MIPS DNA FunKey micro-array mRNA

REDUCE GEO

Chapter 6

T-base

PAC Gene Ontology gene groups rRPE

General Discussion

consensus motif *Saccharomyces cerevisiae*

PAGE ACO1 GSEA T-profiler Hac1

Transcriptomics MIAMI ChIP-chip Genome

General Discussion

After the completion of the first genome sequence, which was from the bacteriophage phi X-174, in 1977, more than 1800 genomes have been sequenced with two important milestones; first, the completion of the first eukaryotic genome (*Saccharomyces cerevisiae*) and, then, in 2001 the completion of the human genome. Data from the genome sequencing projects have now been used for high throughput methods like transcriptomics, proteomics and metabolomics. These methods, also called functional genomics, have changed molecular biology from a relatively data-poor to a data-rich field. Especially, the use of the micro-array technique that enables researchers to simultaneously measure the abundance of all transcripts has grown exponentially. For example, the Gene Expression Omnibus (GEO)[152], a repository for microarray data that started in 2000 now contains 168,222 single microarray data sets. Before the genomics revolution, biologists were mainly analyzing the individual components or aspects of an organism. Nowadays, researchers are able to focus on the systematic study of the complex interactions in biological systems. A new field of study called systems biology has emerged. Systems biology can be defined as the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions that give rise to life (www.systemsbiology.org). The scope of this thesis is the development, implementation and use of gene expression profile (and fitness profile) data analysis tools, which can be regarded as systems biology methods.

T-profiler; analysis of gene expression profiles using pre-defined gene groups

At the start of this thesis, microarray data were mainly analyzed using clustering [75] algorithms and only a limited number of data analysis tools were available. Since then, the number of data analysis tools has, in parallel with the number of datasets, grown exponentially. The majority of these tools measure enrichment of a selection of genes from a gene expression profile in a pathway or a functional group. To this end, statistical methods like the hypergeometric distribution or Fischer's exact test are applied. Onto-express [75] and Pathway processor [78] were the first tools that used this method for the analysis of gene expression profiles; since then dozens of similar tools were developed (www.geneontology.org).

A disadvantage of such tools is that significantly up- or down-regulated genes have to be selected. Since microarray experiments are relatively expensive, the number of experimental duplicates is limited. In yeast expression data an arbitrary up- or down regulation of 2-fold for the selection of genes with an altered expression level has often been used. However, the remainder of these expression profiles could still contain useful information and therefore methods have been developed that do not use prior selection of genes but instead analyze the whole gene expression profile.

Pavlidis *et al*. [102] was one of the first who used class scores for gene expression analysis on whole expression profiles; later, Quontology [81] the forerunner of T-profiler, followed. Another tool is named PAGE (Parametric Analysis of Gene set Enrichment)[82] that, like Quontology, calculates the significance of the difference between the average expression in a gene set and the genome mean. Although the statistics of T-profiler and PAGE are similar, the t-statistic performs better when using large gene groups. A disadvantage of both PAGE and T-profiler is that they assume that the datasets are normally distributed, which is not always the case. Statistical tests like hypergeometric distribution and Fischer's exact test are non-parametric and therefore make no assumptions about the distribution of data. One of the currently most used pathway analysis tools is Gene Set Enrichment Analysis (GSEA), which is developed at the Broad Institute. GSEA uses the non-parametric Kolmogornov-Smirnoff (KS) test to identify significant changes of predefined gene sets. However, since non-parametric tests use ranks instead of measured values, they tend to be less powerful and flexible than corresponding parametric tests [82]. An advantage of PAGE and T-profiler above GSEA is that the t-value or z-value of gene groups can be used to compare multiple gene expression profiles. Such flexibility is for example demonstrated in Chapter 4 where we discuss the development of T-base, a database in which we compare multiple gene expression profiles analyzed by T-profiler. An obvious disadvantage of all gene group-based analysis methods is the limitation of predefined gene groups. Although T-profiler combines Gene Ontology, MIPS, motif and ChIP-chip based datasets, novel transcriptional responses that are not described in one of the gene sets are missed in the analysis.

Large-scale analysis of gene expression profiles using T-profiler

Since the introduction of the microarray technique in 1995 [22] its use has become more and more common. In many of these studies the analysis of the microarray data was far from complete. Fortunately, most research groups make their data available in public repositories for microarray data like GEO [207] or ArrayExpress [92]. Bioinformaticists are then able to use these datasets for re-analysis and meta-analysis. Standard clustering methods have limited use in the analysis of large-scale expression data, mainly owing to their assignment of a gene to a single cluster [208].

One of the first methods that approached this problem was developed by Ihmels *et al*. [208]. Applying their method on a large-scale expression data set revealed modular organization of the transcriptional network of *Saccharomyces cerevisiae*. Other methods used integration of TF ChIP-chip data and large-scale gene expression data to identify the true target genes of TFs [89, 180]. Zhang *et al*. [209] applied network analysis that combined protein-protein interaction, ChIP-chip, gene expression, fitness and sequence homology data to decompose an integrated yeast interaction network into modules [68]. Another approach called SANDY (Statistical Analysis of Network DYnamics) extended this analysis to reflect the dynamic properties of the transcriptional network under various environmental conditions. Finally, Tanay *et al*. [210] used a biclustering algorithm to extract biological modules from large-scale heterogeneous omics data, varying from gene expression, protein-protein interaction, ChIP-chip and fitness profiling data.

Although these methods perform well in integrating all kinds of –omics data, it is important to realize that the physiological meaning of the modules obtained in this way may differ. For example, it is important to make a distinction between the modules obtained by using transcript profiling data and the modules obtained from fitness profiling data as there often is no obvious relationship between the response of transcriptional modules to stress and the fitness modules identified under the same stress conditions. We initially analyzed large-scale expression data using T-profiler to make a distinction between general and specific transcriptional responses. To this end we created T-base (Chapter 4). Basically, T-base could be considered as a gene expression and fitness profile interpretation database. In this thesis we present three more specific applications of T-base.

(1) Condition-specific activation of gene sets

In chapter four we demonstrate that gene groups can be queried for their specific activation. As an example we showed that the specific activation of the Hac1p- (TFO) gene group occurs specifically in gene expression conditions resulting in the accumulation of defective protein in the ER. Thus, the experimental conditions provide information about the physiological role of the activated gene group; in this case, most of them are expression profiles of cells with partially repressed genes that function in the ER. The context of such defined conditions also gives information about similar expression profiles obtained using less well defined conditions (such as uncharacterized genes or treatments with compounds that have an unknown mode of action). Thus, T-base may also be queried to gain insight into the mode of action of compounds via their expression profiles. A similar method was applied by Lamb *et al*. [70], who created a reference collection of gene-expression profiles from cultured human cells treated with bioactive small molecules together with pattern-matching software to mine these data. This ‘‘Connectivity Map’’ resource can be used to find connections between expression signatures of mode of action of compounds, physiological processes, or diseases.

Our method also enables to separate specific and general responses. T-base revealed that the STRE (AGGGG/CCCCT) gene group is active in almost 50% of the expression profiles and therefore can be regarded as a general response. Zakrzewska *et al*. (Thesis 2007, chapter 4) showed that this approach is not limited to gene expression profiles only. She applied it to a set of fitness profiling experiments and found that the GO-term of vesicle-mediated transport is active in the majority of the fitness profiles and therefore could be regarded as a general fitness feature.

(2) Co-modulation network of Transcription Factor activity

Secondly, we present a novel approach that we used to build a co-modulation network. This approach is based on the assumption that the t-value of the TFO (transcription factor occupancy) gene group might be considered as a proxy for the activity of a TF. Other approaches make networks that are based on the correlation of individual genes. In Chapter 4 we built such a network by using correlation analysis of the activation profiles of the TFO gene groups. Our method differs in two important ways from that of Luscombe *et al*. [175]. First, their network is created based on the relation between TF activity (measured by the differential transcription of a TF) and their gene targets, whereas our network is based solely on the gene targets of a TF. Since the activity of many transcription factors is regulated on the posttranscriptional level rather than on the level of transcription, our approach seems more generally useful. Secondly, Luscombe *et al*. (arbitrarily) separate their experiments in five different classes based on the conditions used whereas our approach does not need this.

(3) Prediction of gene functions

An important goal after sequencing the genome of *S. cerevisiae* is to complete the functional description of all yeast genes [6]. In Chapter 5 we used the information from T-base in combination with correlation analysis of the modulation of individual genes to make predictions of poorly annotated genes. The results of this method are presented in a web-application named FunKey. In this analysis we combine gene expression and fitness profile datasets but also present the predictions based on the separate data. This combination allows making prediction of genes that are not transcriptionally regulated but show a clear fitness effect. For example, genes that are involved in DNA repair show clear fitness effects but are poorly regulated on a transcriptional level [211]. Secondly, genes that show different relationships in genomic transcription and fitness experiments could be multifunctional genes. The TCA cycle gene *ACO1*, shows for example high correlations to the TCA cycle GO-term based on gene expression data but the GO-term of mitochondrial maintenance gives the highest score based on fitness profiles. Just recently it has been shown that *ACO1* indeed is multifunctional and participates in both functions [206].

T-profiler, T-base, FunKey and future perspectives

Gene expression profiles can be considered as a snapshot of the physiological state of a cell under a certain condition at the moment of sampling. The interpretation of this snapshot, the list of genes that are up- or down regulated, sometimes reflects a Rorschach figure; dependent on the researcher, genes that are useful to explain a certain hypothesis are used while others that are not are neglected. Bioinformatic approaches like T-profiler can help researchers to interpret their data in an objective way.

T-profiler analysis has already been applied in several studies: Zakrzewska *et al*. [69] used it to analyze the transcriptional response to the plasma membrane perturbing compound chitosan. T-profiler analysis predicted activation of the cell wall integrity pathway, the calcineurin pathway and a Cin5p-mediated response. All these predictions could be validated using biological assays. A similar experimental set-up was used to measure the response to chitosan by fitness-profiling analysis [65]. A comparison between this study and the one based on gene expression profiling revealed a poor comparison on the gene level but a much better comparison on the gene group level. Furthermore, T-profiler analysis was used to study the transcriptional effects of strains mutated in the glucose regulatory network [212], the transcriptional response to a heat shift (30°C to 39°C) in a time series experiment (Mensonidis, Thesis chapter 5; manuscript in preparation) and the transcriptional response to sorbic acid (Resende *et al*., unpublished results). In collaboration with Unilever, T-profiler was also adapted to work with gene expression profiles from *Bacillus subtilis*. Ter Beek *et al*. (submitted) used this T-profiler version to study the transcriptional response to sorbate in *B. subtilis*.

T-profiler is also used within the Netherlands Toxicogenomics Centre (NTC) at TNO and the

University of Maastricht to study the effect of toxic compounds in mouse, rat and humans both *in vivo* and *in vitro* (Kuper *et al*., submitted; van Leeuwen *et al*., submitted). Interestingly, T-profiler can also be used to make an interspecies comparison between the transcriptional response to acetaminophen (paracetamol) in rat livers, *in vivo* and *in vitro* and in human *in vitro* liver cells on the pathway level (Kienhuis *et al*., unpublished).

Compared to other bioinformatics tools the ones presented in this thesis are conceptually simple, powerful and easy to implement. As mentioned before, T-profiler is already used for the analysis of rat, mouse and human transcription data and the Bussemaker lab plans to extend this to every organism with GO annotations. For the moment, FunKey is only developed for *S. cerevisiae* genes but, if enough transcription data become available, the method can also be applied to other organisms. Finally, it would be highly useful to extend T-base with data from other organisms. If meta-data of such data sets would be given proper descriptions like suggested by MIAME (Minimal Information of A Microarray Experiment), this would also allow coupling of such a database with other data sources. Together, this would create a higly relevant gene group based interpretation database.