Dissection of transcriptional regulation networks and prediction of gene functions in Saccharomyces cerevisiae

Boorsma, A.

Summary

Molecular biology aims to unravel the functions of cells by studying cellular processes at the molecular level. A model organism that is well established in molecular biology is bakers yeast (*Saccharomyces cerevisiae*). Bakers yeast cells are remarkably similar to human cells, but much easier to grow and manipulation of its DNA is straight-forward. In 1996, the complete DNA sequence of the yeast genome has been deciphered, revealing that the whole genome contains 12 million basepairs and that the estimated amount of genes is around 6000. In comparison, the recently sequenced human genome contains 3 billion basepairs and the number of genes is estimated to be between 20.000 and 25.000.

To translate genes into functional proteins, the gene-DNA is first copied to messenger RNA (mRNA) during a process called transcription, and subsequently the mRNA is translated to proteins. One of the important questions of the molecular biology is how transcription is organized. It is already known that during this process a very important role is played by transcription factors, which are proteins that bind small specific stretches of DNA (called motifs) to enable transcription of genes. The scope of this thesis is the regulation of transcription; when are which genes transcribed to mRNA and which transcription factors are involved?

An important new technique that revolutionized the study of transcription regulation is microarray analysis. With this technique it is possible to measure transcription of all genes of a certain cell type in a single experiment. Microarray analysis generates large amounts of data that are processed using informatics and that are analyzed by statistical methods. As a result, a new area of biology has emerged, named bioinformatics. This thesis describes the development of several (bioinformatic) methods that help analyze and interpret microarray data.

Although the technique has improved dramatically, there are still some problems associated with microarray data, making it difficult to analyze them. First of all the data are noisy, and since the technique is expensive, it is not possible to repeat experiments many times to reduce the noise. Secondly, microarray experiments are difficult to reproduce; results from identical experiments performed on different array platforms are often not the same. Finally, methods that allow a biological interpretation of microarray data are lacking. At the begin of this study, the method of choice was cluster analysis, for which data from multiple experiments where needed

In chapter two of this thesis we present T-profiler, a microarray analysis method that we developed to address these problems. The idea of T-profiler is not to focus on the transcription of individual genes but instead to look at groups of genes with a common feature. This might be genes that are bound by the same transcription factor, or groups of genes with a similar biological function. A major advantage of measuring transcription of groups of genes is reduction of the influence of noise. In addition, the common feature of the gene groups also provides information about the effect of the experimental condition, for example, which transcription factor or which functional group is active. T-profiler is available through a web application (www.t-profiler.org).

In chapter three we use the microarray technique to measure the transcriptional response of yeast to compounds that cause cell wall stress. Analysis of the data revealed that besides a general stress response, a specific response is triggered. This specific response is regulated by the transcription factor Rlm1 that is known to mainly regulate cell wall related genes. In addition we used our analysis method to compare these data to that of publicly available microarray data of two mutants that constitutively activate the cell wall stress response. Not

unexpectedly, the analysis profiles of these datasets were highly comparable. Surprisingly, we found activation of the transcription factor Sko1p, that is known to be involved in the response of osmo-stress.

In chapter four we take the comparison of public available microarray data a step further by comparing about 1000 different microarray experiments. First we used T-profiler to calculate the activity of the different transcription factors in these studies and then we used this information for correlation analysis. The final results provide several new insights into the basic process of transcription in bakers yeast. For example, we show a strong negative correlation between the so-called PAC and rRPE motifs, and transcription factors of the general stress response. So far, no transcription factors have been assigned yet to the PAC and rRPE motifs, and we hypothesize that they are part of a special class of motifs, the so-called core-promoter elements. Furthermore we used our correlation matrix to built a network of transcription factor activities. We used this network to predict new functions for some transcription factors.

The focus in chapter four is on the activity of transcription factors. We performed T-profiler analysis on the same microarray dataset, using gene groups based on similar biological functions. This information is used in chapter five to make predictions about the biological function of uncharacterized genes. The method has been validated by testing the reliability of predictions on well-characterized genes. A special website has been developed (www.science.uva.nl/~boorsma/funkey) that can be used to generate functional predictions.

The examples from chapter four and five demonstrate the power of new bioinformatic techniques such as T-profiler; it is now possible to compare different datasets and to make functional predictions that were not possible by studying only individual genes and proteins. The techniques that were developed and described in this thesis are now also being used for mouse, rat and human microarray data.

Samenvatting

Moleculaire biologie is het vakgebied waarin de processen in cellen op moleculair niveau worden bestudeerd met als doel de werking van de cellen te doorgronden. Een van de model organismen die hiervoor wordt gebruikt is de eencellige bakkersgist (*Saccharomyces cerevisae*). Bakkersgistcellen lijken qua opmaak sterk op de cellen van de mens, maar zijn veel gemakkelijker te kweken en het DNA is relatief eenvoudig te manipuleren. In 1996 is de basenpaar volgorde van het gist genoom (de complete set chromosomen) ontcijferd; het genoom bestaat uit 12 miljoen basenparen en het aantal genen word geschat op 6000. Ter vergelijking; het genoom van de mens telt ongeveer 3 miljard basenparen en het aantal genen ligt naar schatting tussen de 20.000 en 25.000.

Om genen te vertalen in werkzame eiwitten wordt het gen-DNA eerst gekopieerd in zogenaamd boodschapper RNA, een proces dat transcriptie wordt genoemd, waarna dit boodschapper RNA (of messenger RNA; mRNA) uiteindelijk eiwitten produceert. Een van de belangrijkste vraagstukken in de moleculaire biologie is hoe de aanmaak van eiwitten precies georganiseerd wordt. We weten inmiddels dat hierbij een zeer belangrijke rol is weggelegd voor transcriptie factoren. Dit zijn eiwitten die aan kleine specifieke stukjes DNA (zogenaamde motieven) binden om zo de transcriptie van genen op gang te helpen. Dit proefschrift gaat voor een groot deel over de regulatie van transcriptie; wanneer worden welke genen gekopieerd van DNA naar mRNA en en vooral welke transcriptie factoren zijn hierbij betrokken.

Een belangrijke nieuwe techniek om de regulatie van transcriptie te bestuderen is de microarray analyse. Met behulp van de microarray kan in één experiment de transcriptie van alle genen in een bepaald celtype tegelijkertijd worden gemeten. De microarray techniek genereert grote hoeveelheden data die door middel van de informatica wordt verwerkt en met behulp van statistische methodes worden bestudeerd. Dit heeft een nieuw veld in de biologie opgeleverd: de bioinformatica. In dit proefschrift wordt een aantal (bioinformatica) methodes beschreven die helpen microarray data te analyseren en interpreteren. Hoewel de techniek tegenwoordig sterk is verbeterd, kleven er aan microarray analyse een aantal nadelen die de analyse bemoeilijken. Ten eerste is de ruis in microarray data relatief hoog en is de methode te duur om experimenten vaak te herhalen om ruis te reduceren. Ten tweede zijn experimenten vaak lastig te reproduceren; de resultaten van identieke experimenten uitgevoerd op verschillende microarray systemen en in verschillende laboratoria zijn vaak heel verschillend. Ten slotte ontbrak het in het begin van deze promotiestudie aan methodes die nieuwe biologische informatie uit de data kon halen. Er werd destijds vaak gebruik gemaakt van cluster analyse die een serie van experimenten als input nodig heeft.

T-profiler, de analyse methode die in hoofdstuk twee van dit proefschrift wordt gepresenteerd, is door ons ontwikkeld om deze problemen aan te pakken. De gedachte was om niet naar de transcriptie van individuele genen te kijken maar naar groepen genen die een bepaalde samenhang hebben. Genen die bijvoorbeeld door dezelfde transcriptie factor gebonden worden, of die eenzelfde biologische functie hebben. Een groot voordeel van het meten van de transcriptie van groepen van genen is dat de invloed van ruis vermindert. Een ander voordeel van deze aanpak is dat de samenhang van de groepen genen informatie geeft over de experimentele omstandigheid, zoals bijvoorbeeld welke transcriptie factor of functionele groep actief is. De data wordt daardoor beter interpreteerbaar. T-profiler is beschikbaar gemaakt via een webapplicatie (www.t-profiler.org).

In hoofdstuk drie meten we met behulp van de microarray techniek de transcriptionele respons

van gist op stoffen die celwand stress veroorzaken. Uit de analyse blijkt dat de stoffen naast een algemene stress respons ook een specifieke stress respons veroorzaken. Deze specifieke respons wordt gereguleerd door de transcriptie factor Rlm1 die vooral celwand genen tot expressie brengt. Vervolgens konden we deze data met behulp van onze analyse methode vergelijken met eerder gepubliceerde microarray data van twee mutanten die de celwand stress respons permanent activeren. Zoals verwacht zijn de analyse profielen van deze data sterk vergelijkbaar met die van de stoffen die celwand stress veroorzaken. Een verrassende vinding is echter de activatie van de transcriptie factor Sko1p die normaal betrokken is bij de respons op osmolariteits stress.

In hoofdstuk vier gaan we een stap verder met het vergelijken van gegevens van andere datasets door ongeveer 1000 verschillende microarray experimenten met elkaar te vergelijken. Daarvoor gebruiken we eerst T-profiler om de activiteit van de verschillende transcriptie factoren te berekenen waarna we correlatie analyse toepassen op deze data. Dit levert vooral inzichten op over het basale transcriptie proces van bakkersgist. We laten bijvoorbeeld zien dat transcriptie factoren die betrokken zijn bij de algemene stress respons een sterke omgekeerde correlatie hebben met groepen genen die gereguleerd worden via de zogenaamde PAC en rRPE motieven. Van deze motieven is overigens niet bekend door welke transcriptie factoren ze worden gebonden. Onze analyse vormt de basis voor de hypothese dat deze motieven zogenaamde core-promotoren zijn. Verder hebben we de correlatie gegevens gebruikt om een netwerk van de activiteit van transcriptie factoren te bouwen. Op basis van dit netwerk voorspellen we een nieuwe rol voor een aantal transcriptie factoren.

De nadruk in hoofdstuk vier ligt voornamelijk op de activiteit van transcriptie factoren. Dezelfde 1000 microarray experimenten zijn echter ook door T-profiler geanalyseerd met behulp van groepen genen met een zelfde biologische functie. Deze gegevens gebruiken we in hoofdstuk vijf om voorspellingen te doen over de biologische functie van niet gekarakteriseerde genen. De methode is hiervoor eerst gevalideerd door te laten zien dat de voorspelling van de functie van goed gekarakteriseerde genen over het algemeen klopt. De voorspellingen voor alle genen zijn uiteindelijk gemakkelijk te controleren op een speciaal hiervoor gemaakte website (www.science.uva.nl/~boorsma/funkey).

De voorbeelden uit hoofdstuk vier en vijf laten zien hoe krachtig nieuwe bio-informatica technieken zoals T-profiler zijn; het is nu mogelijk om verschillende datasets met elkaar te vergelijken en voorspellingen te doen die ondenkbaar zijn wanneer alleen individuele genen of eiwitten worden bestudeerd. In dit proefschrift hebben we de technieken voornamelijk gebruikt voor gist microarray data, maar momenteel worden ze ook toegepast op muis, rat en humane micrarray data.