known transcription factors and regulatory motifs

Information about transcription factors and their binding sites are gathered into many databases (Wingender *et al.* 2000; Dwight *et al.* 2002; Zhu & Zhang 1999; Teixeira *et al.* 2006; Peterson 2004). These databases describe

Table 3.2: The distribution of the number of orthologs for *Saccharomyces cerevisiae* ORFs

Number of orthologs	How many ORFs
0	162
1	116
2	168
3	444
4	798
5	1734
6	1378
7	506

the relations between motifs and transcription factors, as well as relations between genes and regulators. Most of the databases incorporate only *in vivo* or *in vitro* verified binding sites and do not hold *in silico* predicted motifs. However the BiGeR database contains the *in silico* predicted motifs as well.

The known transcription factor target gene sets were extracted from YEASTRACT database (Teixeira *et al.* 2006). For 143 different transcription factors there was 12328 relations to 4248 different target genes. The average number of target genes per transcription factor is 84.8. Maximum number of target genes were connected to Arr1 (712) and least to Spt23, Rds1, Otu1, Hpc2 (1). In average there is 2.9 relations to TFs per gene, with maximum 20 TFs connected to YGR088W. Figure 3.2 displays a histogram for known regulators per gene.

In the current project we used motifs from YEASTRACT and BiGeR databases to compare the predicted motifs to previously known motifs (Teix-

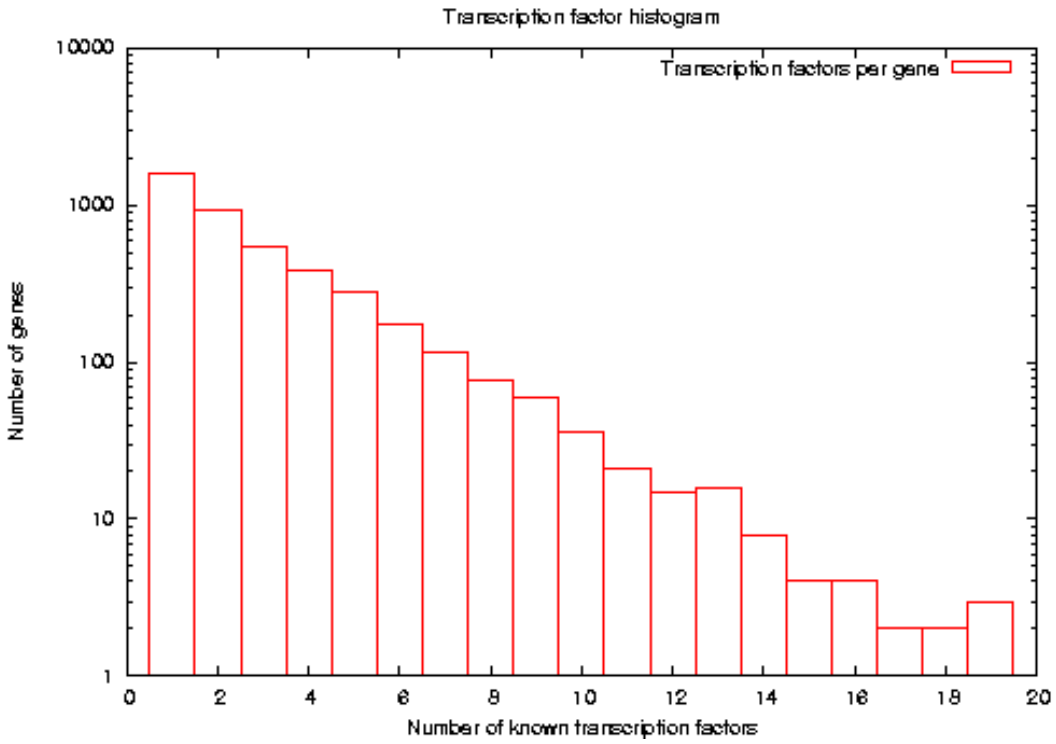

Figure 3.2: Histogram of known transcription factors per gene.

eira *et al.* 2006; Peterson 2004).

3.1.3 GO groups

In the current study we used all GO annotations except the very general ones that are either the top nodes of the GO tree or have more than 670 genes associated to it. The limit was set to 670 because larger GO groups are very general, but smaller groups, like *'regulation of biological process GO:0050789'*, may already share common regulatory motifs. We excluded 42 annotations e.g. *biological process GO:0008150*, *biological process unknown GO:0000004*. The gene groups describing the annotations consist of the genes belonging to

the annotation itself plus the genes that belong to the annotation's children. By default, not all the genes having specific annotations have the more general annotations given. We used bottom-up annotation addition to get full annotations for each gene and GO node. All annotations irrespective of the evidence codes were used to annotate each GO group with maximum number of genes.

We used 3977 different GO groups with the size ranging from 1 to 658 genes. For the GO group size distribution, see Figure 3.3.

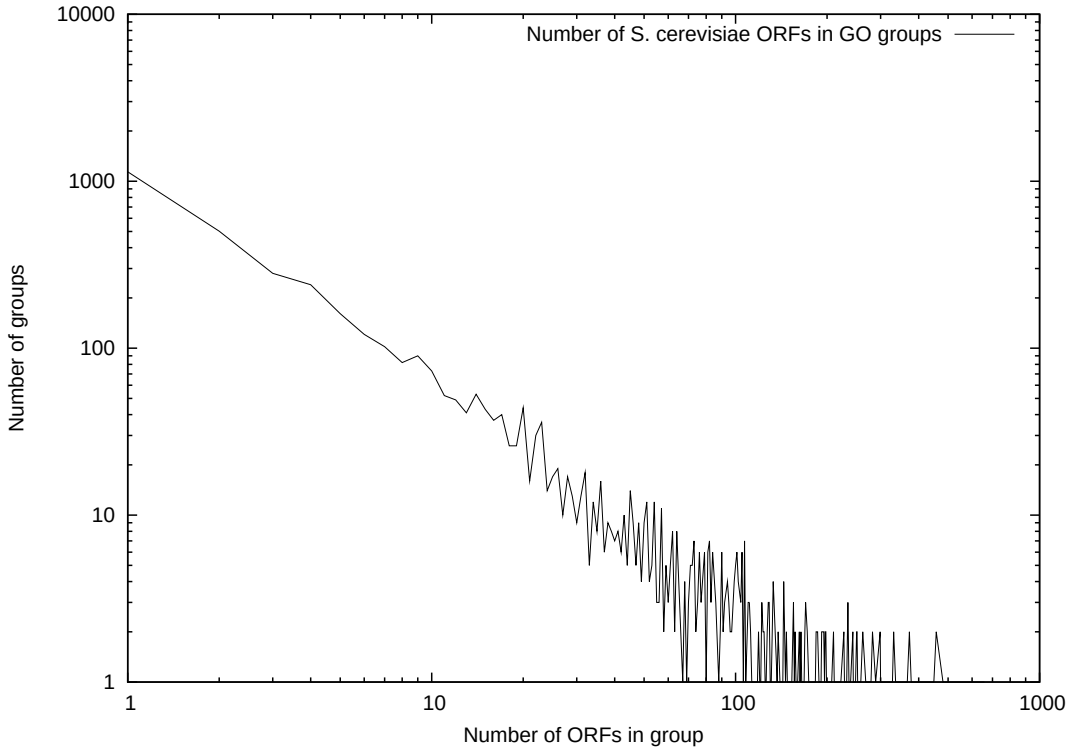

Figure 3.3: The distribution of the size of the GO groups. The size of the groups is highly variable

3.1.4 Protein-protein interaction data

The interactions between proteins form stable complexes, many of these act in transcription control or are parts of a cellular machinery like for example the proteasome. In current thesis PPI data is used to relate proteins with unknown or poorly described functions with more specific annotations. The PPI data comes usually from yeast two-hybrid or tandem affinity purification experiments. We use data from both types of experiments and do not prefer one to another.

The three datasets used are Kemmeren (Kemmeren *et al.* 2002), Gavin (Gavin *et al.* 2006) and Krogan (Krogan *et al.* 2006) that describe interactions between 3334 different ORFs. The Gavin and Krogan datasets are from two independent tandem affinity purification experiments (Krogan *et al.* 2006; Gavin *et al.* 2006). The Kemmeren dataset, on the contrary, covers several datasets from yeast two-hybrid and TAP experiments that are verified by expression analysis (Kemmeren *et al.* 2002). For the data source and ORF numbers overview look at Table 3.3

Table 3.3: Protein-protein interaction datasets overview

Dataset	Experiment type	Data source	Nr of ORFs
Kemmeren	Two-hybrid; TAP	(Gavin *et al.* 2002; Uetz *et al.* 2000; Ito *et al.* 2001; Ho *et al.* 2002; Hughes *et al.* 2000; Spellman *et al.* 1998; Roberts *et al.* 2000; Chu *et al.* 1998; Travers *et al.* 2000; Gasch *et al.* 2000)	1309
Gavin	TAP	(Gavin *et al.* 2006)	1709
Krogan	TAP	(Krogan *et al.* 2006)	2186

3.2 Pattern discovery

In this work we used SPEXS algorithm from Expression Profiler tool-set to find overrepresented patterns in all of the analyzed datasets (Expression Profiler ; Vilo 2002; Vilo *et al.* 2000). SPEXS is a program for finding common patterns from input sequences using user defined background sequences.

While using the SPEXS we set the following parameters:

- **-ms 1**, the motif has to occur at least once in our input set
- **-genorder 2**, the patterns were generated by the most frequent first
- **-binomial_prob 1.0e-04**, the output pattern has to have a probability less than 1.0e-04

The parameters were set to a low end with the aim of not losing any true results. The primary results were later depleted using different p-value and occurrence filters. The background sequences in the current project were all *Saccharomyces cerevisiae* upstreams or, if we looked for evolutionary conserved patterns, all the upstreams from eight yeasts.

3.3 Statistical evaluation

The statistical evaluation of pattern discovery is highly needed to exclude the false positive motifs from the predictions. We used randomized data to get the p-values of randomly occurring patterns and used these for filtering our original results.

Each of the result sets were filtered with the p-value threshold to minimize the false positive results. The threshold was calculated with expectation of 0.01% of results to be random. For most of the GO groups we believe to have few if any real random motifs.

Algorithm 1 Calculating p-value threshold T_g for input size g

Require: $r > 0$ {*Number of calculation runs*}

Require: $g > 0$ {*Number of genes in input set*}

Require: U {*The set of all upstreams*}

Require: p {*The pre-defined p-value threshold for SPEXS*}

$P = []$ {*Declare empty array for storing values*}

for all $i \in (1, 2, ..., r)$ **do**

$U_{g,i} = \text{rand}(g, U)$ {*Create g random upstreams from all upstreams U*}

$P_{g,i} = \text{SPEXS}(U_{g,i}, p = 1.0)$ {*Run SPEXS with random input and p-value threshold = 1.0*}

$P_{g,all} = \min(P_{g,i})$ {*Store the minimum p-value from SPEXS output*}

end for

$T_g = \text{avg}(P_{g,all})$ {*The average p-value from r runs for the input size g*}

return T_g {*Return p-value threshold T_g for input size g*}

After having the p-value threshold from randomized data we filter the SPEXS output according to algorithm 2.

Algorithm 2 Filtering the SPEXS output

Require: $g > 0$ *{The group size has to be bigger than zero}*

Require: $1 \leq T_g \leq 0$ *{Significance threshold for given group size g}*

```
for group with size g do
    use T_g threshold
    for all patterns in group g do
        if pattern p − value < T_g then
            keep the pattern
        else
            remove it
        end if
    end for
end for
return g {Return all the important patterns}
```

3.4 Expansion of groups by PPI data

We know that protein complexes, like proteasome, are regulated by the same factor. Knowing the protein complex specific regulatory motif we are able to find other proteins that could be members of these complexes. A contrary hypothesis can be stated: if genes interact and share a common motif then they can be regulated by the same factor. To study this idea we propose an expansion algorithm 3.

This approach can be applied at various datasets e.g. GO groups or known transcription factor target sets. With this method we are able to find

Algorithm 3 Expansion of gene groups by PPI data

Require: G *{GO group}*

Require: I *{All PPI pairs}*

Require: m *{GO specific regulatory motif}*

$G' = []$ *{Declare empty array for expanded GO group}*

$C = []$ *{Declare empty array for genes that interact with genes inside GO and have the same m pattern}*

for all $g \in G$ **do**

{For all genes in GO group}

$g \rightarrow G'$

{Add g to expanded GO group}

if (g,i) $\in$ I **then**

{If there exists an interaction between g and i}

$i \rightarrow G$ *{Add i to expanded GO group}*

if $upstream(i) =\sim /m/$ **then**

{i has GO specific motif in the upstream}

$i \rightarrow C$ *{Add i to possible GO candidate genes}*

end if

end if

end for

return (C, G') *{Return expanded GO group G' and group of interacting ORFs that share input motif C}*

putative members of input sets based on protein-protein interactions and common regulatory motifs.

Chapter 4

Results

The regulatory complexity of an organism can be studied step by step. One of the first steps is to find relatively simple relations between regulators and their targets. The regulators can be transcription factors that act alone or through protein-protein interactions. The targets can be genes having previously described binding with the factor or genes acting similarly in expression, having similar GO annotations or interacting with each other. In the current thesis we look at the both type of, known and non-verified, targets.

4.1 Randomization

It is easy to make lots of predictions on functions, binding sites, relations between genes etc with *in silico* methods. The important step is evaluation of the predictions to exclude as many false-positive results as possible, keeping still most of the true-positives. The extensively used approach is to estimate the false-positive rate by using random data.

In the current thesis we randomized the upstream regions into groups of various sizes. The group sizes were taken the same as the size of real

target sets. Using the pattern discovery approach on the random groups we calculated the thresholds to filter the primary results. The limit between the random and non-random motifs were calculated from the average p-value of the best motifs from ten motif discovery runs. The method for randomization is described in previous section 3.3 with algorithm 1.

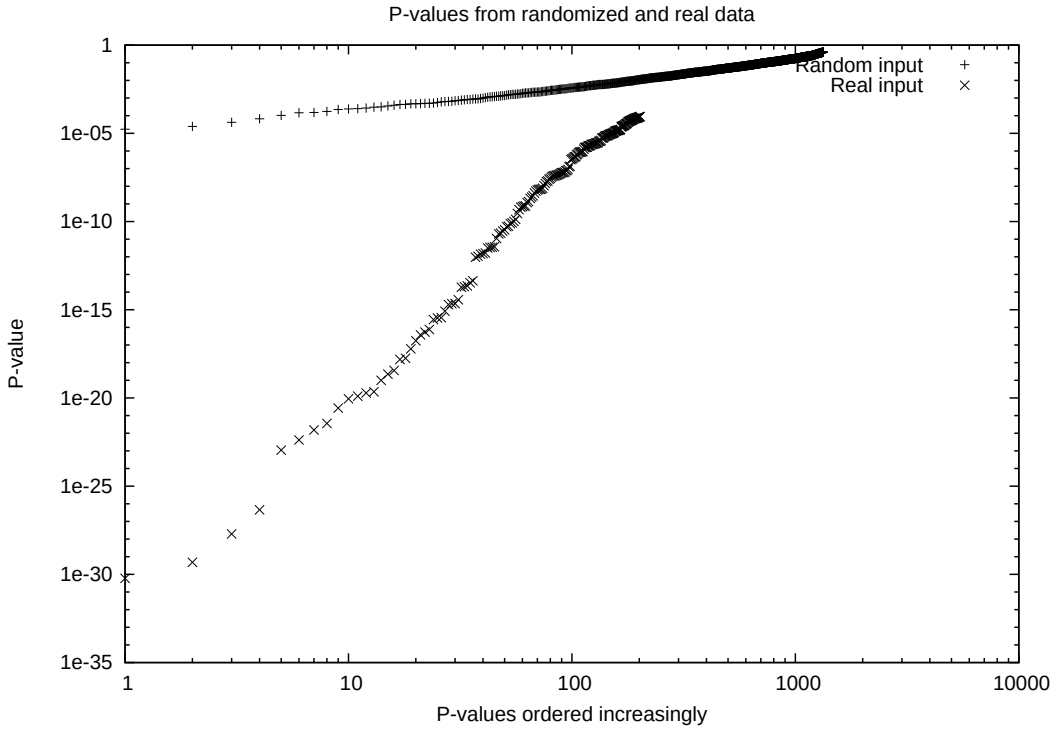

Figure 4.1: p-values from real and random data with the input size of 160 genes

The comparison of the p-values for real and random set is given on the Figure 4.1. The given figure illustrates the distribution of p-values for motifs discovered from group size of 160. From the graph we can conclude that the best motifs from the real dataset have approximately 10^{20} smaller p-values than the motifs from random data. We do believe that such a difference between the random and real data supports the trustworthiness of the best

motifs from real datasets.

The random p-values are between $9.1e-05$ and $9.3e-06$. The p-values for different group sizes do not differ much, with an average of $3.9e-05$. The distribution of the p-values for different group sizes is illustrated in Figure 4.2.

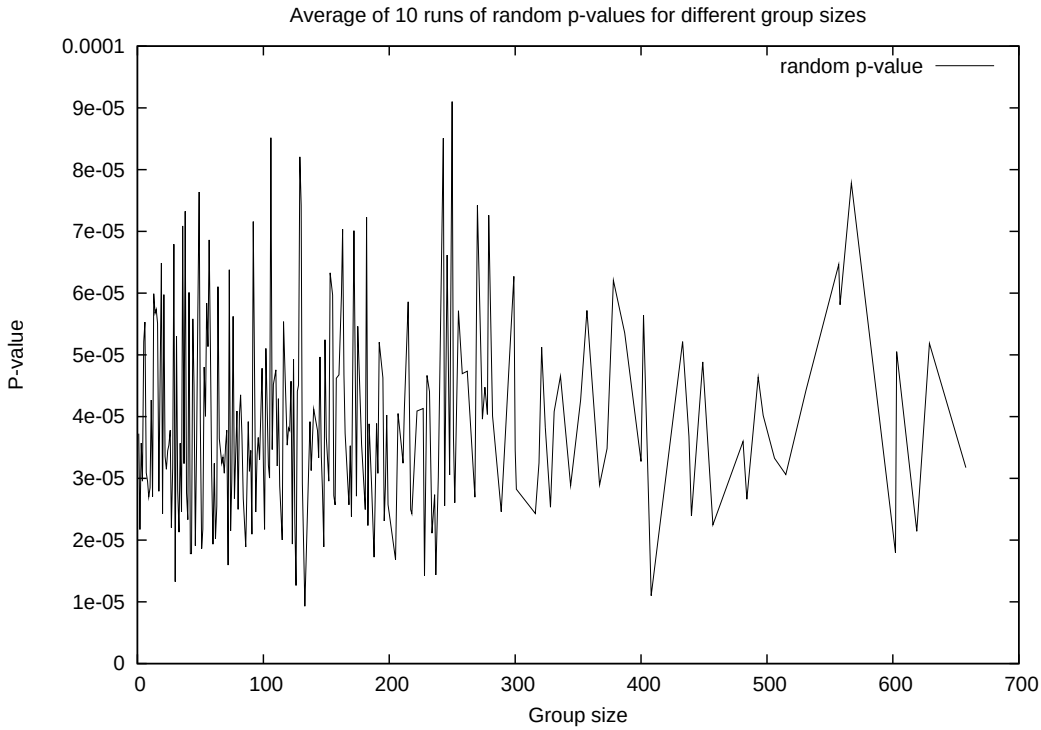

Figure 4.2: P-values from random data for each GO group size

The p-values for each group sizes were used in the following steps to filter out the non-random signals. In the next steps of analysis we considered true motifs only the signals that had p-values lower than the random threshold for given group size.

4.2 Pattern discovery for known transcription factors' target groups

Many of the transcription factors have already been identified with their relative binding sites (Dwight *et al.* 2002). The *in vivo* and *in vitro* methods like DNA footprinting or ChIP-on-chip have given more or less specific binding sites for most of the factors. Still there are TFs with a number of known target genes and uncharacterized binding sites. These groups of targets and TFs are very challenging datasets for bioinformatics and especially for pattern discovery methods.

During the project we used target gene sets for 143 transcription factors as described in previous section 3.1.2. Each of the group was analyzed with two datasets. The first set consisting only *Saccharomyces cerevisiae* upstreams and the second with the addition of orthologous upstreams. The motifs were filtered with the random threshold and only putatively non-random signals were analyzed further. For 138 TFs we could find at least one motif with the p-value smaller than the random threshold.

We found 7381 patterns for the 143 TFs, 2020 of them are distinct [1]. For almost half of the TFs we are able to find the known motif as the strongest from our prediction. In some cases we were able to find a motif for a transcription factor that did not had any motif related before.

The best example is Gat3 with very strong motif TACTTCGAAGC in *Saccharomyces cerevisiae* (p-value $1.6e-26$) that is also conserved in orthologs (p-value $2.1e-31$). The motif is not characterized in the databases before. From the Gene Ontology datasets we find that the motif overlaps

[1] Distinct - edit distance is at least one and motif is not a sub-motif for another motif (e.g. CACGTG and CTCGTG are distinct but CACGTG and CACGT are not)

well with *telomerase-independent telomere maintenance (BP)* and *helicase activity (MF)* specific motifs. The Gat3 is not well characterized transcription factor. Based on our results we propose that it can be related to the regulation of the genes with helicase activity taking part in telomere maintenance.

We were able to find motifs for other weakly characterized transcription factors as well. The newly found signals quite often overlap with already known motifs belonging to well described TFs. This suggests that these transcription factors can act as protein-protein complexes for example. For more results look the Table 1 in Appendix.

4.3 Pattern discovery in GO dataset

The Gene Ontology is a major knowledge base of biological data. The GO annotations relate genes with similar molecular function, genes participating in the same biological process and genes having the same location in the cell. All the annotations describe a small subset of all the genes. The subsets can often be regulated by the same transcription factors.

The fact that functionally related genes can be regulated by same regulators has been used earlier to find function specific regulatory motifs and to relate these to known transcription factors (Kellis *et al.* 2003; Cliften *et al.* 2003). In the current project we look for overrepresented patterns for each GO group. The groups are analyzed similarly to TF target sets. Both *Saccharomyces cerevisiae* and orthologous sequences are used to detect interesting motifs.

For 2704 GO groups out of 3977, we found at least one pattern below the SPEXS threshold $1.0e-04$ and 2006 groups remain after filtering out the

GO groups with all patterns having better (i.e. smaller) p-value than the random threshold.

The GO groups with patterns having the smallest p-values are often covering a small subtree from the GO tree. This was however expected with our bottom-up annotations, i.e. if the most specific GO group has highly overrepresented patterns, then these patterns occur more frequently than random in the more general parent GO groups as well. Usually in this case the p-value gets worse (i.e. increases) while going up to root in the GO tree, indicating that the motif was child-node specific.

In the set of top 100 GO groups with the top p-values the following GO annotations are represented in the Table 4.1. The GO groups showed in the Table 4.1,ranking between 10 and 100, are the first groups for each TF motif. If there is several groups with the same (sub)motif, then only the first GO group is shown.

Table 4.1: Patterns with the smallest p-values from GO groups

Nr	GO	CC|MF|BP	GO description	p-value	ORFs	pattern	related TF
1	GO:0000943	CC	retrotransposon nucleocapsid	2.57319e-136	91	TGTTGGAATA	Mot3
2	GO:0006319	BP	Ty element transposition	2.12078e-135	92	TGTTGGAATA	Mot3
3	GO:0006313	BP	DNA transposition	9.80204e-126	105	TGTTGGAATA	Mot3
4	GO:0006310	BP	DNA recombination	1.21499e-95	186	TGTTGGAATA	Mot3
5	GO:0003723	MF	RNA binding	4.21101e-75	316	TGTTGGAATA	Mot3
6	GO:0003964	MF	RNA-directed DNA polymerase activity	1.45907e-60	48	TGTTGGAATA	Mot3
7	GO:0005515	MF	protein binding	2.37398e-57	493	TGTTGGAATA	Mot3
8	GO:0006259	BP	DNA metabolism	5.76027e-56	557	GAGGAGAACTTCTA	Mot3*
9	GO:0003676	MF	nucleic acid binding	1.53731e-53	567	TGTTGGAATA	Mot3
10	GO:0005730	CC	nucleolus	3.73244e-53	222	AAAATTTT	
21	GO:0008233	MF	peptidase activity	1.09457e-34	155	GCAAGGATTGATAAT	
22	GO:0005830	MF	cytosolic ribosome (sensu Eukaryota)	1.89356e-31	160	CCGTACA	Rap1
23	GO:0008652	BP	amino acid biosynthesis	5.62027e-31	102	TGACTCA	Gcn4
27	GO:0000502	CC	proteasome complex (sensu Eukaryota)	5.39292e-28	407	GGTGGCAAA	Rpn4
31	GO:0016788	MF	hydrolase activity, acting on ester bonds	1.75131e-25	274	ATAATGTAATA	Hcm1?

Continued on Next Page...

Table 4.1 – Continued

Nr	GO	CC|MF|BP	GO description	p-value	ORFs	pattern	related TF
34	GO:0016772	MF	transferase activity, transferring phosphorus-containing groups	1.10842e-23	331	GATTGATAATG	
36	GO:0031974	CC	membrane-enclosed lumen	1.74176e-23	658	GCGATGAG	Esr1/Mec1?
50	GO:0006260	BP	DNA replication	4.16142e-21	105	ACGCGT	Mbp1
70	GO:0004386	MF	helicase activity	1.53663e-14	90	CCTCGACTAA	Xbp1
98	GO:0009068	BP	aspartate family amino acid catabolism	1.91911e-11	10	AGCACGTGAC	Pho4

? marks probable regulator, the relation between the motif and TF is not known before

4.4 Expansion of GO groups by PPI

The Gene Ontology groups were expanded based on the protein-protein interaction data from Kemmeren, Gavin and Krogan datasets described in 3.1.4. Expanding the group with interacting proteins gives us hints about the proteins both in and outside of the group. The example of GO:0008652 *amino acid biosynthesis* expansion by PPI and the additional information of Gcn4 binding site TGACTC is illustrated in Figure 4.6

The genes belonging to the GO group are noted as rectangles and proteins interacting with them are shown as circles. If a gene has the TGACTC motif in the upstream, then the figure is colored red. On the picture one can see that there are two proteins interacting with the GO members and share the Gcn4 motif - YLR058C and YDR172W. The first protein is part of a lysine degradation and glycine, serine and threonine metabolism pathways and related to amino acid metabolism. Therefore the protein could be a target of Gcn4 TF and belong to this GO group as well. The second protein has translation termination function and about this protein we have currently no clue if it could be regulated by GCN4 or does it have a function related to amino acid biosynthesis.

4.5 From protein-protein interactions to regulatory motifs and GO annotations

Previously we showed how known GO group can be expanded with PPI and known regulatory motifs. In this step we show how to start analyzing genes from the opposite direction. As we have mentioned earlier, interacting proteins are often regulated by the same transcription factor. We also showed

that genes with common regulator share similar motifs in their regulatory regions and often interact with each other. With this knowledge we can show how to examine genes that interact with each other.

Using the Kemmeren protein-protein interaction data we look for groups of genes that interact with each other and form connected graphs (PPI graphs). The ORFs are divided into groups recursively starting from a random ORF.

The groups are then searched for non-random Gene Ontology annotations with GOSt tool set (Reimand 2006). For each of the PPI graphs the GO annotations that have the probability less than random threshold are kept. The sets of ORFs are then taken as input for pattern discovery step with SPEXS to look for potential regulatory motifs. In this step of pattern discovery we look for non-discrete motifs. We allow wildcard[2] positions in the motifs, at most 2 wildcards per motif. The motifs are sorted by their p-values increasingly and the motif with the smallest p-value at the given input size is taken.

4.5.1 Pipeline example

A set of ORFs without any previous knowledge is analyzed. The input size is 5 ORFs: YER099C, YKL181W, YBL068W, YHL011C, YOL061W. The ORFs are connected with each other in the following way, Figure 4.3. The five ORFs make connected graph because they are pairwise connected.

The input set is analyzed with Gene Ontology tool GOSt (Reimand 2006). The tool is used to find all GO annotations mapping the input ORFs. The annotations with probability value smaller than GOSt analytical threshold is given in the output. From the Figure 4.4 we can see that the best GO

[2] wildcard is a special character representing more than one character

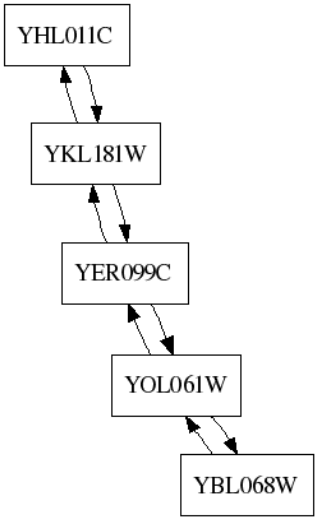

Figure 4.3: The connected graph of input ORFs. The arrows indicate the known interactions between genes

annotation is *GO:0004749 ribose phosphate diphosokinase activity.*

From the pattern discovery we find 3 motifs occurring in all input sequences. The motif with the smallest p-value is taken into further analysis. In this case it is AATG.TTA, where . denotes the wild character i.e. A, C, G or T can be at that position.

The pattern can be matched back to all *Saccharomyces cerevisiae* upstream sequences to check if the pattern is highly specific to the input set or it belongs to some more general GO node.

The expression dataset can be checked to evaluate the potential co-expression of the ORFs belonging to the connected graph set. One still has to keep in mind that the Kemmeren dataset had already verified the data with expression analysis, so a similar expression should be no big surprise, but rather an expected result.

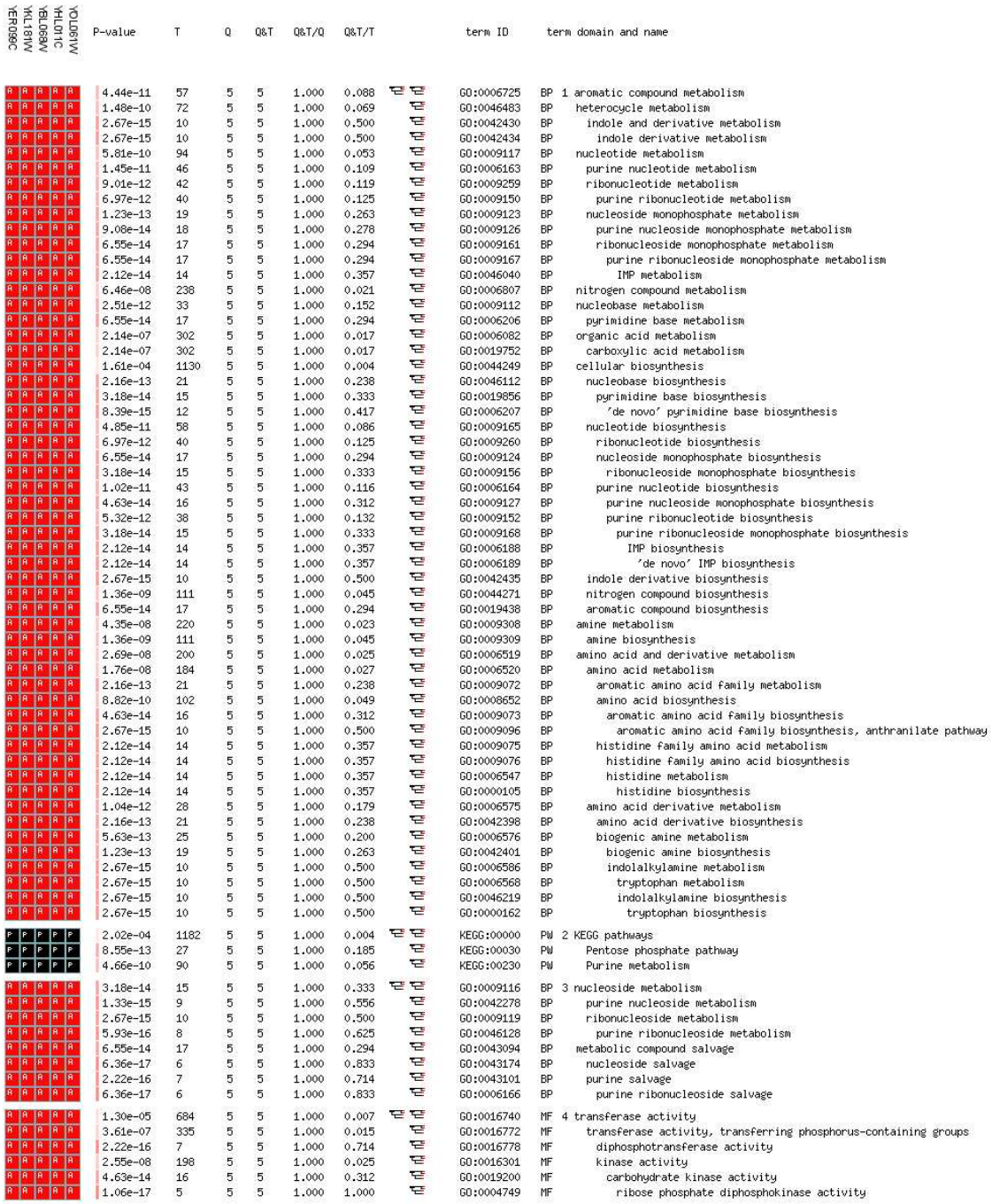

Figure 4.4: The GO annotations for the 5 ORFs with GOSt tool (Reimand 2006). The bottom line on the figure describes the most relevant GO annotation for the input set

From the Figure 4.5 we can see that the expression of the interacting ORFs is very similar in various stress conditions (Gasch *et al.* 2000).

The motif has no relationships to any GO annotations and is not known regulatory motif before. The very similar expression clusters suggest that the proteins might be regulated by a same factor and work in protein complexes.

We checked from the literature that the our analyzed genes are five phosphoribosyl diphosphate synthase-homologous genes. These genes have been previously described as proteins being active only in a complex of at least 3 subunits out of 5 (Hove-Jensen 2004).

4.6 Web-tool Gviz-PPI

The method for expansion a group of genes by addition of interacting proteins is incorporated to a web-tool Gviz-PPI and is publicly accessible at `http://bioinf.ebc.ee/u/peterson/gviz/`. The tool allows to input a set of *Saccharomyces cerevisiae* ORFs and a motif to find out all the interactions the input gene set has in Kemmeren, Gavin and Krogan datasets. The ORFs that have input motif in their upstream regions are visualized to give a fast and easy overview of the data. It is possible to use regular expressions for the motifs e.g. to use motif like TGA.TC. The tool is implemented in Perl and visualization is enabled by the Simple Web Object Graphics language SWOG (Hansen 2005).

4.7 Data update to BiGeR

The BiGeR database incorporates data about *Saccharomyces cerevisiae* transcription factor binding sites, both predicted by bioinformatics methods and

from the wet-lab experiments (Peterson 2004). In the database the TFBS are described as regular DNA motifs, regular expressions, consensus sequences or with position weight matrices. The binding sites are related to known transcription factors and genes they regulate, if the relations are known.

With the previously described analysis a number of binding sites get additional descriptions and relations from GO. The updated motifs contribute to more deeper knowledge in the database and help to distinguish more relevant patterns for each transcription factor.

From GO pattern discovery motifs that had p-value smaller than $1.0e-07$ were considered as interesting. The even stricter p-value threshold was chosen because of the large number of motifs with p-values around random threshold. The motifs chosen for BiGeR update should be more reliable. The motifs were then filtered to find distinct motifs. Finally 700 distinct motifs with 5714 connections to 178 different GO annotations were added to the database.

The updated and newly discovered regulatory motifs have been made publicly available in the BiGeR database at `http://bioinf.ebc.ee/biger/`. An example output of a query describing TGACTC motif is given in Figure 4.7.

Figure 4.5: The clustering based on stress response (Gasch 2000)

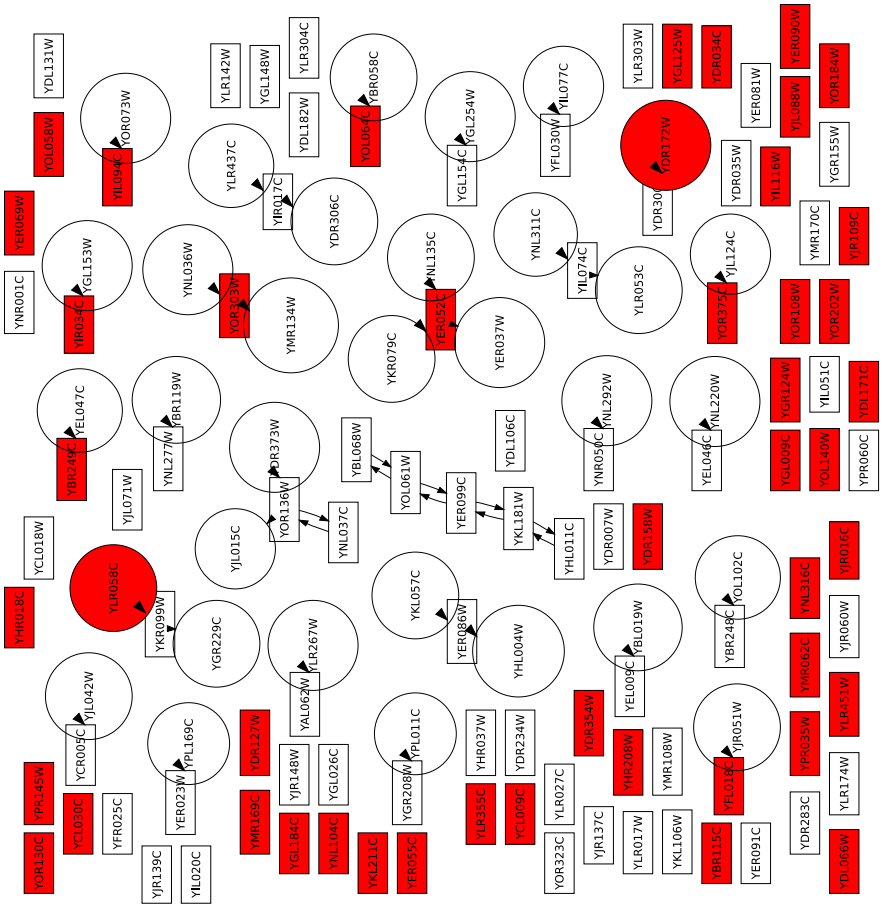

Figure 4.6: The expansion of GO:0008652 using Gcn4 motif TGACTC. The rectangles are ORFs belonging to the GO group, the circles are interacting proteins. The red fulfil indicates that the TGACTC pattern is in the upstream region of the ORF

nr	signal	coord start	coord end	description	signal type	dbref	Regulated Gene	Regulating factor
1	TGACTC	-126	-108	DNase I footprinting	oligo	TRF:R00646	HIS3/YOR202W	GCN4
2	ATGACTCAT	-102	-92	DNase I footprinting direct gel shift/cancil interference	oligo	TRF:R00649	HIS3/YOR202W	SKO1
3	TAATAGTGACTCCGGTAAATT	-240	-229	DNase I footprinting gel retardation	oligo	TRF:R00650	HIS4/YCL030C	RAS1
4	TGACTC	-250	-????	DNase I footprinting	oligo	TRF:R00651	HIS4/YCL030C	GCN4
5	TAATAGTGACTCCGGTAAATTAGTTAATTAA	-240	-217	DNase I footprinting gel retardation	oligo	TRF:R00655	HIS4/YCL030C	RAS1
6	TGACTC	-203	-190	DNase I footprinting	oligo	TRF:R00656	HIS4/YCL030C	GCN4
7	TTGACTCTCaaaaATGATTCAT	-140	-108	DNase I footprinting gel retardation	oligo	TRF:R02022	TRP4/YDR354W	GCN4
8	TTGACTCTT	-350	-350	DNase I footprinting	oligo	TRF:R04022	ADE4/YMR300C	GCN4
9	SNSNWNWNWTGACTCATNS				consensus	TRF:R04800	ADE4/YMR300C	GCN4
10	TAATAGTGACTCCGGTAAATT	-240	-229	Science 246:931-935 (1989)	oligo		HIS4/YCL030C	RAS1
11	TCGAACTGACTCTCTAATAGTGAC	-261	-240	PNAS 1992, 89:6746-6750	oligo		HIS4/YCL030C	RAS1
12	TAATAGTGACTCCGGTAAATTAGTTAATTAA	-240	-219	Science 246:931-935 (1989)	oligo		HIS4/YCL030C	RAS1,PHO2
13	TGACTC	-199	-194	Proc. Natl. Acad. Sci. USA 83:8516-8520 (1986)	oligo		HIS4/YCL030C	GCN4
14	ACAGTGACTCACGTTT	-203	-188	Mol Cell Biol 1991, 11:3642-3651	oligo		HIS4/YCL030C	GCN4
15	TGACTC	-255	-250	Proc. Natl. Acad. Sci. USA 83:8516-8520 (1986)	oligo		HIS4/YCL030C	GCN4
16	TTTGACTCTTC	-166	-158	EMBO J 9:2651-2657 (1990)	oligo		TRP4/YDR354W	GCN4
17	TGACTC	-184	-179	J Biol Chem 1997, 272:13348-13354	oligo		ADE5,7/YGL234W	RAS1
18	GCCGACTGACTCGTGTCCTGGT	-190	-169	PNAS 1992, 89:6746-6750	oligo		ADE5,7/YGL234W	RAS1
19	TGACTC	-217	-212	J Biol Chem 1997, 272:13348-13354	oligo		ADE5,7/YGL234W	RAS1
20	TTCAGTTGACTCGCCCCCGTCGG	-223	-202	PNAS 1992, 89:6746-6750	oligo		ADE5,7/YGL234W	RAS1
21	TTGACTCTT	-358	-350	J. Biol. Chem. 266:20453-20456 (1991)	oligo		ADE4/YMR300C	GCN4
22	TGACTCA	-342	-336	Yeast 1995, 11:1367-1380	oligo		ARG1/YOL058W	GCN4
23	TGACTCA	-180	-174	Yeast 1995, 11:1367-1380	oligo		ARG8/YOL140W	GCN4
24	ACTCATTGACTCTCTCGCTGACCT	-169	-130	PNAS 1992, 89:6746-6750	oligo		ADE2/YOR128C	RAS1
25	GACAAATGACTCTCTTGTTGCATG	-202	-181	PNAS 1992, 89:6746-6750	oligo		ADE2/YOR128C	RAS1
26	TGACTC	-122	-117	Proc. Natl. Acad. Sci. USA 83:8516-8520 (1986)	oligo		HIS3/YOR202W	GCN4
27	ATGACTCTT	-123	-115	Mol Cell Biol 1995, 15:7059-7066	oligo		HIS3/YOR202W	GCN4
28	ATGACTCWT			5 prime; MCS:6.1; Bet category:ChIP: Gcn4, CCS:44; Interpreted:Known, Gcn4Ras1	regexp		HIS3/YOR202W	GCN4

Figure 4.7: BiGeR output of all patterns matching known Gcn4 binding site TGACTC. The top motifs come from various *in vivo* and *in vitro* analyzes. The last two rows describe motifs from ChIP-on-chip and Gene Ontology specific data

Chapter 5

Discussion

Combining various data sources like protein-protein interactions with known TF binding sites or with Gene Ontology data we are able to verify the connections between data sources, make new function or annotation predictions and gain additional knowledge about the complex mechanism of a cell. In the current scientific world there is huge data abundance and therefore a lot of effort has been made to combine the data and find the reliable connections between methods and datasets.

In the current thesis we looked at regulatory motifs with the help of protein-protein interactions, Gene Ontology and known transcription factor target sets. Firstly, we were able to find the previously known motifs, which assured that our methods are working well. The large scale analysis on Gene Ontology gave us a great amount of motifs related to one or many ontologies.

With the help of known facts like Gcn4 is the main regulator of amino-acid biosynthesis and knowing the binding sites for Gcn4, we could verify the known ontology specific motifs and start to predict the unknown.

The new patterns were predicted with SPEXS algorithm and related to GO nodes. The probable motifs were filtered with random p-value distri-

butions. With a very small false-negative approximation we were able to find GO specific motifs for almost half of the input annotations. Few of the groups were analyzed more deeply and verified with the data from previously published studies.

We found several clusters of motifs that differ by one or two bases. The changes of the binding sites may affect the binding affinity of TF. For example changes in proteasome specific motif GTGGCAAA are related to changes in gene expression similarity (data not shown).

The protein-protein interaction datasets gave us a possibility to use the previously known fact that interacting proteins tend to be regulated by the same factors and have alike functions with the pattern discovery and GO annotations to expand the GO groups.

The GO group expansions were made to find genes that do interact with the members of the GO group, but do not have the relationship with the annotation. We used putative regulatory motifs to describe the interacting proteins and add them as possible members of GO groups.

In many cases we saw that same motif belongs to different GO groups. If the groups belong to the same GO subtree, then this is observation is expected. We think it would be interesting to look for motifs that belong to not connected GO annotations.

In the future work we propose to use less discrete patterns to describe the GO groups, because of the variability of binding sites for the TFs. The usage of position weight matrices or regular expressions will make it easier to describe similar motifs in a compact way. The datasets could be filtered according to some quality thresholds, e.g. GO evidence codes or the similarity of orthologous genes. The datasets would be smaller but the predictions could be more reliable. The protein-protein interactions could be filtered according

to the expression dataset to have more trustworthy data i.e. to use the same approach as Kemmeren et al (Kemmeren *et al.* 2002).

Chapter 6

Conclusions

The combination of datasets, methods, quality assessments is a real challenge in current bioinformatics. A big number of datasets can give lots of information about the topic, however it can be hard to find the most important data and relations from it. We combined data with evolutionary background, pattern discovery, the functional annotations from GO and protein-protein interactions to develop methods and pipeline for large scale analysis of transcriptional regulation in *Saccharomyces cerevisiae* . The pattern discovery step was quite successful and the usage of random data and phylogenetic data helped to deplete the probable false-positive results even more. The PPI data gave us hints about ontologies that could be expanded and how to predict functions to unknown proteins.

The outcome of the thesis is knowledge about how to combine relevant data sources to understand better the transcription factor binding sites and the complexity of the regulatory mechanism. We propose a few approaches towards the function prediction and describe a large variety of putative regulatory regions.

Summary

We know the whole genome of *Saccharomyces cerevisiae* for 10 years already. The baker's yeast has been studied extensively and thoroughly, but still not all the genes and molecular mechanisms of this fairly simple eukaryote has been described. The challenge for understanding this one cell organism is more and more drifted to bioinformatics. The large datasets covering different aspects of regulatory mechanisms and regions have been published and the main challenge nowadays is to to put all the information together, to connect the small pieces of this huge puzzle.

In this thesis we gave an overview of the possibilities to join the widely used and up-to-date source of Gene Ontology, the sets of protein-protein interactions and pattern discovery methods. The approaches were used to study the connections one can find between the sets, to make predictions of regulatory motifs and widen the Gene Ontology groups.

We were able to show how to find GO specific putative regulatory motifs using PPI data or how to broaden the known GO annotation based on PPI and known regulatory motifs. These experiments helped us to understand the ways how one can start annotating functions or regulatory regions from a set of genes with a little or no previous knowledge.

The thesis gave a glimpse of the complexity of connecting large datasets, analyzing the results and predicting new knowledge. The data sources grow

rapidly, so there will be more and more challenges to solve with a similar approach.

Summary in Estonian

Saccharomyces cerevisiae regulatoorsete motiivide ennustamine ja kirjeldus

Saccharomyces cerevisiae genoom sekveneeriti juba kümme aasta tagasi. Hoolimata pagaripärmi põhjalikust teaduslikust uurmisest viimase kahe kümnendi jooksul on siiski paljude geenide ja molekulaarsete mehhanismide funktsioonid teadmata. Eksperimenditulemuste üha suurenev kasv loob võimalusi bioinformaatika laialdaseks kasutamiseks funktsioonide ennustamisel. Tänapäeva bioinformaatika suurimaid väljakutseid on erinevate andmehulkade sidumine nii, et kõik kokku moodustaks tervikliku pildi rakus toimuvast.

Käesolevas töös anti ülevaade kuidas siduda laiadlaselt kasutusel olevat geeniontoloogia andmestikku, valk-valk interaktsioone ning mustrite otsimismeetodeid. Töös loodi seosed andmestike vahel ning kasutati neid regulatoorsete motiivide ennustamiseks ning geeniontoloogiate laiendamiseks.

Töös on toodud näited, kuidas kasutades valk-valk interkatsioone on võimalik leida regulatoorseid motiive. Samuti, kuidas laiendada geeniontoloogiaid spetsiifilisi regulatoorseid motiive ning valk-valk interaktsioone kasutades. Teostatud eksperimendid aitasid leida viise annoteerimaks funktsioone või regulatoorseid regioone kasutades vähest varasemat teadmist sisendgeenide hulga kohta.

Antud töö andis lühiülevaate suurte bioloogiliste andmehulkade sidumise

keerukusest, analüüsivõimalustest ning uue teadmise ennustamisest. Andmehulkade jätkuva kasvu taustal on kindlasti kirjeldatud teadmistest ja meetoditest edaspidises uurimistöös kasu.

Acknowledgements

I would like to express my sincere gratitude to my supervisor dr. Jaak Vilo for introducing the interesting world of bioinformatics to me, for all the support and guidance throughout the years and for believing in me.

I am really grateful to my friend and fellow BIITer Jüri with whom I have shared those long working nights during the last two years. Thanks for all the tips and ideas. It has and will continuously be fun to work with You.

Many thanks go to my friends in BIIT, especially to Asko, Jelena and Lemps.

I wish to thank also Jaanus, Kostja, Meelis, Pavlos and Raivo, who have supported me with tools, helped out with statistics or just given good advice.

Financial support from Estonian Science Foundation grant no 5724, EU FP6 STRE ATD and Kristjan Jaak scholarship foundation is acknowledged.

Last, but not least, I would like to thank my friends from outside of the science world and my family for their continuous support during the studies.

References

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Suzanna Lewis, John C. Matese, J. E. R. M. R.; Rubin, G. M.; and Sherlock, G. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics* 25(1):25–29.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, M. J.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; and Sherlock, G. 2001. Creating the Gene Ontology Resource: Design and implementation. *Genome Research* 11:1425–1433.

Brazma, A.; Jonassen, I.; Vilo, J.; and Ukkonen, E. 1998. Predicting gene regulatory elements in silico on a genomic scale. *Genome Research* 8:1202–1215.

Chu, S.; DeRisi, J.; Eisen, M.; Mulholland, J.; Botstein, J.; Brown, O.; and Herskowitz, I. 1998. The transcriptional program of sporulation in buddying yeast. *Science* 282(5389):699–705.

Cliften, P.; Sudarsanam, P.; Desikan, A.; Fulton, L.; Fulton, B.; Majors, J.; Waterston, R.; Cohen, B. A.; and Johnston, M. 2003. Finding functional

features in *Saccharomyces* geomes by phylogenetic footprinting. *Science* 301(5629):71–76.

Dujon, B.; Sherman, D.; Fischer, G.; et al. 2004. Genome evolution in yeasts. *Nature* 430(6995):35–44.

Duret, L., and Bucher, P. 1997. Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology* 7(3):399–406.

Dwight, S.; Harris, M. A.; Dolinski, K.; Ball, C. A.; Binkley, G.; Christie, K. R.; Fisk, D. G.; Issel-Tarver, L.; Schroeder, M.; Sherlock, G.; Sethuraman, A.; Weng, S.; Botstein, D.; and Cherry, J. M. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene ontology (GO). *Nucleic Acids Research* 30(1):69–72.

Expression Profiler. Webpage `http://www.ep.ebi.ac.uk/EP`.

Fields, S. 2005. High-throughput two-hybrid analysis. The promise and the peril. *The FEBS Journal* 272(21):5391–5399.

Gasch, A.; Spellman, P.; Kao, C.; Carmel-Harel, O.; Eisen, M.; G.Storz; Botstein, D.; and Brown, P. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of Cell* 11(12):4241–4257.

Gavin, A.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Scultz, J.; Rick, J.; Michon, A.; Cruciat, C.; et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868):141–147.

Gavin, A.-C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche1, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dümpelfeld, B.; Edelmann, A.; Heurtier, M.-A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon,

A.-M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; and Superti-Furga, G. 2006. Proteome survey reveals modularity of the yeast cell machiner. *Nature* 440(7084):631–636.

Goffeau, A.; Barrell, B.; Bussey, H.; Davis, R.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J.; Jacq, C.; Johnston, M.; Louis, E.; Mewes, H.; Murakami, Y.; Philippsen, P.; Tettelin, H.; and Oliver, S. 1996. Life with 6000 genes. *Science* 274(5287):563–567.

Hansen, J. 2005. Graphics language SWOG, Bachelor thesis, University of Tartu.

Harris, M.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G.; Blake, J.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J.; Hill, D.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J.; Christie, K.; Costanzo, M.; Dwight, S.; Engel, S.; Fisk, D.; Hirschman, J.; Hong, E.; Nash, R.; Sethuraman, A.; Theesfeld, C.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T.; and White, R. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32(Database issue):D258–261.

Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G.; Moore, L.; Adams, S.; Millar, A.; Taylor, P.; Bennet, K.; Boutilier, K.; et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectometry.

Nature 415(6868):180–183.

Hong, E.; Balakrishnan, R.; Christie, K.; Costanzo, M.; Dwight, S.; Engel, S.; Fisk, D.; Hirschman, J.; Livestone, M.; Nash, R.; Park, J.; Oughtred, R.; Skrzypek, M.; Starr, B.; Theesfeld, C.; Andrada, R.; Binkley, G.; Dong, Q.; Lane, C.; Hitz, B.; Miyasato, S.; Schroeder, M.; Sethuraman, A.; Weng, S.; Dolinski, K.; Botstein, D.; and Cherry, J. Saccharomyces genome database. `http://www.yeastgenome.org/ 26.03.2006`, year = 2006,.

Hove-Jensen, B. 2004. Heterooligomeric phosphoribosyl diphosphate synthase of *Saccharomyces cerevisiae*: combinatorial expression of the five prs genes in *Escherichia coli*. *J. Biol. Chem* 279(39):40345–40350.

Hughes, T.; Marton, M.; Jones, A.; Robets, C.; Stoughton, R.; Armour, C.; Bennett, H.; Coffey, E.; Dai, H.; He, Y.; et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102(1):109–126.

Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; and Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. In *Proc. Natl. Acad. Sci. USA*, volume 98, 4569–4574.

Kellis, M.; Patterson, N.; Endrizzi, M.; Birren, B.; and Lander, E. S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.

Kemmeren, P.; van Berkum, N.; Vilo, J.; Bijma, T.; Donders, R.; Brazma, A.; and Holstege, F. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell* 9(5):1133–1143.

Krogan, N.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Ignatchenko, A.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A.; Punna, T.; Peregrin-Alvarez, J.; Shales, M.; Zhang, X.; Davey, M.; Robinson, M.; Paccanaro, A.; Bray,

J.; Sheung, A.; Beattie, B.; Richards, D.; Canadien, V.; Lalev, A.; Mena, F.; Wong, P.; Starostine, A.; Canete, M.; Vlasblom, J.; Wu, S.; Orsi, C.; Collins, S.; Chandran, S.; Haw, R.; Rilstone, J.; Gandi, K.; Thompson, N.; Musso, G.; Onge, P. S.; Ghanny, S.; Lam, M.; Butland, G.; Altaf-Ul, A.; Kanaya, S.; Shilatifard, A.; O'Shea, E.; Weissman, J.; Ingles, C.; Hughes, T.; Parkinson, J.; Gerstein, M.; Wodak, S.; Emili, A.; and Greenblatt, J. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440(7084):637–643.

Lenhard, B.; Sandelin, A.; Mendoza, L.; Engström, P.; Jareborg, N.; and Wasserman, W. W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology* 2(13).

Li, W.; Luo, C.; and Wu, C. Evolution of DNA sequences. In *Molecular Evolutionary Genetics*.

Liti, G., and Louis, E. J. 2005. Yeast evolution and comparative genomics. *Annual Review of Microbiology* 59:135–153.

Peterson, H. 2004. Gene regulation database BiGeR, Bachelor thesis, University of Tartu.

Puig, O.; Caspary, F.; Rigaut, G.; Rutz, B.; Bouveret, E.; Bragado-Nilsson, E.; Wilm, M.; and Séraphin, B. 2001. The tandem affinity purification (TAP) method: A general procedure of protein complex purification. *Methods* 24(3):218–229.

Qiu, P. 2003. Computational approaches for deciphering the transcriptional regulatory network by promoter analysis. *Biosilico* 1(4):125–133.

Reimand, J. 2006. Gene ontology mining tool GOSt. Master's thesis, University of Tartu.

Roberts, C.; Nelson, B.; Marton, M.; Stoughton, R.; Meyer, M.; Benet, H.; He, Y.; Dai, H.; Walker, W.; Hughes, T.; et al. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287(5454):873–880.

Spellman, P.; Sherlock, G.; Zhang, M.; Iyer, V.; Anders, K.; Eisen, M.; Brown, P.; Botstein, D.; and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of Cell* 9(12):3273–3297.

Stormo, G. D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16(1):16–23.

Teixeira, M.; Monteiro, P.; Jain, P.; Tenreiro, S.; Fernandes, A.; Mira, N.; Alenquer, M.; Freitas, A.; A.L Oliveira, A.; and Sá-Correia, I. 2006. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Research* 34(Database issue):D446–451.

Tompa, M. 2001. Identifying functional elements by comparative DNA sequence analysis. *Genome Research* 11(7):1143–1144.

Travers, K.; Patil, C.; Wodicka, L.; Lockhart, D.; Weissman, J.; and Walter, P. 2000. Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* 101(3):249–258.

Uetz, P.; giot, L.; Cagney, G.; Mansfield, T.; Judson, R.; Knight, J.; Lockshon, D.; Narayan, V.; Srinivasan, M.; and Pochart, P. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403(6770):623–627.

van Helden, J.; André, B.; and Collado-Vides, J. 1998. Extracting regula-

tory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281:827–842.

Vilo, J.; Brazma, A.; Jonassen, I.; Robinson, A.; and Ukkonen, E. 2000. Mining for putative regulatory elements in the yeast genome using gene expression data. In *ISMB-2000*, 384–394. AAAI press.

Vilo, J. 2002. *Pattern Discovery from Biosequences.* Ph.D. Dissertation, University of Helsinki.

Wingender, E.; Chen, X.; Hehl, R.; Karas, H.; Liebich, I.; Matys, V.; Meinhardt, T.; Prüß, M.; Reuter, I.; and Schacherer, F. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research* 28(1):316–319.

Xie, Y., and Varshavsky, A. 2001. RPN4 is a ligand, substrate, and transcriptional regulator of the 26s proteasome: a negative feedback circuit. *Proc Natl Acad Sci USA* 98(6):3056–6.

Zhu, J., and Zhang, M. Q. 1999. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae. Bioinformatics* 15(7/8):607–611.

Appendix A

Tables

Table 1: TF target genes and their relative motifs

TF	Nr	Known BS	S.cerevisiae	Prob	Orthologs	p-value	Cons.	Comments
ABF1	102	TCCCCATTAACG	CGCTGAAT	1.27101e-05	*NA*	*NA*	NO	weak motif, no overlap
ACA1	2	*NA*	GCACCTTA	3.73022e-05	*NA*	*NA*	NO	weak motif, no overlap
ACE2	83	TGCTGGT	TGCTGGC	3.95396e-07	AACCAGCA	3.02497e-13	YES	medium strong motif, good overlap; strong (CA) repeats in *S.cerevisiae* upstreams
ADR1	43	GGRGK	TTGGGGTA	1.22318e-07	TTGGGGTA	3.88413e-15	YES	medium strong motif, improvement of a motif
AFT2	46	YRCACCCR	GGGTGCA	3.07926e-11	GGGTGCA	5.94077e-20	YES	strong motif, improvment of a motif
ARG80	19	CCTCTAAAGG	TTTCACTTA	1.80249e-07	TTTTGCCACCC	3.0292e-12	NO	medium strong motif, no overlap
ARG81	21	AAGTACAGTTAATAACGA	TTTCACTTA	3.11774e-07	TTTTGCCACCC	2.09836e-12	NO	medium strong motif, no overlap
ARO80	30	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	no motif found
ARR1	712	TTAATAA	ATCCGTACA	6.71412e-06	*NA*	*NA*	NO	weak motif, no overlap
ASH1	5	YTGAT	*NA*	*NA*	GTCTCCCACATCACCA	2.66303e-17	NO	no motif from *S.cerevisiae*, strong motif from orthologs with overlap

Continued on Next Page...

Table 1 – Continued

TF	Nr	Known BS	S.cerevisiae	p-value	Orthologs	Prob	Cons.	Comments
AZF1	8	TTTTTCTT	TTTTCTTTT	7.0558e-07	GCCCTCATCT	5.27129e-11	NO	medium strong motif from *S.cerevisiae* with overlap, known motif not conserved in orthologs
BAS1	70	TGACTC	TGACTC	3.91535e-11	TGACTC	1.06697e-23	YES	strong motifs from both input sets, 100% overlap with known motif
CAD1	314	TTACTAA	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found
CAT8	78	NCCDTYNVNCCGN	CCGAGC	1.43146e-06	GTATATAA	2.92261e-17	NO	weak *S.cerevisiae* motif, strong orthologous motif
CBF1	56	RTCACRTG	CACGTGA	4.49396e-39	CACGTG	4.44005e-83	YES	very strong motifs overlapping known motif
CHA4	10	CGGN10CCG	*NA*	*NA*	*NA*	*NA*	*NA*	no motif found
CIN5	133	TTACRTAA	TTACATAA	4.20119e-12	TTACATAAT	3.55377e-14	YES	strong overlapping motif, not very strongly conserved in orthologs
CRZ1	147	GNGGCKCA	GAGGCT	7.49694e-07	AGCCTC	3.81303e-16	YES	medium strong overlapping motifs
CST6	2	TTACGTAA	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found
CUP2	2	TCTTTTGCTG	TTCTTTTGCT	2.61691e-10	AATTAGTAAGC	6.91919e-11	NO	strong overlapping motif, not conserved

Continued on Next Page...

Table 1 – Continued

TF	Nr	Known BS	S.cerevisiae	p-value	Orthologs	Prob	Cons.	Comments
DAL80	26	GATAA	CTTATC	7.74645e-14	CTTATC	1.12426e-33	YES	very strong overallpig motifs from both sets
DAL81	45	GAAAATTGCGTTT	GGCCCGC	2.29767e-06	AAAAAGAAACGG	5.98394e-14	NO	weak, not overlapping motifs
DAL82	15	GATAAG	TGACTCGA	1.35422e-05	CCTGCCGTTA	3.7337e-12	NO	weak, not overlapping motifs
ECM22	4	TCGTATA	CCCACGG	3.77335e-06	TCGTATAAG	1.12228e-13	NO	weak *S.cerevisiae* motifs but medium motif from orthologs
FKH1	97	GTAAACAA	GTAAACA	1.37214e-18	GTAAACA	9.11558e-48	YES	strong conserved motif
FKH2	136	GTAAACA	GTAAACA	2.15905e-16	GTAAACA	5.25653e-46	YES	strong conserved motif
FLO8	5	TTTGCN97GCAAA	AGGCATCAT	2.15306e-05	ATCACCACCA	1.69306e-21	NO	very weak *S. cerevisiae* motif
FZF1	22	CGTATCGTATAAGGCAACAATAG	AGTACAAG	1.15037e-07	TTATCATCAG	1.79077e-10	NO	not very strong motifs, not overlapping
GAL4	62	CGGN11CCG	GGAAATGTAA	1.51717e-09	*NA*	*NA*	YES	medium-strong motif with weak overall with known motif
GAT1	19	AGATAAG	GCTTATC	1.09203e-07	CTTATC	1.16657e-15	YES	medium-strong overlapping motifs
GAT3	59	*NA*	TACTTCGAAGC	1.64527e-26	TACTTCGAAGC	2.10251e-31	YES	very strong conserved motif
GCN4	291	TGASTCA	TGACTCA	1.06791e-30	TGACTC	4.18126e-66	YES	very strong overlapping motif

Continued on Next Page...

Table 1 – Continued

TF	Nr	Known BS	S.cerevisiae	p-value	Orthologs	Prob	Cons.	Comments
GCR1	128	GGCTTCCWC	ATATACGGT	1.64991e-18	ATACGGTGTT	6.27441e-19	YES	conserved strong motif, not overlapping
GCR2	82	*NA*	CCGGGCA	6.18481e-06	ATTGAGCCAATC	1.49211e-13	NO	weak non-conserved motifs
GIS1	24	TWAGGGAT	TAAATAGGA	3.00845e-06	TAAGGG	1.63583e-11	NO	weak non-conserved motif, orthologous motif overlapping with known motif
GLN3	56	GATAAGA	CTTATC	2.0965e-14	CTTATC	2.10289e-32	YES	strong conseerved overlapping motif
GTS1	14	*NA*	GTTGATGC	3.13764e-06	*NA*	*NA*	*NA*	weak motif found
GZF3	6	GATAAG	GTAATCAG	4.56157e-08	TCGCTTATC	4.72237e-17	NO	the orthologous motif overlaps with the known motif
HAA1	10	*NA*	ATAACTAAT	1.65476e-05	TTAATATTTAATT	1.76094e-13	NO	weak non-conserved motif
HAP1	56	CGGNNNTANCGG	GCCCC	7.42444e-08	TTGGTTGGTGG	1.54924e-12	NO	non-conserved non-overlapping motifs
HAP2	127	CCAAT	CCAATCA	8.26735e-12	CCAATCA	1.8314e-34	YES	strong conserved overlapping motif
HAP3	127	CCAAT	CCAATCA	8.26735e-12	CCAATCA	1.8314e-34	YES	strong conserved overlapping motif
HAP4	136	GNCCAATCA	CCAATCA	3.74938e-11	CCAATCA	3.3177e-33	YES	strong conserved overlapping motif
HAP5	126	CCAAT	CCAATCA	6.92853e-12	CCAATCA	1.22063e-34	YES	strong conserved overlapping motif
HCM1	244	WAAYAAACAAW	GTACGCCAAA	5.36295e-06	*NA*	*NA*	*NA*	weak non-overlapping motif

Continued on Next Page...

Table 1 – Continued

TF	Nr	Known BS	S.cerevisiae	p-value	Orthologs	Prob	Cons.	Comments
HMS1	10	*NA*	GGGCTC	5.5389e-05	GATGGTACAT	2.09612e-12	NO	weak non-overlapping motif
HOT1	11	*NA*	CTTGATCTG	6.50105e-06	AATAATAAGAAG	4.07536e-10	NO	weak non-overlapping motif
HPC2	1	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found
HSF1	114	cTTCtaGAAgcTTCtaGAAg	TTCTAGAA	2.85502e-27	TTCTAGAA	1.97021e-64	YES	very strong conserved and overlapping motif
IFH1	194	*NA*	TCCGTACA	2.41217e-27	TACTAAC	2.35358e-50	NO	very strong *S.cerevisiae* motif
IME1	9	TTTTCHHCG	TTTCCCTTTTC	1.34873e-06	TTAGCCGC	4.7683e-15	NO	medium strong, partly overlapping motif
INO2	76	CACATGC	TTCACATG	9.42167e-11	TTCACATG	3.67712e-32	YES	strong conserved overlapping motif
INO4	76	CATGTGAA	TTCACATG	1.32319e-13	TTCACATG	1.57806e-37	YES	strong conserved overlapping motif
IXR1	2	*NA*	AACGATAAG	7.18972e-06	TTTCTTCAAGA	1.27713e-12	NO	weak non-conserved motif
KAR4	2	*NA*	AGAGCTATT	8.87619e-06	TTGTTTACCATTT	7.30219e-11	NO	weak non-conserved motif
LEU3	155	CCGGTACCGG	TGACTCA	8.85301e-18	TGACTCA	2.37607e-52	YES	very strong conserved non-overlapping motif
LYS14	6	TCCRNYGGA	AATTCCG	2.67637e-08	AAATTCCG	4.07014e-20	YES	strong, partly overlapping motif
MAC1	39	TTTGCTC	AACAAGACA	2.85045e-06	GGCAGTTTCC	1.52828e-11	NO	weak, non-overlapping motif
MAL13	4	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found
MAL63	3	CGGN9CGG	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found

Continued on Next Page...

Table 1 – Continued

TF	Nr	Known BS	S.cerevisiae	p-value	Orthologs	Prob	Cons.	Comments
MBP1	194	ACGCGT	ACGCG	1.42434e-27	ACGCG	1.6667e-67	YES	very strong conserved and overlapping motif
MCM1	130	CCTAATTAGG	TTTCCTAA	5.24175e-09	ATTAGGAA	1.14065e-23	YES	strong conserved overlapping motif
MET28	34	*NA*	CACGTGA	1.32878e-16	CCACAGTT	2.75285e-43	NO	strong motifs
MET31	34	AAACTGTGG	CACGTGA	8.70367e-14	CACGTGA	4.12471e-27	YES	strong conserved non-overlapping motif
MET32	8	AAACTGTGG	CACGTG	5.58093e-08	AACTGTGGC	5.88414e-15	NO	strong non-conserved overlapping motifs
MET4	36	TCACGTG	CACGTG	4.12474e-17	CCACAGTT	3.36324e-42	NO	very strong motifs
MGA2	3	*NA*	TGTTGAGC	4.38491e-07	TGTTGAGC	1.32824e-23	YES	strong conserved motifs
MIG1	50	ATTTTTGTGGGG	CCCCAC	1.58549e-06	AGTTTTTCCG	6.12217e-13	NO	weak, partly overlapping motifs
MIG2	22	ATTTTTGTGGGG	GGGGAAGG	1.32309e-06	GTTTTTCCG	3.30121e-17	NO	weak, partly overlapping motifs
MIG3	6	AATTRTCCGGGG	CTAACCAG	7.982e-07	TCCGGGGAAA	3.54364e-17	NO	medium-strong, partly overlapping motifs
MOT3	34	HAGGYA	CGTGCCTT	1.16731e-08	AGACTGACA	3.05372e-11	NO	strong, overlapping *S.cerevisiae* motif, non-conserved
MSN1	14	*NA*	AACGTAT	1.20969e-05	TCTCCCACATCACCA	5.06786e-16	NO	weak, non-conserved motifs
MSN2	267	MAGGGGSGG	AGGGG	1.60505e-19	CCCCCT	1.26991e-34	YES	very strong conserved, overlapping motifs

Continued on Next Page...

Table 1 – Continued

TF	Nr	Known BS	S.cerevisiae	p-value	Orthologs	Prob	Cons.	Comments
MSN4	264	AAGGGG	AGGGG	4.2888e-17	CCCCCT	9.27985e-34	YES	very strong conserved, overlapping motifs
MSS11	23	TTTGCN97GCAAA	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found
NDT80	17	GNCRCAAAW	GACACAAA	9.01781e-07	CACAAAA	1.9436e-15	YES	medium-strong conserved, overlapping motif
NRG1	222	GGACCCT	CCGCGGA	1.61155e-08	*NA*	*NA*	*NA*	medium-strong non-overlapping motif
NRG2	144	*NA*	CGCTCGGA	2.17966e-06	*NA*	*NA*	*NA*	weak motif
OAF1	26	CGGN3TNRN8,12CCG	AACTCCG	1.43239e-05	TAAACTCCGA	1.07697e-13	YES	weak, conserved, partly overlapping motifs
OPI1	23	TCGAAYC	CATGTGA	3.29699e-11	TTCACATG	4.89146e-21	YES	conserved non-overlapping motifs
OTU1	1	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	no motifs found
PDC2	2	*NA*	GCTCTTTCA	3.03788e-05	ATTTTTGCAT	4.55004e-15	NO	weak non-conserved motifs