KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

COMPUTATIONAL DISCOVERY OF *CIS*-REGULATORY MODULES IN ANIMAL GENOMES

Promotoren:
Prof. dr. ir. B. De Moor
Prof. dr. ir. Y. Moreau

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Stein AERTS

Mei 2004

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

COMPUTATIONAL DISCOVERY OF *CIS*-REGULATORY MODULES IN ANIMAL GENOMES

Jury:
Prof. dr. ir. P. Verbaeten, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. Y. Moreau, co-promotor
Prof. dr. B. De Strooper
Prof. dr. ir. J. Vanderleyden
Prof. dr. ir. S. Vanhuffel
Prof. dr. ir. D. Roose
Prof. dr. ir. J. van Helden (ULB)
Prof. P. Rouzé (INRA, VIB, U.Gent)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Stein AERTS

U.D.C. 681.3*F11:575.113

Mei 2004

© Katholieke Universiteit Leuven – Faculteit Toegepaste Wetenschappen
Arenbergkasteel, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2004/7515/29

ISBN 90-5682-491-0

Voorwoord

Op dit moment van schrijven is mijn vrouw ongeveer 30 dagen zwanger en het embryo heeft zopas zijn/haar neurale buis gevormd. Enkele weken geleden, toen het nog een blastocyst was, zouden u en ik geen verschil kunnen merken tussen dit embryo en een embryo van, zeg een eekhoorn. Hiermee wil ik ons embryo nu al niet een identiteitscrisis bezorgen, ik wil er enkel het volgende mee zeggen: Wij dieren bestaan allemaal uit zowat dezelfde bouwstenen, en toch lijkt de mens in een volgroeid stadium helemaal niet op een eekhoorntje (voor een gist wel natuurlijk, doch dat geheel ter zijde). De clue is dat wij die bouwstenen op een lichtjes andere manier gaan aanwenden. Het "ontwikkelingsprogramma" zit, net als de bouwstenen zelf, gecodeerd in ons DNA. Het bestaat uit schakelaars die onder gepaste omstandigheden keurig de juiste "genen" moeten aanschakelen die dan de bouwstenen (lees eiwitten) leveren. De plaats van levering in het embryo, de hoeveelheid tegelijk geleverde bouwstenen, en de duur van levering bepalen eenvoudig gesteld of we een eekhoorn of een mens worden. Hoewel dit vanuit een naturalistisch standpunt verder van geen betekenis is, betrapte ik mezelf toch op een licht obsessieve drang om dat programma, en vooral die schakelaars beter te begrijpen. Een beetje zelfkennis was al voldoende om in te zien dat een chemisch-biologische strategie —die overigens in mijn ogen bijzonder efficient is— niet aan mij was besteed. Toen in februari 2001 de volledige menselijke DNA sequentie werd geopenbaard, met daarin 3 miljard "letters" die een zo goed als onleesbare tekst vormden, dienden zich plots mogelijkheden aan om in die letterzee naar onze schakelaars te zoeken, en wel zonder pipetten en proefbuizen, maar met een door mij meer geliefkoosd medium, de computer en het internet. Toen tegelijkertijd de "DNA-chip" technologie wijd verbreid begon te worden werd het ook mogelijk om, laat ons zeggen, alle aanwezige bouwstenen op bepaalde plaats en moment in ons lichaam te inventariseren. Met deze werktuigen voorhanden kon men wereldwijd plots heel wat vooruitgang boeken om de ingewikkelde puzzel van "genregulatie" op te lossen, en ik mocht meedoen.

De mensen (lees experts) die mij de complexe spelregels hebben uitgelegd om aan deze puzzel te werken, ben ik veel dank verschuldigd. Het bleef trouwens niet bij de inwijding, er ontstond al gauw een boeiende samenwerking en velen hebben mij enorm geholpen om aan wetenschap te leren doen. Dit werk mag dan ook gezien worden als het resultaat van gezamenlijk werk, en het is een hele eer voor mij dat ik het mag samenvatten en voorstellen in dit boek.

Vooreerst wil ik mijn promotor, Bart De Moor bedanken om mij de kans, de steun en de financiële middelen te geven om aan dit doctoraat te werken. I thank all the members of the jury, and the reading committee in particular, for their reading effort and valuable feedback. Yves Moreau wil ik bedanken voor de nu al jarenlange boeiende samenwerking, zowel bij Data4s als ESAT. The Polish scientist Mike Dabrowski deserves special thanks for the everlasting confidence in computational biology. The great discussions we had on gene regulation (remember the transcription factories in Liege) always boosted my work. Bedankt Bart De Strooper om open te staan voor de samenwerking met ESAT en om Mike en mij te steunen in ons onderzoek. Ook Hannelore Denys wil ik bedanken voor de samenwerking, het is dankzij het werk met TCF-3 dat ik ingeleid werd in de DNA-sequentie-analyse. I further wish to thank Bassem Hassan for the discussions on gene regulation in fly neurodevelopment. Veel dank gaat uit naar iedereen van BioI@SCD voor de samenwerkingen en de sfeer: Gert, Bert, Patje, Geert, Kathleen, Kristof, Frank, Joke en de hele groep. Ook wil ik Peter Van Loo hartelijk bedanken voor de intensieve samenwerking en de significante bijdrage in dit project tijdens zijn ingenieursthesis. Merci allemaal! En hopelijk volgen er nog veel gelijkaardige samenwerkingen! Op persoonlijk vlak dank ik tenslotte mijn familie, mijn vrienden en in het bijzonder mijn vrouw Nele voor de vlotte chromosomenmengeling en voor alle liefdevolle steun in dit werk. Veel plezier met lezen, en aarzel niet mij te contacteren indien u vragen zou hebben over de inhoud. Stein.

Abstract

The transcriptional regulation of metazoan genes is governed by combinations of transcription factor binding sites in *cis*-regulatory modules. Their central role in gene regulatory networks makes their detection and characterization of great importance for the understanding of the genetic programs encoded in the genome. The availability of complete genome sequences of several metazoan species and of high-throughput expression profiling using DNA microarrays is exploited in the bioinformatics methods described here to detect sets of co-expressed genes on the one hand, and the transcription factor binding sites that govern this co-expression on the other hand. For the former, a case study of gene expression profiling during *in vitro* neuronal differentiation in mice is described. The microarray data are preprocessed, clustered, and functionally analyzed using Gene Ontology associations. The expression data is further compared with expression data from *in vivo* differentiation. A high correlation between the systems was found after mapping the time points of the two data sets by time warping. For the detection of transcription factor binding sites, new algorithms are presented to predict significant occurrences and combinations thereof as *cis*-regulatory modules. The methods combine the statistical over-representation of instances of known motif matrices in gene batteries with evolutionary sequence conservation. Their performance is tested either on artificial data sets, on benchmark data sets, or on proprietary data sets. For module finding, a branch-and-bound and a genetic algorithm are implemented to find the optimal combination of binding sites in a set of co-expressed genes. Genomic searches for such newly found modules then yield putative target genes, for which the functional coherence is measured to give an indication of the validity of the module. The putative target genes are further prioritized computationally by comparing their functional characteristics with the gene battery where the module was found. The methods are integrated into computational analysis strategies using multiple genomic information sources and they are made available as user-friendly software tools. Lastly, a genomic sequence analysis is performed to study the nucleotide composition around the transcription start site in several metazoan species.

Samenvatting

Bij dieren verloopt de transcriptionele regulatie van genen via combinaties van transcriptiefactorbindingsplaatsen in *cis*-regulatorische modules. De centrale rol van dergelijke modules in genregulatorische netwerken maken dat de detectie en de karakterisatie ervan van groot belang zijn voor een beter begrip van de genetische programma's die gecodeerd zijn in ons genoom. De beschikbaarheid van volledige genoomsequenties van verscheidene dierlijke species en van "high throughput" expressieprofilering met DNA microarrays worden aangewend in de beschreven bio-informatica methoden voor de detectie van enerzijds groepen van genen die samen tot expressie komen (genbatterijen), en anderzijds van de transcriptiefactorbindingsplaatsen die deze co-expressie veroorzaken. Betreffende de genbatterijen wordt een casus beschreven van genexpressieprofilering tijdens neuronale differentiatie *in vitro* in muizen. De microarray data worden voorbehandeld, gegroepeerd, en functioneel geanalyseerd gebruik makende van "Gene Ontology" associaties. Een vergelijking van de expressiegegevens met gegevens van neuronale differentiatie *in vivo* toont een hoge correlatie aan tussen beide systemen. Betreffende de detectie van transcriptiefactorbindingsplaatsen worden nieuwe algoritmes voorgesteld om significante voorkomens en combinaties ervan te vinden. De methoden combineren de statistische over-representatie van voorkomens van gekende motiefmatrices met de evolutionaire conservering van de sequenties. De performantie wordt ofwel getest op artificiële datasets, of op "benchmark" datasets, of op zelf ontworpen datasets. Betreffende het vinden van modules werden een "branch-and-bound" en een genetisch algoritme geïmplementeerd om de optimale combinatie van bindingsplaatsen te vinden in een genbatterij. Het zoeken naar voorkomens van op die manier ontdekte modules in het hele genoom levert dan potentiële doelgenen op, en om de geldigheid van de module na te gaan wordt de functionele coherentie van deze doelgenen gemeten. De mogelijke doelgenen worden verder computationeel geprioritiseerd door hun functionele karakteristieken te vergelijken met de genbatterij waar de module werd gevonden. De methoden werden geïntegreerd tot computationele analyse-strategieën gebruik makende van verscheidene genomische informatiebronnen en ze worden beschikbaar gemaakt onder de vorm van gebruiksvriendelijke software programma's. Tenslotte werd een genomische analyse uitgevoerd om de nucleotidesamenstelling te bestuderen rond de transcriptiestartplaats van genen in een aantal dierlijke species.

Notation

Abbreviations

ANOVA	analysis of variance
BLAST	Basic Local Alignment Search Tool
BTA	basal transcription apparatus
CDS	coding sequence
CNS	conserved non-coding sequence
CRE	*cis*-regulatory element
CRM	*cis*-regulatory module
DAG	directed acyclic graph
DNA	deoxy-ribonucleic acid
DPE	downstream promoter element
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
GUI	graphical user interface
HMM	hidden Markov model
IBC	intergenic background composition
IDF	inverse document frequency
ISM	information submodel
IUPAC	International Union for Pure and Applied Chemistry
JWS	Java Web Start
GFF	general feature format
GO	Gene Ontology
GRN	gene regulatory network
GTF	general transcription factor
LRA	logistic regression analysis
MGED	Microarray Gene Expression Data
MGI	Mouse Genome Informatics
MIAME	Minimum Information About a Microarray Experiment
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information (US)
ncRNA	non-coding RNA
PDB	Protein Data Bank

PF	phylogenetic footprinting
PSFM	position specific frequency matrix
PWM	position weight matrix
RMI	Remote Method Invocation
RNA	ribonucleic acid
rRNA	ribosomal RNA
RNAP	RNA polymerase
SOAP	Simple Object Access Protocol
SNF	single nucleotide frequency
SNP	single nucleotide polymorphism
TAF	TBP associated factor
TATA	TATA-box, see glossary
TBP	TATA-binding protein
TCF	transcription co-factor
TF	transcription factor
TFBS	transcription factor binding site
TLS	translation start site
tRNA	transport RNA
TSS	transcription start site
UTR	untranslated region

IUPAC ambiguous DNA characters

These characters are often used in consensus DNA binding sites:

`M`	`A` or `C`
`R`	`A` or `G`
`W`	`A` or `T`
`S`	`C` or `G`
`Y`	`C` or `T`
`K`	`G` or `T`
`B`	`C`, `G` or `T`
`D`	`A`, `G` or `T`
`H`	`A`, `C` or `T`
`V`	`A`, `C` or `G`
`N`	`A`, `C`, `G` or `T`

Gene nomenclature

All gene symbols are italicised and protein symbols are normally the same as the encoding gene symbols but not italicised. Human gene symbols[1] are designated by upper-case Latin letters or by a combination of upper-case letters and Arabic numerals, for example *BRCA1*, *CYP1A2*. To identify human genes we use either HUGO symbols as found in the LocusLink and Ensembl databases or Ensembl gene identifiers (ENS*). Mouse gene symbols[2] begin with an uppercase letter, the rest is normally lowercase, for example *Brca1*, *Cyp1a2*. We use gene identifiers from the Mouse Genome Database (MGD). Lastly, for *Drosophila melanogaster* the genetic nomenclature from FlyBase[3] is used.

---

[1] Guidelines for human gene nomenclature can be found on http://www.gene.ucl.ac.uk/nomenclature/guidelines.html [321].

[2] Guidelines for mouse gene nomenclature can be found on http://www.informatics.jax.org/mgihome/nomen/ [200].

[3] FlyBase URL: http://fly.ebi.ac.uk:7081/docs/nomenclature/lk/nomenclature.html.

Related publications

- **Stein Aerts**, Gert Thijs, Bert Coessens, Mik Staes, Yves Moreau and Bart De Moor (2003) TOUCAN: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Research*, 31(6), 1753-1764.
- Michal Dabrowski*, **Stein Aerts***, Paul Van Hummelen, Katleen Craessaerts, Bart De Moor, Wim Annaert, Yves Moreau, and Bart de Strooper (2003) Gene profiling of hippocampal neuronal culture. *Journal of Neurochemistry*, 85(5), 1279-1288. (* equal contribution)
- **Stein Aerts**, Peter Van Loo, Gert Thijs, Yves Moreau and Bart De Moor (2003) Computational detection of *cis*-regulatory modules. *Bioinformatics*, 19 Suppl. 2, ii5-ii14.
- **Stein Aerts**, Peter Van Loo, Yves Moreau and Bart De Moor (2004) A genetic algorithm for the detection of new *cis*-regulatory modules in sets of coregulated genes. *Bioinformatics, in press.*
- Yves Moreau, **Stein Aerts**, Bart De Moor, Bart De Strooper, and Michal Dabrowski (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends In Genetics*, 19(10), 570-577.
- Bert Coessens, Gert Thijs, **Stein Aerts**, Kathleen Marchal, Frank De Smet, Kristof Engelen, Patrick Glenisson, Yves Moreau, Janick Mathys, and Bart De Moor (2003) INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Research*, 31(13), 3468-3470.
- Hannelore Denys, Ali Jadidizadeh, Saeid Amini Nik, Kim Van Dam, **Stein Aerts**, Benjamin A Alman, Jean-Jacques Cassiman and Sabine Tejpar (2004) Identification of IGFBP-6 as a significantly downregulated gene by beta-catenin in desmoid tumors. *Oncogene*, 23(3), 654-664.
- Kathleen Marchal, Kristof Engelen, Jos De Brabanter, **Stein Aerts**, Bart De Moor, Torik Ayoubi, and Paul Van Hummelen (2002) Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA arrays. *Journal of Biological Systems*, 10(4), 409-430.
- **Stein Aerts**, Gert Thijs, Michal Dabrowski, Yves Moreau, and Bart De Moor (2004) Comprehensive analysis of the base composition around the transcription start site in Metazoa. *Under revision.*

Contents

Voorwoord **i**

Abstract **iii**

Samenvatting **v**

Notation **vii**

Related publications **xi**

Contents **xiii**

1 Context and scope **1**

2 An overview of gene regulation: biology and bioinformatics **7**
- 2.1 Gene regulation and development . . . 7
- 2.2 Gene regulation and evolution . . . 8
- 2.3 Gene regulation and disease . . . 9
- 2.4 Transcriptional regulation in eukaryotes . . . 9
  - 2.4.1 The eukaryotic gene . . . 11
  - 2.4.2 The core promoter . . . 13
  - 2.4.3 Transcription factor binding sites . . . 16
  - 2.4.4 Transcription factors . . . 21
  - 2.4.5 Transcription co-factors . . . 22
  - 2.4.6 *Cis*-regulatory modules . . . 23
  - 2.4.7 Putting it all together . . . 29
- 2.5 Gene batteries . . . 30
  - 2.5.1 Genome-wide expression analysis . . . 31
  - 2.5.2 Detecting common motifs in gene batteries . . . 33
- 2.6 Gene regulatory networks . . . 35
- 2.7 Regulatory evolution . . . 37
  - 2.7.1 Mutations in *cis* . . . 38
  - 2.7.2 Mutations in *trans* . . . 38
  - 2.7.3 Phylogenetic footprinting . . . 38

3 Microarray data analysis: a case study in neurobiology **43**
- 3.1 Introduction . . . 43
- 3.2 Neuronal differentiation *in vitro* . . . 44
- 3.3 The microarray experiment . . . 45
- 3.4 Data preprocessing . . . 46
  - 3.4.1 Normalization . . . 46
  - 3.4.2 Filtering . . . 48

3.5 Clustering . . . 49
3.6 Analysis by gene function: the synaptic vesicle cycle genes . . . 53
3.7 Functional exploration using Gene Ontology . . . 55
3.8 Comparing two microarray data sets . . . 56
3.9 Discussion . . . 58
3.10 Perspectives on *cis*-regulatory sequence analysis . . . 62
3.11 Perspectives on the comparison of microarray data . . . 63
  3.11.1 Data access and exchange . . . 63
  3.11.2 Microarray standards and repositories . . . 64
  3.11.3 Microarray analysis in the era of repositories and compendia 65

4 Detecting transcription factor binding sites in metazoan genes 67
4.1 Introduction . . . 67
4.2 Constructing sets of putative regulatory sequences . . . 68
  4.2.1 Proximal promoters . . . 68
  4.2.2 Distal regulatory sequences . . . 70
4.3 Detection of transcription factor binding sites . . . 71
  4.3.1 PSFM databases . . . 71
  4.3.2 Higher-order background models . . . 72
  4.3.3 MotifLocator . . . 74
  4.3.4 MotifScanner . . . 74
  4.3.5 Discussion . . . 75
4.4 Statistical test for over-representation . . . 77
4.5 TOUCAN . . . 78
4.6 Case studies . . . 81
  4.6.1 E2F target genes . . . 83
  4.6.2 Liver and muscle genes . . . 84
  4.6.3 TCF3-β-catenin target genes . . . 87
  4.6.4 Binding site detection without gene batteries: a case study in neurogenesis . . . 88
4.7 Related work . . . 90
4.8 Conclusions . . . 91

5 Detecting *cis*-regulatory modules 93
5.1 Introduction . . . 93
5.2 Module score function . . . 94
5.3 The A^* search algorithm . . . 95
5.4 Validation of the A^* ModuleSearcher . . . 97
  5.4.1 Semi-artificial sequence sets . . . 97
  5.4.2 Sensitivity to PSFM scoring . . . 98
5.5 ModuleSearcher on real gene batteries . . . 98
  5.5.1 Gene Ontology statistics . . . 102
  5.5.2 Genomic searches . . . 102
  5.5.3 Biological validation of the ModuleSearcher . . . 104
5.6 Genetic Algorithm version of the ModuleSearcher . . . 107
5.7 Availability within Toucan . . . 108
5.8 Discussion . . . 108

6 Data integration for module and target validation 111
6.1 Introduction . . . 111
6.2 Data sources . . . 112
  6.2.1 Vector data . . . 112
  6.2.2 Non-vector data . . . 114
6.3 Order statistics and overall ranking . . . 114
6.4 ENDEAVOUR . . . 115
6.5 Cross validation . . . 116
6.6 Case studies . . . 116

6.7 Conclusions 118
6.7.1 Perspectives 118
6.7.2 Perspectives on the computational prioritization of candidate disease genes 118

7 Comprehensive analysis of the base composition around the transcription start site in Metazoa 121
7.1 Background 121
7.2 Data and methods 122
7.3 Results and discussion 123
7.3.1 Comparing Ensembl and DBTSS human gene start annotations 123
7.3.2 Variations in base composition in different phyla 123
7.3.3 Nucleotide composition and CpG islands 125
7.3.4 Nucleotide composition and gene expression 126
7.3.5 GC and AT skews around the TSS 129
7.4 Conclusions 131

8 General discussion 133

A. Glossary 139

B. Software and databases 141

Nederlandse samenvatting 147

Bibliography 163

Curriculum vitae 181

Chapter 1

Context and scope

DECADES of reductionistic research in molecular biology, following the discovery of the DNA double helix in 1953 [319], have yielded a tremendous knowledge about the components of biological systems (genes, proteins, etc.). Today, the genomics revolution allows for a new research approach that will deepen our understanding of evolution, development, and life. *Systems biology* uses complete genome sequences and massive amounts of data from high-throughput technologies to understand the components, the linkages between the components, and the dynamic behavior of biological systems [166, 182].

A key challenge of systems biology is to understand the functioning of the entire gene regulatory network (GRN) of each organism, including human, together with its origins and adaptations through evolution. For each cell type in our body, the particular function, shape, location, developmental stage, mitotic phase, age, communication abilities, future state, responsiveness to stimuli, and evolutionary trace is reflected by its set of active genes. The urge to comprehend the regulatory program that controls gene activation is therefore obvious.

The architecture of a GRN is determined by causal *cis*-regulatory interactions [146]: internal genes in the network are transcriptional regulatory proteins (transcription factors) that recognize specific *cis*-regulatory sequences of other internal genes and of batteries of peripheral genes (e.g., differentiation genes). The *cis*-elements are therefore the central elements of a GRN. Moreover, not only do they implement the linkages between the components, but they also implement how the linked components interact dynamically. The latter is done through "*cis*-regulatory logic" [330]. The logic can be modeled as a combination of Boolean and more complex rules that integrate all upstream inputs and produce a scalar output that (stochastically) determines the number of mRNA molecules being transcribed. The availability of complete genome sequences has opened the door towards the detection and characterization of the *cis*-regulatory system of each gene in an organism.

The immediate output of a GRN are messenger RNA (mRNA) molecules that have been transcribed from the activated genes in the network. Although a significant aspect of gene regulation may be represented by subsequent post-transcriptional and post-translational controls that lead to the ultimate protein output, the transcriptional control itself often plays the most prominent role. With the advent of DNA microarrays, the mRNA output levels of essentially all genes in a genome can be measured simultaneously. Such data, together with genome sequences, can provide a means to reversely engineer network linkages and network dynamics. The circumstantial data that is required for the analysis and interpretation of microarray data, like unambiguous clone identification and functional gene annotations, are currently under continuous development and curation.

Although today it may seem a distant aim to reconstruct the complete GRN of an organism, the data and tools of the genomics revolution is allowing for the first steps to be taken [78]. The role of bioinformatics or computational biology in this respect cannot be underestimated. In the light of GRNs, bioinformatics has a long history regarding the research of the network components: gene prediction, detection of homolog sequences, protein structure recognition, biological data management, etc. Today, in the genomic era, new roles are emerging like comparative genomics, expression profiling, proteomics, and system theory approaches for the dynamical modeling of the network.

Aims and rationale in this work

The work presented here is performed in the light of GRNs as explained above. Particularly it is focused on (1) the analysis of mRNA output levels of a GRN measured with DNA microarray technology and (2) the detection of *cis*-regulatory sequences that control the transcriptional process. Unlike most of the published work regarding the detection of transcription factor binding sites, we will work on metazoan sequences. This involves special considerations regarding low signal to noise ratios: small regulatory elements are located in enormous intergenic or intronic regions. This is different from prokaryotes or lower eukaryotes like yeast, where most regulatory elements are located within a few hundreds of base pairs upstream of the translation start site of a gene.

Transcription factor binding sites are short, and they can occur every few hundred base pairs in a sequence, just by chance. To select only those sites that have a high probability of being a real functional site *in vivo*, we will apply and *combine* the following ideas into new methods, strategies, and generic software tools:

1. Genes that are co-regulated by the same factors (i.e., gene batteries), share similar binding sites. The discovery of sites that are present in all or many of the genes in a co-regulated set, has been applied on prokaryotic and yeast sequences since the 1990s, but barely on metazoan sequences. Microarray data clustering allows us to construct gene groups that are *co-expressed*. Depending on the quality and resolution of the data, on the clustering itself, and on the usage of supporting data, the assumption that tightly co-expressed genes are also co-regulated is often valid, and we will often work under this assumption.

2. A second feature of regulatory sequences that can be used to reduce the search space and that increases the confidence of a prediction, is their evolutionary conservation between orthologous genes. We will use this so called *phylogenetic footprinting*, most often by aligning genomic sequences of human and mouse orthologs, in combination with gene co-expression.
3. The transcriptional regulation in higher eukaryotes is of combinatorial nature. A consequence thereof is that the transcription factor binding sites that receive the multiple regulatory inputs, are often clustered within a confined region of DNA. We will use this binding site clustering in our site prediction methods.

The philosophy that we will adopt regularly in this work, and to which our algorithms will be optimized is shown in Figure 1.1. As depicted in this figure, we will also deal with the detection of target genes for certain transcription factors in the full genome sequence. For all our goals, there are several important sources of genomic data that help us to achieve them. The data we will use extensively are gene expression data as measured by DNA microarrays, DNA sequences of the fully sequenced metazoan genomes (human, mouse, fish, etc.), and functional gene annotation data based on the Gene Ontology vocabulary. Our aim is, on the one hand, to understand and to analyze these data sources individually, and on the other hand to integrate and mine these heterogenous data to find new biological hypotheses. We will validate, or at least illustrate all developed methods, tools, and strategies with one or more biological cases. To this end we will use either existing data sets from the literature, newly compiled data sets from publicly available databases, or data that originates from collaborations with molecular biologists of research groups of the university. A last, more general aim is to bring the developed bioinformatics methods and strategies closer to molecular biologists by making them available via intuitive user-friendly software tools.

Achievements

The main achievements of this work are summarized in Table 1.1.

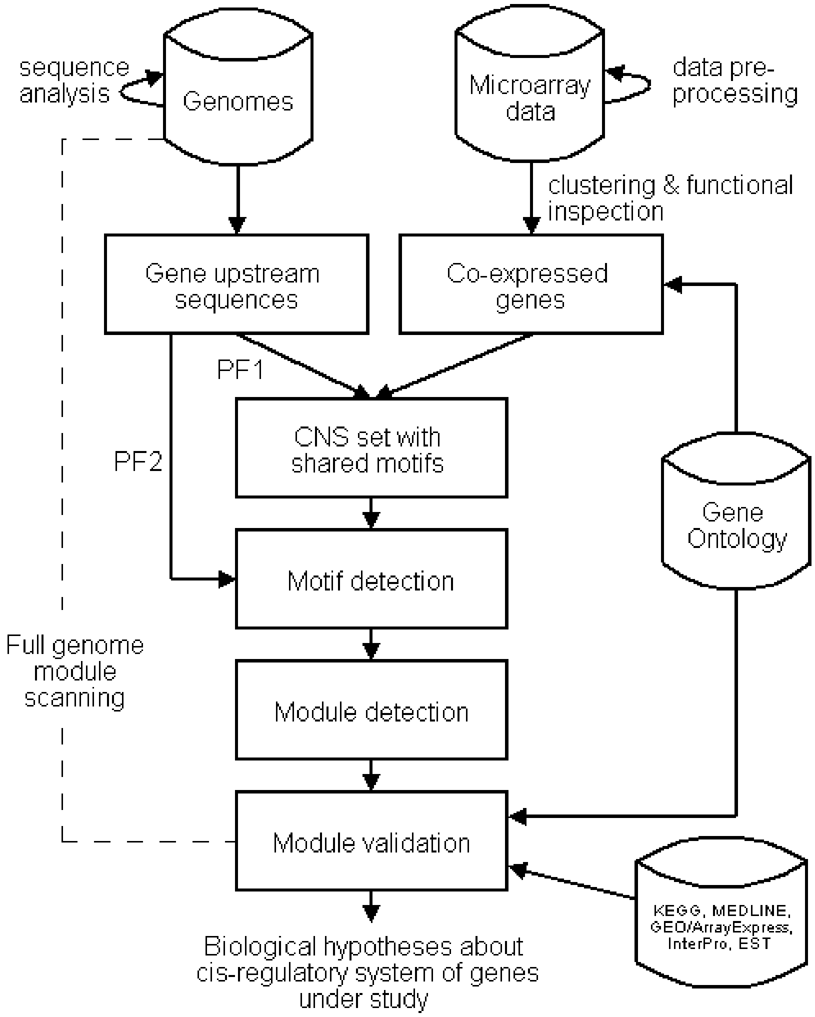

sequence analysis
Genomes
Microarray data
data pre-processing
clustering & functional inspection
Gene upstream sequences
Co-expressed genes
PF1
PF2
CNS set with shared motifs
Gene Ontology
Motif detection
Full genome module scanning
Module detection
Module validation
KEGG, MEDLINE, GEO/ArrayExpress, InterPro, EST
Biological hypotheses about cis-regulatory system of genes under study

Figure 1.1: Schematic overview of the analysis pipeline that is proposed in this work. For several tasks, and also for the integration of multiple tasks into a pipeline, new algorithms, strategies, and software tools are presented in this work. An explanation of these achievements can be found in Table 1.1. PF1 a phylogenetic footprinting approach to detect larger blocks of conserved non-coding sequences (CNS) between two or more orthologous sequences. The CNSs may carry *cis*-regulatory potential because of their conservation. PF2 is another PF approach to directly detect motifs in sets of orthologous sequences.

Table 1.1: Achievements in this work.

Achievement	Ch.	Pub.	Software
Literature survey on the biology and bioinformatics of eukaryotic gene regulation, including original contributions, for example a summary of methods for *cis*-regulatory module detection.	2		
Literature survey on the sharing of microarray data, including standards and compendia.	3	[215]	
Microarray data analysis in collaboration with the Center for Human Genetics (K.U.Leuven and VIB): (1) preprocessing, including state-of-the-art dye-normalization and original filtering methods; (2) clustering; (3) functional analysis with Gene Ontology; (4) data management with home-grown data model and MySQL database; (5) intelligent data retrieval and visualizations.	3	[202, 77]	NEURODIFF, GO4G
Motif detection: (1) original combination of gene co-expression and phylogenetic footprinting; (2) contribution to the development of a new approach to score a sequence with a position weight matrix [296]; (3) original integration of the motif detection method with user-friendly visualizations and with the Ensembl database for sequence retrieval; (4) statistical testing for motif overrepresentation using a published method [307]; (5) validation of the system on benchmark data sets; (6) usage of the system in two collaborations with the Center for Human Genetics (K.U.Leuven and VIB).	4	[4, 83]	TOUCAN

"Ch." is the Chapter reference. "Pub." are the related publications. The software tools in the last column are implemented specifically for this work. For the URLs where the software can be used or downloaded, see Appendix B. This table is continued on the next page.

Table 1.1: Achievements in this work (continued).

Achievement	Ch.	Pub.	Software
Module detection: (1) original combination of gene co-expression, phylogenetic footprinting, and binding site clustering; (2) construction of a database of conserved non-coding sequences in the promoters of human-mouse orthologs; (3) genome-wide screening for modules; (4) module validation using a measure for functional coherence of putative target genes; (5) validation of the system on artificial data and on biological data.	5	[7, 6]	ModuleSearcher, ModuleScanner, GO4G
Module validation: (2) original system for the integration of multiple genomic information sources to rank a set of test genes according to their similarity with a set of training genes; this strategy can be applied to validate putative target genes of a *cis*-regulatory module (e.g., found by the ModuleScanner) or to prioritize putative disease genes.	6	[3]	ENDEAVOUR
Genome sequence analysis: (1) re-evaluation of the nucleotide composition around the transcription start site of human genes; (2) original comparison of the nucleotide compositions among several metazoan species; (3) analysis of the composition profiles in relation with gene expression.	7	[5]	

Chapter 2

An overview of gene regulation: biology and bioinformatics

THE recent completion of various genome projects (human [173, 311], fly [2], mouse [318], rat [119], etc.) has led to estimates of the numbers of genes much lower than expected, and the number of genes that has been found in our own genome (~25,000), is only two times larger than in the fruit fly genome. Furthermore, more than 60% of human genes are related to particular genes in the fly and the worm. It is now believed that the heritable genomic regulatory programs largely determine the morphological differences between species and that they underlie both evolution and development. The motivation to understand how genes are regulated has therefore never been stronger [146, 78, 59].

The role of bioinformatics in the study of gene regulation has become greater during the last decade, both because of the huge amount of sequence and annotation data that are becoming available—and that make for example computational studies feasible on a genome-wide scale and across species—and because of the use of the high-throughput measurements of gene expression using microarrays that require computational analysis methods.

In this introductory chapter we will walk through the biology of gene regulation and through several computational techniques that are helping to unravel and understand it.

2.1 Gene regulation and development

Development, in which a single fertilized egg cell grows into an entire organism, produces a certain morphology. The view that development can be seen as a process that is harmoniously organized by gene products is now generally accepted thanks to a better understanding of the nature of genes and of the mechanisms of gene regulation. Although the DNA of almost all cells in an animal is identical, different cells can acquire different forms, structures and functionalities in the diverse organs of the body. This is possible through differential gene expression: different cells express different subsets of genes. The regulatory program encoded in the genome accurately specifies when genes are turned on and off over the course of development. The accuracy is illustrated by the fact that the outcome of the regulatory program, which is the completed organism, is always the same. An example of differential expression and of genetic subprograms during development is specification, the process by which cells acquire the identities or fates that they and their progeny will adopt. For specification to occur, genes have to make decisions, depending on the inputs they receive (see the information processing capacities of *cis*-regulatory systems in 2.4.6). As stated by Davidson [78], this is because "development depends on creating new spatial and temporal domains of gene expression from preexisting information".

2.2 Gene regulation and evolution

"If morphological diversity is all about development, and development results from genetic regulatory programs, then is the evolution of diversity directly related to the evolution of genetic regulatory programs?" is an intriguing question asked, among others, by Carroll et al. [59] and by Davidson [78]. Both authors explain why the answer to this question is—simply put—yes. Before the advent of molecular biology there were two theories to explain how diverse forms of animal life arose during evolution. The first said that new forms arose *because* the environment changed. But "while changes in climate or other changes definitively presented selective forces, they do not generate heads or appendicular forms; only genes do that" [78]. The second one was that point mutations in DNA coding sequences (causing changes in the protein sequences) accumulated little by little, providing the opportunity for selection. However, the differences between animals cannot be explained by differences in key regulators of development—transcription factors and signalling pathways—because these are all "panbilaterian": they are highly similar among the bilaterally symmetrical animals and their functional conservation can often be illustrated by the potential to be exchanged between different animals (e.g., *Drosophila* Atonal fully rescues the phenotype of Math1 null mice [315]). Thanks to the advancements in regulatory molecular biology, the interpretation of evolutionary change is taking the form that morphological differences are generated largely by alterations in developmental regulatory sequences. Such alterations can have several causes, such as stepwise mutational changes in *cis*-regulatory DNA, transpositional insertions of regulatory modules or of genes in the vicinity of these modules, sequence deletions, local genomic rearrangements, replication of genes or their *cis*-regulatory target sites, gene conversion, etc [78, 59] (see also Section 2.7).

2.3 Gene regulation and disease

As correct gene expression underlies all physiological processes, aberrant gene expression can be a major cause for disease, including various forms of cancer. Indeed, alteration of transcription factor function as a result of either gain or loss of function mutations has now been established as a frequent cause of neoplastic transformation and tumor progression in humans. These mutations can be of any kind, like point mutations, deletions, insertions, or chromosomal translocations.

Some examples where transcriptional regulation is out of control can be found in human acute leukemias where chromosomal translocations rearrange the regulatory and coding regions of a variety of transcription factor genes [190]. For example, a translocation can cause a transcription factor that is normally expressed at low levels to be placed under the control of a powerful enhancer. *IG* (immunoglobulin) or *TCR* (T-cell receptor) genes are examples of highly expressed genes for which the enhancers have driven the expression of TF's like MYC (e.g., in B-cell leukemia and Burkitt's lymphoma). Chromosomal breakpoints can also occur within introns between two transcription factor genes on different chromosomes, producing a fusion gene that encodes a chimeric transcription factor with altered function, for example the CBF^{β}*-MYH11* fusion genes lead to alterations in the CBF transcription complex in acute myeloid leukemias.

Other types of cancer can also be caused by malfunctioning regulatory control. For example, *PLAG1* (pleomorphic adenoma gene 1), which is developmentally regulated, has been shown to be consistently rearranged in pleomorphic adenomas of the salivary glands. *PLAG1* is activated by the reciprocal chromosomal translocations involving 8q12 in a subset of salivary gland pleomorphic adenomas (summary from LocusLink).

A better understanding of normal and aberrant gene expression could lead to the identification of potential new targets for therapeutic intervention. Altered gene expression of transcription factors can be a cause of disease, but altered gene expression is often also a consequence of the disease. This fact makes it possible to characterize tumors by the gene expression profiles of multiple genes (i.e., molecular fingerprints), and DNA chip technology offers great promise for diagnostic, prognostic and pharmacogenomic applications [251].

2.4 Transcriptional regulation in eukaryotes

Eukaryotes employ diverse mechanisms to regulate gene expression, including chromatin condensation, DNA methylation, transcriptional initiation, alternative splicing of RNA, mRNA stability, translational controls, several forms of post-translational modification, intracellular trafficking, and protein degradation [183]. Of these broad categories, the most common point of control is the rate of transcriptional initiation [178]. For virtually every eukaryotic gene where relevant information exists, transcriptional initiation appears to be the primary determinant, or one of the most important determinants, of the overall gene expression profile [325].

Only some of the genes in a eukaryotic cell are expressed at any given moment. The proportion and composition of transcribed genes changes considerably during the life cycle, among cell types, and in response to fluctuating physiological and environmental conditions. Given that eukaryotic genomes contain on the order of five to fifty thousand genes, regulating this differential gene expression requires an exceptionally complex array of specific physical interactions among macromolecules. The form of the machinery that controls transcription is that of a *gene regulatory network* (GRN). The GRN determines the transient regulatory states in a cell and the batteries of downstream genes they will express [325, 146].

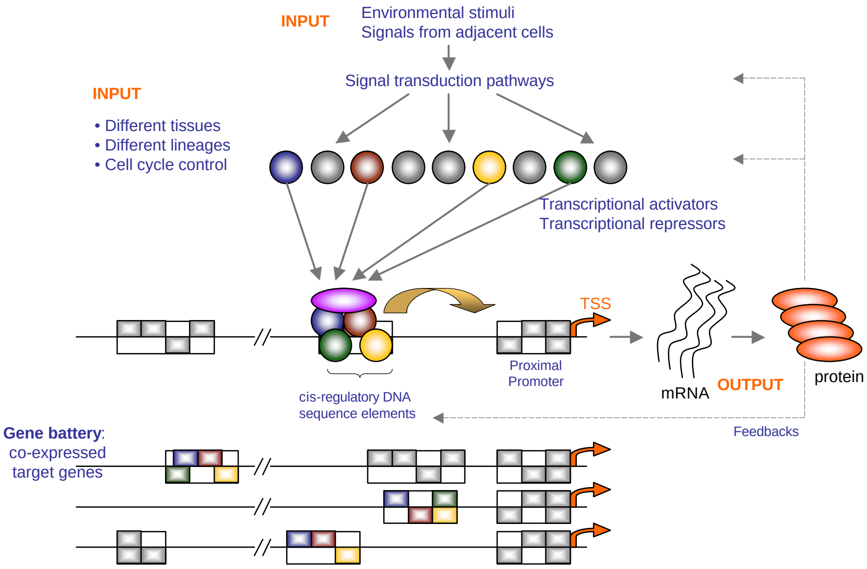

INPUT
Environmental stimuli
Signals from adjacent cells
Signal transduction pathways
INPUT
• Different tissues
• Different lineages
• Cell cycle control
Transcriptional activators
Transcriptional repressors
TSS
Proximal Promoter
cis-regulatory DNA sequence elements
mRNA
OUTPUT
protein
Feedbacks
Gene battery: co-expressed target genes

Figure 2.1: An imaginary gene regulatory network where the central elements are *cis*-regulatory modules.

Figure 2.1 depicts all the elements of a GRN: several signalling pathways that transduce network inputs (e.g., hormone binding on a cell surface receptor) into the (in)activity of certain transcription factors. The central elements of a GRN are *cis*-regulatory elements (CRE) on which TFs and co-activators can assemble. CREs thereby process all the information of the fluid upstream biochemical signalling pathways and direct the rate of transcription initiation by communicating with the basal transcription apparatus.

2.4.1 The eukaryotic gene

The basic nature of the gene was defined by Mendel more than a century ago. Summarized in his two laws, the gene was recognized as a "particulate factor" that passes unchanged from parent to progeny. A gene may exist in alternate forms (alleles).

Now we know that a gene consists of DNA, and that a chromosome consists of a long stretch of DNA representing many genes. A gene is one unit of DNA that performs a function. The RNA that is formed after transcription is either messenger RNA (mRNA) that codes for a protein or polypeptide, or the RNA itself can be functional (i.e., RNA genes, see further). The structure of a generalized eukaryotic gene (we will use the general term "gene" for protein coding genes) is depicted in Figure 2.2. In contrast with prokaryotic genes, eukaryotic genes are often interrupted: exons are the sequences represented in the mature RNA, and introns are the intervening sequences that are removed when the primary transcript is processed to give the mature RNA. For many genes there can be multiple combinations for the recombination of multiple exons during mRNA splicing (i.e., alternative splicing). This results in the fact that one gene can have several distinct transcripts that can also be differentially regulated.

The *cis*-regulatory system of a gene (the *trans* system are the transcription factors and co-factors) consists of a core promoter where the RNA polymerase complex assembles, a proximal module with several transcription factor binding sites (TFBS), and several distal modules, each with several TFBSs. All these elements of the regulatory system will be described in more detail hereafter.

Gene prediction

Ever since the availability of DNA sequences there has been a need for programs to automatically identify the proteins encoded in genomic DNA. Many advances have been made during the 90's, using content sensors (similarity to proteins and transcripts, codon usage, etc.), signal sensors (translation start and stop, splice sites, etc.), and combinations of both. The algorithms are often based on dynamic programming or hidden Markov models. Now most nucleotides can be identified correctly as either coding or noncoding [66, 278, 208]. However, the most difficult part of gene prediction in eukaryotes has always been the prediction of the complete gene structures, and this is still in need for improvement. Methods based on similarity between genomic DNA and EST and cDNA sequences, and methods based on genome comparisons (e.g., comparing the human genome with other complete vertebrate genomes such as those of mouse and fish) are playing a crucial role in current genome annotations.

The leading source of human genome annotation is the Ensembl project (http://www.ensembl.org [149]) that currently (Ensembl version 18 of November 2003) provides a comprehensive source of stable automatic annotation of the following genomes: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), pufferfish (*Fugu rubripes*), fruit fly (*Drosophila melanogaster*), mosquito (*Anopheles gambiae*), worm (*Caenorhab-*

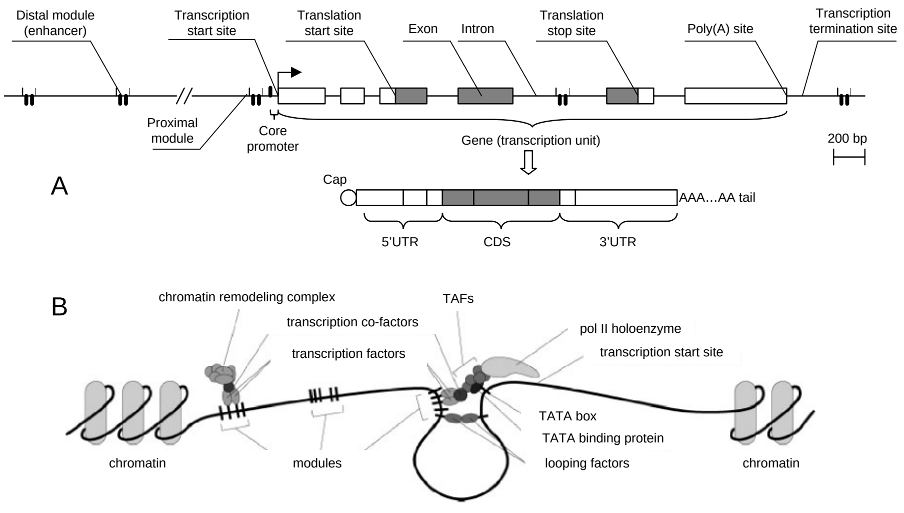
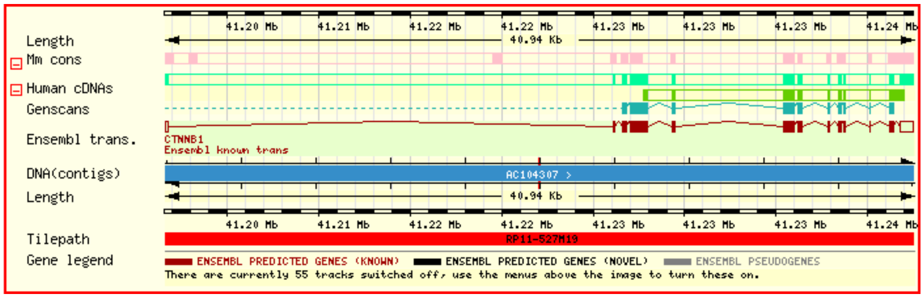


Figure 2.2: The eukaryotic gene and its regulatory regions. (A) Organization of a generalized eukaryotic gene showing all structural and functional components (introns, exons, CDS, UTRs). The gene is shown in relation with the proximal and distal *cis*-regulatory modules that control its transcription. (B) Idealized regulatory system in operation: chromatin modifying factors are bound to a distal module and specific transcription factors are bound to another distal module and interact together with co-factors with the general transcription factors and the basal transcription machinery at the core promoter, thereby initiating transcription. Adapted from [325] and [332].

ditis elegans and *C. briggsae*), and preliminary data of chimpanzee (*Pan troglodytes*) and chicken (*Gallus gallus*). The gene build process of Ensembl uses gene prediction software (GenScan [52] and GeneWise [32] programs), protein and cDNA data, and similarities to other genomes. For the functional annotation of genes, Ensembl uses data from Gene Ontology, InterPro, OMIM, SAGE expression, and other. An example of a “contigview” of a gene is shown in Figure 2.3.

The current estimates of the number of protein coding genes in the human genome are converging to around 25,000 genes. Their DNA (translated and untranslated) represents about 26.55% of the total genomic DNA and the exons alone (i.e., coding sequences + 5' and 3' untranslated regions) represent only 1.48 % of the genome (calculated from Ensembl release 18 using the 22,184 Ensembl stable genes [850,113,396 bp in genes and 47,657,184 bp in exons out of 3,201,762,515 bp]).

Next to automatic annotation there are also initiatives of systematic manual annotation on a gene by gene basis. The best known initiative for vertebrate genomes is the Vertebrate Genome Annotation (VEGA) database at the Sanger Institute (http://vega.sanger.ac.uk/).



There are currently 55 tracks switched off, use the menus above the image to turn these on.

Figure 2.3: "Contigview" in Ensembl of the 40.94 kb large genomic region spanning the β-catenin gene (HUGO = CTNNB1). The transcript structure with exons and introns is denoted as "Ensembl trans", and above it are a selected number of annotated features (out of dozens of available features). "Mm cons" are conserved regions with the mouse CTNNB1 homolog. From the difference between the GenScan prediction and the Ensembl prediction, it can be seen that cDNA mapping is useful for gene prediction.

Non-coding RNA genes

Non-coding RNA genes (ncRNA) produce functional RNA molecules rather than encoding proteins. The above mentioned methods for gene prediction (cDNA cloning and EST sequencing, identification of conserved coding exons by comparative genome analysis, and computational gene prediction) work best for large, highly expressed, evolutionarily conserved protein coding genes, and they probably underestimate the number of other genes. They essentially do not work at all for RNA genes. Classical examples of ncRNA are transfer RNA and ribosomal RNA, but recently, several groups have carried out systematic searches for ncRNA genes. All of them indicate that the prevalence of ncRNA genes has been underestimated, and new RNAs in different flavors continue to appear, with control functions at the transcriptional or post-transcriptional level (for review, see [95]). To our knowledge, so far there has not been a single thorough study on the transcriptional regulation of ncRNAs. The methods in this dissertation will deal with the transcriptional regulation of protein coding genes.

2.4.2 The core promoter

The enzyme RNA polymerase II (RNAPII) together with the auxiliary general transcription factors (GTF, usually described as TFIIx) constitute the basal transcription apparatus (BTA) that is needed to transcribe any gene. The BTA assembles at the core promoter and positions the start of transcription relative to coding sequences. Transcription that is initiated by this minimal set of proteins is referred to as basal transcription.

If all genes use the same machinery to initiate transcription, we may expect to find certain conserved sequence components involved in the binding of RNA polymerase II and the general factors in *all* genes. Unfortunately for computational biologists that strive to recognize them, this is not the case. There appear to be several classes of core promoters. One important class only consists of a TATA-box at ∼25 bp upstream of the TSS. The TATA box is found in all eukaryotes and the 8 bp consensus consists entirely of A·T base pairs. Recognition of the TATA box is conferred by the TATA-binding protein (TBP). TBP forms a complex with TBP-associated factors (TAF) and TFIID and the whole complex puts the RNA polymerase at the right position for the initiation of transcription. A second class of core promoters are TATA-less promoters. These may have an initiator (Inr) element around the TSS that may be described in the general form YYANTAYY where the first A is at TSS. In addition to these two promoter classes, there are also promoters which have both TATA and Inr elements, and promoters that have neither [183, 229, 159]. Another promoter element is the downstream promoter element (DPE) that is present in some TATA-less, Inr-containing promoters about 30 bp downstream of the TSS. It was found in both human and *Drosophila* [172].

The core promoter is necessary for transcription but is apparently not a common point of regulation, and it cannot by itself generate functionally significant levels of mRNA [178]. The specificity and the functional activity is conferred by a collection of diverse transcription factor binding sites often organized in modules. Proteins bound to these sites produce a scalar response: the frequency with which new transcripts are initiated [78] (see the modules in Figure 2.2 and Section 2.4.6).

CpG islands

Methylation of DNA by DNA methyltransferases (Dnmt) is one of the parameters that controls transcription in vertebrates. The targets for such methylation are CpG doublets—cytosine (C) bases adjacent to guanine (G) bases (the p in CpG denotes the phosphodiester linkage). In most human somatic cells, about 80% of CGs are methylated and the distribution of methylated and nonmethylated CGs is not random, but conforms to a pattern. The most obvious features of the pattern are large clusters of nonmethylated CGs at the promoters of many genes (CpG islands) [28]. It has been found that DNA methylation has a repressing effect on transcriptional activation, possibly mediated by the binding of a specific methyl-CpG binding protein [183].

CpG islands can be found by directly testing for the absence of cytosine methylation. But there is a simpler way of finding CpG islands. Most CpG dinucleotides in the vertebrate genome are methylated on the C base and spontaneous deamination of C-methyl residues gives rise to T-residues. (Spontaneous deamination of ordinary cytosine residues gives rise to uracil residues that are readily recognized and repaired by the cell.) As a result, methyl-CpG dinucleotides steadily mutate to TpG dinucleotides. Unmethylated CpG islands have a normal frequency of CpG dinucleotides that is roughly 4% (obtained by multiplying the typical fraction of Cs and Gs, which is 0.21) while the rest of the genome has a frequency of about one fifth of the expected frequency. CpG islands are defined as regions longer than 200 bp with over 50% of G+C content and a CpG frequency that is at least 1.667 of that statistically expected.

The CpG density defines two classes of promoter. In the CpG-related class, the frequency of CpGs is the same as the genome average, which is roughly one every 100 bp. This class invariably includes genes whose expression is restricted to a limited number of cell types (last two genes in Figure 2.4). In contrast, the 5' end of the genes belonging to the other group is surrounded by a region of ∼1 kb long where the frequency of CpGs is approximately 10 times higher than the genome average (the first two genes in Figure 2.4). According to [12, 11], approximately 60% of mammalian gene promoters are associated with one or more CpG islands. This includes all the housekeeping genes—those expressed in all cell types—and about half of the tissue-specific genes. Davaluri et al. [80] defined a CpG score using only the CpG dinucleotide percentage in a window and found that about 70% of the first exons in the human genome are CpG-related. The correlation between CG content and promoters is one of the best features in promoter prediction (see Section 2.4.2).

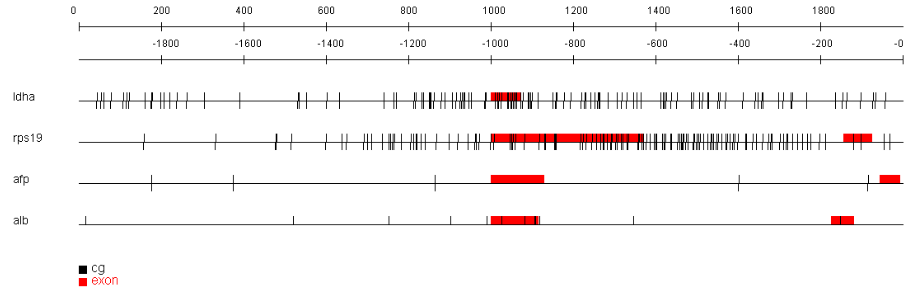


Figure 2.4: CpG content around transcription start site. Two housekeeping genes *LDHA* and *RPS19* with many CpG doublets in the [-1000,+1000] region around TSS and two cell-type specific genes *AFP* and *ALB* with few CpGs in this region. Figure generated with TOUCAN [4].

DNA structure in core promoters

Packaging of DNA into chromatin limits the accessibility of the DNA template for the BTA and has been found to inhibit transcriptional initiation. The derepression of transcription by partial unfolding of chromatin is likely to constitute an important part of gene regulation, and TFs and TCFs can play a role in chromatin remodeling. For example, some are histone acetyltransferases like p300/CBP, which is a coactivator that links an upstream TF (e.g., AP-1, MyoD) to the BTA. p300/CBP acetylates the N-terminal tails of H4 in nucleosomes and acetylation is associated with gene activation (while the absence of acetyl groups is associated with a more condensed, inactive structure). Another example of how TFs can influence DNA three-dimensional structure is the bending of DNA by architectural TFs to facilitate protein binding [183, 228]. The three-dimensional structure of DNA can depend on the DNA sequence itself, and like the CG content, structural information too has been used in promoter prediction algorithms.

Promoter prediction

Algorithms for general promoter prediction can be classified into two groups: search-by-signal and search-by-content [223]. The search-by-signal algorithms make predictions on the basis of the detection of relatively conserved signals and conserved spacing among patterns such as the TATA-box, Inr, DPE, or TFBS outside the core (see further). PROMOTER2.0 [167] uses a combination of neural networks and genetic algorithms, ProScan [236] uses position weight matrices of TFBS. The search-by-content algorithms identity promoters on the basis of the sequence composition. Discriminant analysis has been used in CorePromoter with pentamer frequencies in consecutive 100 bp regions as features [332]. These programs predicted about ∼30-50% of the promoters correctly but predicted one false positive promoter each kilobase [105]. PromoterInspector [255], which is based on context features extracted from training sequences by an unsupervised learning technique, produced only one false positive every 40 kb, a significant improvement.

The more recent algorithms have included other features that improved both sensitivity and specificity: CpG content [153, 80, 133, 91], first splice-donor sites [80], transcript information [187], and structural sequence features such as bendability or conformation [223].

A more direct way to find the TSS and thus the core promoter is to map cDNA sequences to genomic DNA; the 5' end of the cDNA should coincide with the TSS. However, most of the cDNA sequences stored in current databases are imperfect in the sense that they lack the precise information of 5' end termini. Suzuki et al. [286] have developed the *5' oligo-capping* method to obtain full-length cDNAs. The experimentally determined TSSs for 8,793 human genes (as of Jan 2004) are stored in the publicly available database DBTSS [287]. PromoSer is another publicly available database that contains TSSs for human, mouse, and rat genes obtained by aligning a large number of partial or full-length mRNA sequences to genomic DNA [131].

2.4.3 Transcription factor binding sites

Producing functionally significant levels of mRNA requires the *sequence specific* association of transcription factors with DNA sequences outside the core promoter [178, 325]. They can occur both in a region of ∼200-300 bp upstream of the core promoter (i.e., the proximal promoter) and at sites more distal to the core promoter either upstream or downstream of the gene or in introns (see further).

Most transcription factor binding sites (TFBS) span 5-8 bp and they can almost always tolerate at least one, and often more, nucleotide substitutions without losing functionality (in contrast to most restriction enzymes). The sites of recognition are a family of similar sequences, although there can be considerable variability. An understanding of the sequence-specificity of DNA-protein interactions has resulted from studies of the effects of mutations in the DNA-binding sites and the amino acid residues implicated in binding, for which recently also microarrays were used [50]. Regulatory systems can take advantage of this variability in the sites to control the level of transcription because of differences in the affinities between factor and site. For example, low affinity sites compete with high affinity sites for binding to the TF and thus require that more TF be present [276].

TFBSs in the proximal promoter

Some transcription factors are not part of, but very frequently acting in concert with, the BTA. The TFBS for these factors are often present in ∼200-300 bp upstream of the TSS. For example, on the order of half of all vertebrate promoters contain a somewhat conserved CCAAT-box where a large number of factors can bind to. Ohler et al [224] have found several motifs for unknown factors in *Drosophila* proximal promoters using MEME [17] and Gibbs sampling [176].

TFBSs in distal modules

Disjunct regions of DNA of several hundreds of bp in length where TFBSs are clustered together, often produce discrete portions of the total transcription profile. Such a region is called a *cis*-regulatory module (CRM or simply module). They have also been termed *enhancer* (enhancing transcription) or *silencer* (repressing transcription), and in fact the proximal promoter can, according to this definition, also be regarded as a *cis*-regulatory module (i.e., the “proximal module”) in case it produces a discrete portion of the expression—which is often the case. See Section 2.4.6 and further for a discussion on modules.

Protein-DNA interactions

To unravel the stereochemical rules of protein–DNA binding, structures of protein–DNA complexes solved by X-ray crystallography can be used. A recent classification of such complexes was done by Luscombe and colleagues [196]. About two-thirds of the contacts between amino acid side chains and nucleotide bases are van der Waals contacts, about one-sixth are hydrogen bonds and the last sixth are water-mediated bonds [197]. In most studies that have been performed on protein-DNA interactions, there appear to be favored interactions but the consensus is that DNA-binding varies substantially between protein families, and that at present no simple code can adequately describe the recognition of target sites on nucleic acids. Luscombe and colleagues [197] however claim to have found some rules for universal specificity in all complexes and they have constructed a web-based “Atlas of Side-Chain Base Contacts”. The DNA binding domain is often a short alpha helix, sometimes a beta strand or a loop, that inserts into the major groove of double-stranded DNA (see Figure 2.5 for an example).

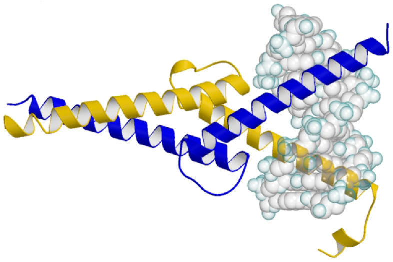


Figure 2.5: DNA-protein interactions. Complex of a helix-loop-helix transcription factor *SREBF1* (sterol regulatory element binding transcription factor 1) bound to the promoter of *LDLR* (low density lipoprotein receptor) (the PDB entry of the complex is 1am9). From [196].

Representation of TFBS

There are basically two ways to represent the range of TFBSs that can bind a particular TF with significantly higher specificity than random DNA under physiological conditions [145, 276]. Both are made from a set of known binding sites that are first aligned to maximize sequence conservation (Fig 2.6.A). The alignment method that is used can already introduce variability in the quality of the model. The simplest and oldest model is the consensus sequence, although the way this is defined is somewhat arbitrary. The consensus sequences match all of the example sites closely, but not necessarily exactly, and there is a trade-off between the number of mismatches allowed, the ambiguity in the consensus, and the sensitivity and specificity of the representation. The alphabet used in consensus sequences is the IUPAC (International Union for Pure and Applied Chemistry) degenerate alphabet (see the *Notation* section). The second possible representation is the matrix model. The simplest form is the alignment matrix or count matrix, which lists the number of occurrences of each letter at each position (Fig 2.6.B). From the count matrix, a position specific frequency matrix (PSFM) can be constructed by calculating the frequencies of each letter at each position and introducing pseudocounts (a zero count means this letter is not observed at this position, but it does not mean it does not exist in the genome) (Fig 2.6.C). The PSFM is used in the MotifScanner and MotifLocator algorithms (see Chapter 4). Instead of a frequency matrix, a weight matrix can also be used (Fig 2.6.D) in which the weights can be calculated using the following formula [145]:

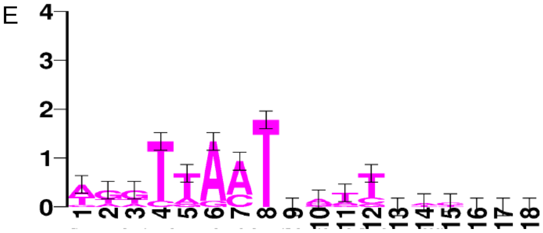
$$\ln\frac{(n_{i,j}+p_i)/(N+1)}{p_i} \approx \ln\frac{f_{i,j}}{p_i}, \tag{2.1}$$

where N is the total number of sequences (eleven in the HNF-1 example), p_i is the *a priori* probability of letter i (in this example 0.25 for all the bases, but p_i's can be calculated from the genome), and $f_{i,j} = n_{i,j}/N$ is the frequency of letter i at position j.

A

taattactaaccaaacta
atgtaaataattttccaa
aggttaatgattggcagc
agttaaatagatatcaga
atatggctggttgaggcc
tgtctactctagcctaca
aagttaattagtaattgt
tgtttaataatcttctgc
aggttaattcttctctaa
ttgttaataattaatact
gggttaatggttaatcgg

Agg**ttaat**aattaacaga consensus
NNN**ttaAt**nn**Tt**nnnNnn alternate consensus



B

A	6	2	2	0	2	10	8	0	5	6	2	0	5	5	1	4	2	5
C	0	0	0	1	0	0	3	0	1	1	1	2	2	1	5	3	3	3
G	1	6	6	0	1	1	0	0	3	3	1	1	2	1	1	1	5	1
T	4	3	3	10	8	0	0	11	2	1	7	8	2	4	4	3	1	2

C

A	0.52	0.19	0.19	0.02	0.19	0.85	0.69	0.02	0.44	0.52	0.19	0.02	0.44	0.44	0.10	0.35	0.19	0.44
C	0.02	0.02	0.02	0.10	0.02	0.02	0.27	0.02	0.10	0.10	0.10	0.19	0.19	0.10	0.44	0.27	0.27	0.27
G	0.10	0.52	0.52	0.02	0.10	0.10	0.02	0.02	0.27	0.27	0.10	0.10	0.19	0.10	0.10	0.10	0.44	0.10
T	0.35	0.27	0.27	0.85	0.69	0.02	0.02	0.94	0.19	0.10	0.60	0.69	0.19	0.35	0.35	0.27	0.10	0.19

D

A	0.73	-0.29	-0.29	-2.48	-0.29	**1.23**	**1.01**	-2.48	0.56	0.73	-0.29	-2.48	0.56	0.56	-0.88	0.35	-0.29	0.56
C	-2.48	-2.48	-2.48	-0.88	-2.48	-2.48	0.08	-2.48	-0.88	-0.88	-0.88	-0.29	-0.29	-0.88	0.56	0.08	0.08	0.08
G	-0.88	0.73	0.73	-2.48	-0.88	-0.88	-2.48	-2.48	0.08	0.08	-0.88	-0.88	-0.29	-0.88	-0.88	-0.88	0.56	-0.88
T	0.35	0.08	0.08	**1.23**	**1.01**	-2.48	-2.48	**1.32**	-0.29	-0.88	**0.88**	**1.01**	-0.29	0.35	0.35	0.08	-0.88	-0.29

Figure 2.6: Representation of transcription factor binding sites. (A) A set of aligned human DNA sequences that are binding sites for the transcription factor HNF-1 (from the TRANSFAC database). (B) Alignment matrix or count matrix generated from A. (C) Position specific frequency matrix (PSFM) generated from (B). (D) Position weight matrix generated from (C). (E) Logo computed using *alpro* and *makelogo* [257] at http://www.bio.cam.ac.uk/seqlogo/logo.cgi. Each of these representations is explained in the text.

The PWM representation is interesting because the logarithms of the frequencies are proportional to the binding energy contribution of the bases [24]. Binding sites can also be viewed from the perspective of their "information content" [258], which also fits with the binding energy analysis. The information content at a position in a site is defined by

$$I_{\text{seq}}(i) = \sum_{b=A}^{T} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}, \tag{2.2}$$

where i is the position within the site, b refers to each of the possible bases, $f_{b,i}$ is the observed frequency of each base at position i, and p_b is the frequency of base b in the whole genome. I_{seq} is between 0, for positions that are 25% of each base, and 2 *bits* for positions completely conserved as one base. I_{seq} is also known as the relative entropy and the Kullback-Leiber distance (to the uniform distribution). It is also a normalized log-likelihood ratio statistic and so can be used to estimate the statistical significance of the pattern [277]. The information content can be represented graphically in a sequence logo (Fig 2.6.E) where the height of each letter in the stack represents the amount of information (in bits) that this position holds and the error bars represent the confidence interval because of the limited sample size.

The TRANSFAC database [323] contains a collection of transcription factors, experimentally determined binding sites and target genes for these factors, and count matrices derived from the alignment of binding sites. The professional release 7.3 of TRANSFAC contains 13112 binding sites and 674 count matrices of which 493 have been created from sites in vertebrate sequences. It is important to note that the full matrix of binding sequences is not yet known for most TFs, even in well-studied species. Recently another database named JASPAR [252] was created and contains 111 curated non-redundant PWMs (as of March 2004).

Weight matrices are based on several assumptions that remain to be firmly established, and their underlying principles may be an over-simplification of the biochemistry of protein-DNA interactions. One limitation is that the recognition sequence is of fixed length. Another assumption of PWMs is that each position of a binding site is modeled as making an independent contribution to the overall binding affinity of the site. Although this provides a good approximation of the true nature of the specific protein-DNA interactions [21], there are more sophisticated methods that model a binding site with dependencies [165, 51].

Experimental detection of TFBSs

Experimental methods to detect TFBSs *in vitro* are *DNAse hypersensitivity studies*, *electrophoretic mobility shift assays*, and *systematic evolution of ligands by exponential enrichment (SELEX)*. SELEX is a high-throughput method to select high-affinity binding sites to a TF of interest from randomized double-stranded DNAs [234]. Recently two technological platforms have been developed for *location analysis*, or the genome-wide detection of TFBSs *in vivo*. That is, in principle all functional binding sites in the genome of a certain TF can be detected in one run, at least those to which the TF is bound. These two platforms are the ChIP-chip [247] and DamID [310] methods. In ChIP-chip, cells under certain conditions are fixed, harvested, and disrupted and the DNA fragments that are cross-linked to a TF of interest are enriched by immunoprecipitation with a specific antibody. After reversal of the cross-links, the enriched DNA is amplified, labeled with fluorescent dye (e.g., Cy5), and hybridized to a cDNA microarray containing intergenic sequences. The positive spots are promoters of genes that are potentially regulated by this particular factor. Iyer et al. [154] and Ren et al. [247] have applied this technique in yeast for SBF and MBF transcription factors and for Gal4 and Ste12 factors respectively. Lee et al. [177] have applied performed such location analysis for 141 transcription factors in yeast and used this wealth of data to find general network motifs in the yeast regulatory network. DamID [310] is based on creation of a fusion protein consisting of *Escherichia coli* adenine methyltransferase (Dam) and the TF of interest. The Adenine in GATC sequences near the binding sites of this TF will be methylated (while methylation of adenines is usually absent in eukaryotes) and can be detected using Southern blot, PCR and microarray assays that take advantage of restriction enzymes that are methylation sensitive.

Computational detection of TFBSs

TFBS can be discovered in sequences by searching for matches to a consensus sequence or by scoring a sequence with a PWM. The latter—more sensitive—method involves simply adding the matrix weights of each occurring letter in a test sequence together and normalizing this for the length of the matrix. The normalized score, between 0 and 1, is calculated as follows:

$$W'(x) = \frac{W(x) - W_{\min}}{W_{\max} - W_{\min}},$$

where $W(x)$ is the score for a given oligonucleotide x, $W_{\min}$ is the sum of the smallest weights at each position and $W_{\max}$ is the sum of all highest weights at each position. In order to decide when a certain oligonucleotide is a "putative hit", a threshold for the normalized score is commonly used. This threshold can either be fixed (e.g., 0.8), it can be different for each PWM, and it can be different for the complete PWM and for a well conserved core of the PWM. Such PWM-specific thresholds are often calculated by comparing the number of hits of the PWM in promoter regions with the number of hits in second exons. Examples of implementations of such a PWM scoring method are Signal Scan [237], Matrix Search [63], MatInspector [241], and Match [163].

Based on random similarity, a PWM can have dozens of instances in each kilobase of genomic DNA because of the fact that TFBSs are so short and imprecise. Many of these consensus matches do not bind protein *in vivo* and have no influence on transcription. Identifying the binding sites that actually bind protein requires either biochemical and experimental tests, or more sophisticated computational strategies. The most commonly used strategies—and central to this dissertation—are the detection of over-represented TFBSs in co-regulated genes (or gene batteries) and phylogenetic footprinting. Both will be described further in this chapter (Sections 2.5 and 2.7) and in several other chapters.

2.4.4 Transcription factors

Transcription factors can bind to DNA via their DNA binding domain and together with transcription co-factors (TCF) they form complexes at particular DNA locations that, through protein-protein interactions with the basal transcription apparatus, influence the frequency with which the polymerase II complex initiate transcription at the transcription start site (TSS).

The most common molecular functions in the human genome are in fact the transcription factors and proteins involved in nucleic acid metabolism [173, 311]. As summarized by Wray et al. [325], most TFs belong to gene families, whose sizes differ considerably among genomes (Table 2.1). The reasons and functional consequences of these differences are not understood. The Zn-finger (C2H2 type) is also the most common InterPro domain in the human genome according to the top 40 InterPro domains in the Ensembl database (Release 18, http://www.ensembl.org).

Table 2.1: Size of selected transcription factor families in three Eukaryotes.

Transcription factor family	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Homo sapiens*
Homeodomain	109	148	267
Nuclear receptor	183	25	59
Zn-finger	437	357	706
Runt domain	2	4	3
Basic HLH	41	84	131
Paired box	23	28	38
Myb	17	18	32

Source: [311, 173, 325]

Most TFs contain several of the following domains: (1) *DNA binding domains*, for example homeodomain, paired-box domain, Zn-finger; (2) *Protein-protein interaction domains*, for example the pentapeptide motif of homeodomain proteins; (3) *Trafficking signal domains*, for example a nuclear localization signal; (4) *Ligand binding domains*, for example nuclear receptor family members can have steroid hormone binding domains.

2.4.5 Transcription co-factors

The transcription factors that can bind to the unique array of *cis*-regulatory sites of a gene are—at least in eukaryotes and especially in metazoans—not enough to directly instruct the BTA to initiate RNA synthesis at a specific core promoter. TFs interact with multiple types of co-factors in both positive (activation) or negative (repression) ways. There are several classes of co-factors with different properties: factors that modulate DNA binding, for example by imposing structural effects on the domains of the TFs; proteins that interact with the BTA; and chromatin-modifying activators and repressors [178]. The interaction between clustered TFs and co-factors generally results in a large protein complex, sometimes referred to as enhanceosome [58], that is capable of communicating with the BTA through protein-protein interactions to influence transcription (see also Figure 2.2.B).

2.4.6 *Cis*-regulatory modules

A module is operationally defined as a cluster of TFBSs that produces a discrete aspect of the total transcription profile. The most common terms in the literature are enhancer and silencer. A single module typically contains about 6 to 15 binding sites and binds 4 to 8 different TFs [14, 78]. As stated by Wray et al. [325], a single module may carry out one or a combination of the following: (1) *initiate transcription*, often in a highly specific manner such as within a single cell type or region of an embryo or at a specific time point during development; (2) *boost transcription*; (3) *mediate* signals from outside the cell, by binding a TF that either contains a receptor for a hormone or that is post-translationally modified by a signal transduction system; (4) *repress transcription*, again in a highly specific manner; (5) *restrict the effect of another module* to a single basal promoter; (6) *selectively "tether" other modules*, by bringing them into proximity with a single basal promoter; or (7) *integrate* the status of other modules by influencing transcription differently depending on what proteins are bound elsewhere [330, 78, 325].

A well-known example of a module is the human interferon-β (*IFN-β*) enhancer, which drives transcription of the *IFN-β* gene in response to viral infection [281]. The presence of each TFBS and its precise arrangement within the module are critical for the various regulatory proteins to assemble through cooperative interactions into an enhanceosome (see Figure 2.7).

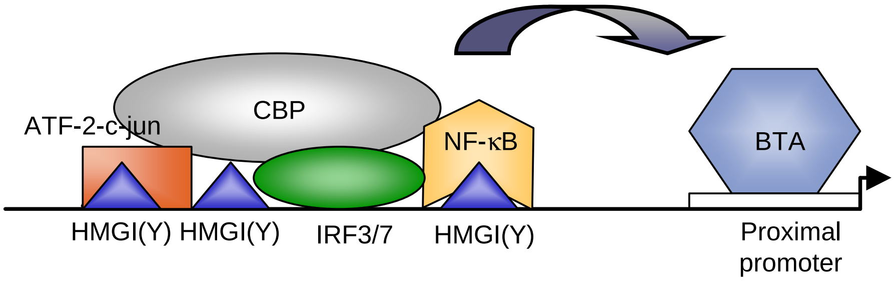


Figure 2.7: *Cis*-regulatory modules and enhanceosomes: the *IFNβ* enhanceosome assembled on the *IFNβ* module [72]. Multiple transcription factor binding sites are clustered within a sequence region of less than 100 bp, upstream of the *IFNβ* gene. Upon viral infection, specific binding (see 2.4.3) of the transcription factors IRF, NF-κB, and HMGI(Y) to these binding sites, and their interaction with each other and with the co-factor CBP, creates an enhanceosome that can interact with the basal transcription apparatus (BTA), looping out the intermediate DNA and initiating transcription of *IFNβ*.

Autonomy of modules

Another example of a well studied module is the *eve* (*even-skipped*) stripe 2 enhancer. This module is responsible for a discrete part of the expression of *eve*, a "pair-rule" gene, namely its localized expression in the second of seven alternating (on/off) "stripes" along the anterior/posterior axis of the embryo. From the stripes the full complement of body segments is ultimately generated (see for review [249]). The pair-rule genes that are expressed in stripes are among the genes that initiate the process of metamerization early in the embryonic development. The expression of *eve* in the seven stripes is controlled independently; that is, different modules control the expression of different stripes. The *eve* stripe 2 enhancer is roughly 700 bp long and contains multiple binding sites for four transcription factors. The Bicoid and Hunchback proteins are broadly distributed activators, and the boundaries of the stripe are sharpened via repression by the Giant and Krüppel proteins (see Figure 2.8). The other stripe modules are bound by different combinations of TFs found at other positions in the embryo [78, 59].

When the stripe 2 enhancer is placed in a construct with the *lacz* gene as a reporter, the endogenous stripe 2 expression is perfectly mimicked; the stripe 2 enhancer works autonomously.

Arrangement of binding sites

The stripe 2 enhancers found in different species of *Drosophila* are of similar size (750-950 bp) and bind the same regulators. But the precise arrangements and affinities of the TFBS in the modules are not the same. Despite these differences, a module from one species directs the proper pattern of gene expression (i.e., in stripe 2) when introduced into another species. Thus, there is more than one arrangement of TFBS that can bind the appropriate activators and exclude repressors to activate the *eve* gene in that region of the embryo. The pattern of stripes is slightly different in more distant relatives of *Drosophila* and this has important morphological consequences. This is concluded by Ptashne [238] as: "Nature can readily throw up functional variants for selection to work on" (see Section 2.7).

There are other cases where the arrangement of binding sites *is* important, and this has mostly to do with specific synergistic or antagonistic actions and related structural requirements for interactions among TFs, as in the IFN-β enhancer. Here, the individual factors have no effect on transcription, but in combination they produce a strong effect. Such interactions involve binding sites that typically lie no farther apart than the size of the proteins that they bind (in practice, up to a few tens of base pairs apart). Some interactions are precisely phased to lie on the same side of nucleosomes ($\sim$40-bp multiples) or completely decondensed DNA ($\sim$10-bp multiples) [183].

Another type of specific arrangement are overlapping or adjacent binding sites for different factors, as in the *eve* stripe 2 enhancer. When a repressor is present, active, and bound to its cognate site, an activator can no longer bind and transcription is off, at least in case the repressor has a higher binding affinity for its site than the activator or in case the repressor concentration is higher than the activator concentration. A consequence of the latter is that concentrations of factors that interact with adjacent or overlapping sites can

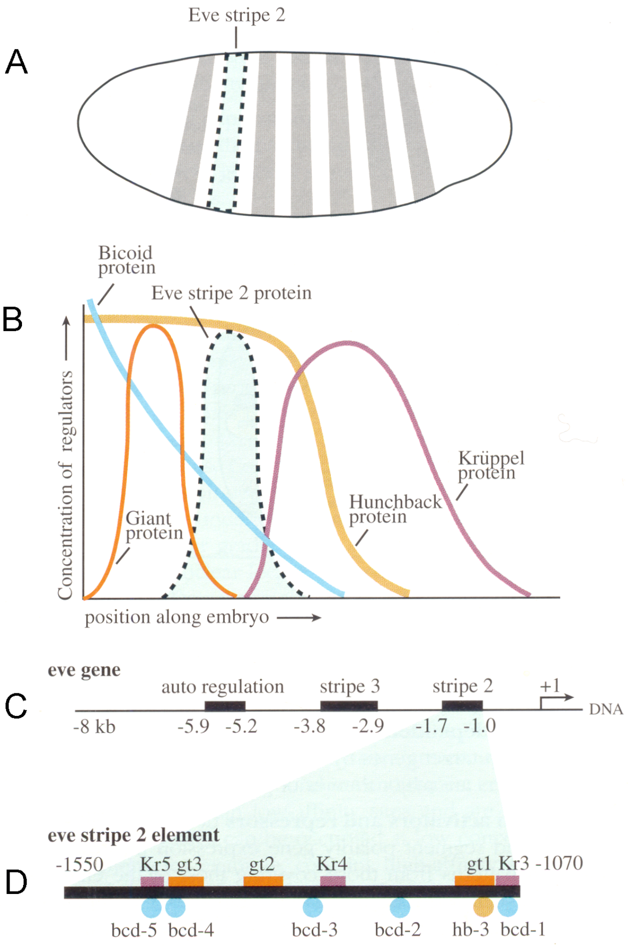


Figure 2.8: Autonomy of *cis*-regulatory modules. (A) The *even-skipped* stripe 2 enhancer (D) controls the expression of the *eve* protein within the second segment polarity stripe of the *Drosophila* embryo. There are functional binding sites for four different TFs, two activators (Bicoid and Krüppel) and two repressors (Giant and Hunchback). (B) The spatial expression of these TFs and their relative concentrations determine the expression profile of *eve*. (C) The expression of *eve* in the other stripes is controlled by other independent enhancers. Taken from [59].

have a significant impact on transcription.

Modular *cis*-regulatory information processing

A *cis*-regulatory module can be seen as a control device that becomes active when the TF's for which the module contains binding sites are present and active. Each module is an information processing regulatory device: it integrates multiple, diverse inputs (e.g., four factors for the *eve* stripe 2 enhancer) and produces a single, scalar output, namely the rate of transcriptional initiation. As mentioned in [325], one could draw an analogy with a neuron that receives input from many sources and integrates the inputs into one output, namely how often it fires. Figure 2.9 shows a generalized system with two distal modules and one proximal module.

A well studied example is the *cis*-regulatory system of the *endo16* gene of the sea urchin *Strongylocentrotus purpuratus*. Within the 2.3 kb upstream of the TSS, six modules have been found that carry out discrete functions. All together they contain 34 TFBS for 13 different transcription factors and co-factors. The collaboration among modules in this system results in complex logical structures as can be seen in Figure 2.10.

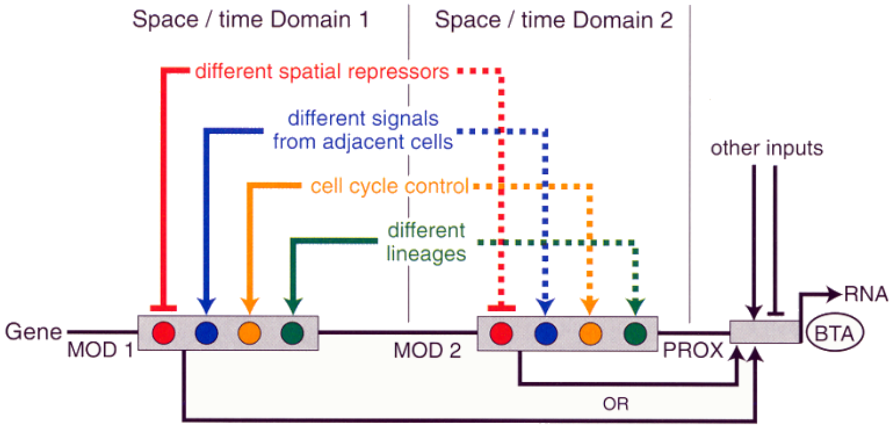


Figure 2.9: Information processing by modules. An imaginary set of distal modules showing the integration of different signals in two independent space/time domains and the communication with the proximal module. Taken from [78].

Computational detection of modules

The major bottleneck in the study of gene regulatory networks is that the short *cis*-regulatory modules lie within the expanse of genomic DNA that surrounds each gene—flanking sequence upstream and downstream, and introns. Conventionally, modules have been found by building and testing expression constructs

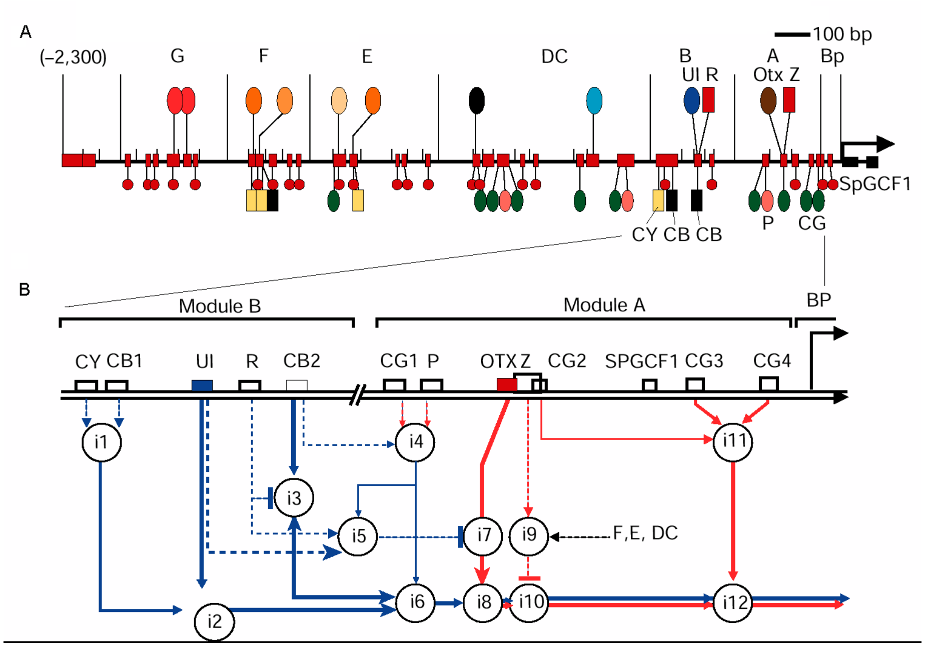


Figure 2.10: Examples of *cis*-regulatory modules. (A) The 2.3 kb upstream region of the *endo16* gene of *S. purpuratus* contains six modules (A-G, Bp=proximal promoter) with 34 binding sites in total for 13 different TFs. They carry out discrete functions during sea urchin development. (B) The regulatory logic or computational model for modules A and B is shown. The graph is accompanied by a set of rules, like IF($P = 1$ AND $CG1 = 1$) THEN $\beta = 2$ ELSE $\beta = 0$ and so on. From [330, 147].

that contain successive deletions of the DNA flanking the gene or by other methods like *DNAse hypersensitivity studies*, *location analysis*, and *electrophoretic mobility shift assays*. These approaches are labor-intensive and they are not suited for large-scale analysis of complete GRNs. Help from the bioinformatics side is therefore needed.

Literature on computational module detection has only begun to appear during the course of this work. I propose nine categories of module detection, and combinations of these approaches are also possible.

1. Detection of *single TFBSs* (they reside in modules). Single TFBS detection has been described extensively earlier in this chapter. Once a real TFBS is identified, flanking regions can be analyzed to detect other module elements. To our knowledge this approach has not been described in the literature. Chapter 4 contains a case study analyzing gene regulation in desmoid tumors using this approach.

2. Detection of *dense clusters of unknown TFBSs*. Marsan and Sagot [206] use suffix trees to detect motif co-occurrences with conserved spacing be-

tween them in a complete genome, and they used it to detect promoters in bacterial genomes (as a combination of the -35 site and the TATA box). Argos [242] detects genomic sequence windows that have multiple occurrences of one motif by testing for statistical over-representation of all motifs of a certain length.

3. Detection of similar genomic sequence regions as in a training set of known modules using *classification methods*. Logistic Regression Analysis (LRA) combined with phylogenetic footprinting (here the identification of conserved sequences between human and mouse) was applied in [316] and [168] to find modules in the human genome that control muscle- and liver-specific genes respectively. The coefficients are determined by a maximum likelihood procedure to maximize the discrimination between a positive and a negative training set. Fisher Kernel Support Vector Machines (SVM) have been applied to predict promoters (modules) based on combinations of motifs and the spacing between them in yeast [227].
4. "Module scanning", or the detection of a *joint occurrence of a known combination of known TFBSs* within a confined sequence window in a single sequence or on a genomic scale:
   - In the "sliding window approach" a program detects genomic sequence windows where all or most of the *given* PWMs or consensus sequences have predicted instances, without regard to order or spacing. Wagner [313] developed a statistical measure starting from the Poisson distributions of the instances of the individual PWMs, to detect clusters of TFBSs of two or more given PWMs and tested it on the yeast genome. CIS-ANALYST [26] examines sequence windows of length *wind_size*, retaining only those containing at least *min_sites* binding sites, and collapsing all overlapping windows into a single cluster. FlyEnhancer [205] works similarly and was used to detect clusters of Dorsal binding sites in *Drosophila*. Halfon et al. [132] and Rebeiz et al. [244] developed respectively cooccur_scan.pl and SCORE, also to find clusters of predicted binding sites for multiple TFs in *Drosophila*, and they used Monte Carlo simulations to test whether certain combinations occur less than expected (and thus could have functional consequences). Ahab [242] computes an optimal probabilistic segmentation [53] of a sequence S into binding sites and background (modeled by a local higher-order Markov model) for a fixed set of PWMs. MSCAN [161] is a module scanner that calculates the combined statistical significance (an upper bound for the p-value) of the instances of a given set of PWMs in a window.
   - Several other module scanners use a hidden Markov model (HMM) implementation that take distance constraints between TFBSs into account: the first implementation was done by Crowley et al. [74], and other flavors are Cister [115], COMET [116] with statistical signifance (E values), MCAST [18], and Stubb [265].

Also see our own algorithm ModuleScanner in Chapter 5 and related work.

5. Detection of a *joint occurrence of newly discovered motifs by their over-representation in a set of co-regulated genes.* See Section 2.5.2.
6. Detection of a *joint occurrence of newly discovered motifs by the conservation in orthologous regulatory sequences.* There exist no methods of this kind to our knowledge.
7. Discovery of a *joint occurrence of a new combination of known TFBSs by their over-representation in a set of co-regulated genes.* See Chapter 5 for our algorithm ModuleSearcher and related work.
8. Discovery of a *joint occurrence of a new combination of known TFBSs by the conservation in orthologous regulatory sequences.* There exist no methods of this kind to our knowledge.
9. Detection of *conserved non-coding sequences.* A conserved non-coding sequence block is assigned a putative regulatory role and TFBSs are detected thereafter. See Section 2.7.3 and the ModuleSearcher in Chapter 5.

2.4.7 Putting it all together

Despite many advances in discerning interactions between factors, the detailed mechanisms by which the transcription initiation rate of an individual gene is determined remain poorly understood.

Transcription factories

Lemon and Tijan [178] discuss several possible models. The model with the most supporting evidence is based on cytological studies that suggest that some co-regulators and components of the general transcription machinery may be segregated in the nucleus [248] and exist as organized compartments that are also referred to as transcription factories [152]. The authors propose a model that integrates nuclear compartmentalization and transcription. The first step involves regional chromatin remodeling events by complexes associated with the nuclear matrix together with primary activators with some limited DNA target accessibility. After initial remodeling events, a distal module can become accessible for binding factors that then recruit other appropriate chromatin remodeling factors, including histone acetyltransferases (HAT). Secondary chromatin remodeling including nucleosome shifting could thereafter provide access of other activators to other modules. Subsequent cooperative interactions between sequence-bound TFs and associated TCFs, and perhaps mobilization of the nuclear matrix, could promote the directed association of the promoter with a transcription factory where pre-assembled parts of the BTA are present. Following the selective recruitment to a localized factory, cooperative interactions between TFs and TCFs could initiate promoter binding and subsequent events leading to the formation of an active initiation complex. Re-initiation can follow by concerted signals from TFs that dynamically interact with target modules and co-factors. Subsequent events might signal for promoter de-activation. It is even postulated that in the same factories RNA processing and even initial translation (proofreading?) is facilitated [69, 152, 68]. In this model a gene, with its array of *cis*-regulatory modules, competes with other genes for access to a transcription factory. If all necessary TFs and TCFs are present and activated, a functional enhanceosome can be formed that, through interactions with proteins in a factory, can bring the core promoter in close proximity with the BTA components in the factory, thereby looping out the sequence in between the respective module en promoter. The efficiency of this process then determines the initiation rate.

Transcription factories can also explain some aspects of chromosomal location effects of gene expression (for example in *Drosophila* [270]): co-localized genes can be co-expressed if whole chromosomal domains are exposed to such factories.

Enhanceosome versus information display

Apart from the discussion whether or not the transcription factory model is appropriate, Kulkarni and colleagues [169] recently proposed that modules work by information display as opposed to the enhanceosome model. In the enhanceosome model, the module serves as an information processing center, receiving inputs from multiple TFs that bind it. The enhanceosome creates a stereospecific interface for docking with and recruiting the BTA. Here the module serves as a molecular computer, resolves multiple inputs and provides a single output to the BTA. With such a module, the target gene would be activated only upon the assembly of a complex, providing a precise on/off binary switch. Graded responses from such an element could be achieved by varying the stability of the entire complex, possibly in response to activator concentrations (this can be translated into various probabilities of getting access to a transcription factory depending on factor concentrations). In the "information display module" subelements can display contrasting information, which is then interpreted by the BTA. The BTA "samples" discrete regions of the enhancer each composed of a small number of TFBSs, either iteratively or simultaneously. Successive or multiple interactions with the BTA, and the biochemical consequence of these interactions, would dictate the overall output of the enhancer [169].

2.5 Gene batteries

Wray et al. [325] demonstrate with a simple calculation that most eukaryotic transcription factors must bind to the promoters of many downstream genes. They describe it as follows. Eukaryotic genomes contain on the order of 5000 to 50000 genes, only a small fraction of which encode transcription factors (see 2.4.4). Because the expression of all genes requires that transcription factors bind to their promoters, and because most promoters contain binding sites for at least five different transcription factors (and often many more), transcription factors must on average interact with the promoters of tens to hundreds of genes [325]. Such a set of coregulated genes is called a gene battery or a regulon and they will be of great importance in this dissertation because if the output of a GRN can be measured for all genes at the same time, then gene batteries can be selected by grouping together the genes with similar expression profiles, and since the *cis*-regulatory elements around these genes cause their expression profile, we should be able to find the same regulatory elements in all the genes of a battery. So a gene battery is actually defined as a set of genes that share a given set of regulators [78].

2.5.1 Genome-wide expression analysis

Microarray technologies have revolutionized biological research in the past few years in the sense that instead of studying genes one by one, thousands of genes can now be studies simultaneously.

Microarray technology

There are two types of microarray platforms depending on the method of nucleic acid deposition on the chip surface. One technology involves slides with robotically spotted probes that can be PCR products amplified from cDNA libraries (cDNA microarrays) [254, 189] or oligonucleotides (∼50 to 70mers). In most experiments one compares two samples, the test and the reference sample. These are labeled with two different fluorescent-dyes and co-hybridized to the same array (i.e., two-channel measurements). The ratio between the two dyes indicates the relative abundance of a gene in these two samples. The other technology is also used in computer chips fabrication, and is commercially available from Affymetrix^{TM}. It involves the photolithographic synthesis of short oligonucleotides (∼25-mers) [186]. The two samples under comparison are labeled with the same dye and individually hybridized to different arrays.

Extraction of raw intensity data

After the microarray hybridization experiments, laser scanning of the slides produces images from which the raw intensity data has to be extracted. The four basic steps in image acquisition and analysis are summarized in a review by Leung et al. [181]: (1) *scanning*; (2) *spot recognition* or gridding, which locates each spot on the microarray image; (3) *segmentation*, which differentiates the pixels within a spot-containing region into foreground (true signal) and background; and (4) *intensity extraction*, which calculates the foreground (signal) and background intensities from the pixels after the segmentation process.

Data preprocessing

For two-channel microarray data, the background corrected intensities of the test sample are divided by the background corrected intensities of the reference sample to obtain expression ratios. These are log transformed so that (1) upregulated and downregulated values are of the same scale and comparable [240] and (2) multiplicative effects are converted into additive effects that are easier to model [75]. Consistent sources of variation should—if present—be removed in a *normalization* step such that, for each gene, the measured value reflects the mere expression level as caused by the condition tested [202, 203]. For cDNA microarrays these consistent sources of variation include array, dye, condition and spot effects. Array effects refer to the differences in hybridization efficiency between different slides. Condition and dye effects reflect differences in respectively mRNA isolation and labeling efficiencies between two distinct samples while spot effects refer to the difference in amount of cDNA spotted on the array. Global normalization assumes that only a small fraction of the total number of genes on the array alters its expression level and that symmetry exists in the number of genes that is upregulated versus downregulated. Remark therefore that the assumption of global normalization applies only to microarrays that contain a random set of genes and not to dedicated arrays. Under the assumption of global normalization the average intensity of the test genes should be equal to the average intensity of the reference genes. Based on the hypothesis of global normalization, for the bulk of the genes the $\log_2(\text{test}/\text{reference})$ ratio should equal 0. Normalizing the data consists of finding the right transformation factor that allows centering the $\log_2(\text{test}/\text{reference})$ for the bulk of the genes around zero. Linear normalization assumes a linear relationship between the measurements in both conditions (test and reference) and uses a constant transformation factor that can either be the mean or median of the log intensity ratios or a regression factor as determined by linear regression. However, the relationship between dyes can depend on the measured intensity. More sophisticated normalization techniques are needed to remove such intensity-dependent dye effects (see Chapter 3). An alternative way of preprocessing cDNA microarray data is to use an ANOVA model with one factor for each of the above mentioned effects [164].

Exploratory data analysis

Exploratory data analysis essentially aims at finding genes with similar expression profiles, using the expression data only. Commonly used techniques are principal component analysis (PCA) or singular value decomposition (SVD) [9] for dimensionality reduction, and several algorithms for clustering like hierarchical clustering [98], K-means [292] and self-organizing maps (SOMs) [289].

Detecting differentially expressed genes

Statistical methods for the detection of differentially expressed genes are different for the comparison of two conditions and for the comparison of more than two conditions (see [75] for a review). For two conditions, the simplest method is the "fold change" cut-off where all genes that differ by more than an arbitrary cut-off value between the conditions are taken to be differentially expressed. Of the statistical methods, the t test is the simplest, where the standard error can be computed per gene (in replicated experiments) or by combining data across all genes, although the latter is effectively a fold-change test. For small sample sizes, the error variance may be hard to estimate and may be subject to erratic fluctuations. There are several modifications of the t test that can obtain more stable estimates of the error variance, such as SAM [303] and the regularized t test [19]. If p-values are calculated from the t statistic, one should account for multiple testing (e.g., using Bonferroni correction) since false positives (type I errors) may accumulate when thousands of tests are conducted. Alternatively, the false discovery rate (FDR) can be used [246]. In Chapter 3 we describe some simple correlation measures and ANOVA approaches to detect differentially expressed genes in an experiment with more than two conditions.

2.5.2 Detecting common motifs in gene batteries

All genes of a gene battery are co-regulated and thus their *cis*-regulatory systems share common TFBSs (motifs) or complete modules. There is a plethora of techniques to discover common motifs in a collection of the regulatory sequences of a set of co-regulated genes. But the construction of a good sequence set is not trivial. One often has to rely on sets of co-expressed genes that are *putatively* co-regulated, such as clusters of gene expression profiles from microarray data. The motif detection algorithms therefore have to take into account that some of the genes might be secondary response genes (or genes that have accidentally the same expression profile) without the shared motif. Another source of noise is the selection of putative regulatory sequences for the selected genes since promoters and modules cannot easily be identified. Promoter or TSS prediction algorithms can be used, sequences upstream of the translation start can be used (possibly containing long 5'UTRs), or sequences can be selected from databases with TSS annotations (Ensembl, DBTSS, PromoSer). Many of the motif discovery algorithms have been tested on sets of yeast genes where promoter sequence selection is less a problem. Since yeast has no 5'UTR, and since the intergenic regions are short, the regulatory regions commonly lie just upstream of the translation start codon (ATG). Therefore, when selecting $\sim$800bp upstream of the TLS most regulatory regions are represented in the sequence set. When applying the same algorithms on higher eukaryotes, special attention should be given to sequence selection and some approaches are given in the results chapters.

Once a set of sequences is constructed, there are basically two possible approaches for motif discovery (for review see [276]). The consensus approach uses a consensus sequence as motif model (i.e., to represent the common motif), and has been applied to sets of co-expressed yeast genes (based on microarray data analysis). van Helden and colleagues [307] used the binomial formula to calculate p-values for the number of matches of the 4096 possible hexanucleotides. Hexanucleotides with significant p-values are statistically over-represented and those may be the elements that are responsible for the common gene expression profile. Brazma and colleagues [46] used suffix tries for the discovery of patterns in a sequence set.

The second approach uses a position weight matrix as motif model. Consensus [279, 145] uses a greedy algorithm with I_{seq} (see 2.2) as the criterion for identifying the best motif model (i.e., having the best alignment of potential sites). MEME [17] uses Expectation-Maximization (EM) in which, in each iteration, a weighted alignment is obtained from the instances of the current motif model, and from which a new motif model (PWM) is derived for the next round until convergence. A Gibbs sampling algorithm using a similar sequence model was developed by Lawrence and colleagues [176]. This algorithm iterates between refining a description of the motif and aligning sites in the sequences that may represent instances of the motif. AlignACE [150] and MotifSampler [298] are other Gibbs sampling implementations. MotifSampler uses species-specific higher-order background models (Markov models) to improve the robustness of the algorithm to noise (i.e., lower the variability of the outcome of the algorithm). A background model is a mathematical representation of the areas of the sequence that do not contain motifs. The better the representation of the background, the higher the efficiency of detecting true positive motifs in the presence of noise [297, 204].

Instead of starting with a cluster of microarray data, the expression data can be used more intensively during the motif search. Bussemaker et al. [54] used a regression model in which upstream motifs contribute additively to the log-expression level of a gene to predict statistically significant motifs. Caselle et al. [61] first clustered genes according to over-represented motifs and then compared the expression of the cluster with genome-wide averages to detect significant differences.

Combinations of motifs in gene batteries

Pilpel et al. [232] used the occurrence of pairs of TFBSs in yeast sequences instead of single TFBSs and aimed at detecting significantly synergistic combinations using expression coherence scores. Co-bind [128] models the synergy between two TFs, with conserved spacing between the binding sites. A PWM is looked upon as a simple, single layer, neural network (perceptron). Two perceptrons are combined to model cooperativity. The detection of spaced dyads [309] can in fact also be seen as a co-occurrence of two motifs with a conserved spacing.

The above mentioned technologies have been successfully demonstrated mainly in prokaryotes and yeast. One of the goals in this work was to develop a method to detect over-represented combinations of TFBSs in sets of co-regulated *metazoan* genes. Independent from our work, and at the same time, Elkon and colleagues [100] also demonstrated that similar approaches can be applied for human genes. They revealed eight TFs whose binding sites are significantly over-represented in promoters of genes whose expression is cell-cycle dependent. Later, more regulatory sequence analysis studies in metazoan genes appeared (see further).

2.6 Gene regulatory networks

Gene regulatory networks (GRN) consist of genes and of the linkages between genes that are implemented as *cis*-regulatory systems governing the expression of the genes. Gene batteries as described above are an example of a GRN where the linkages are between the genes encoding TFs and for example genes encoding differentiation proteins. The genes in the battery are called peripheral network elements, and the TF-genes are internal network elements [14]. In general however, there can also be multiple linkages between the internal elements. Figure 2.11 shows an imaginary developmental GRN with three internal genes and six peripheral genes. There are downstream connections from the internal genes to the peripheral genes, an autoregulatory connection and two connections amongst the internal genes. Autoregulatory connections are known in many genes that encode TFs. The first complex real GRN that was deciphered and where all linages were experimentally validated (using perturbation and expression data) was the GRN for endomesoderm development in sea urchin [79], see Figure 2.12.

Note that genes encoding components of signal transduction pathways causally upstream of genes encoding TFs are not part of GRNs because they involve chains of protein-protein interactions and thus describe regulatory connections beyond those immediately represented in the genomic DNA sequence.

Reverse engineering gene networks

Computational methods exist to model GRNs *in silico* and also to computationally construct GRNs from data. For gene network modeling one either uses the Boolean method (expression is ON or OFF), or the dynamical system method where ordinary differential equations describe the rates of change of mRNA or protein concentrations. Terms in these differential equations describe how gene expression rates are modified by changes in the levels of transcription factors or other effector molecules. Examples of such modeling can be found in a review by Goldbeter [123] describing models for the complex oscillatory behavior of gene expression in circadian rhythms [123, 268].

For GRN reconstruction, mostly microarray data has been used since it should theoretically be possible to reverse engineer the architecture of a GRN by measuring the network outputs, namely mRNA levels (for review see [326]). SVD analysis has been used [328] to find underlying patterns or modes in expression data, with the intention of linking these modes to the action of transcriptional regulators. Friedman and co-workers have learned a Bayesian network from expression data [114], and Hartemink and colleagues used Bayesian networks to pick one of several competing models that best fits expression data [137]. Segal et al. developed a method to identify coregulated gene modules from large-scale gene expression data [261]. Reconciliation of expression data with known network structures has been done for *E. coli* [144].

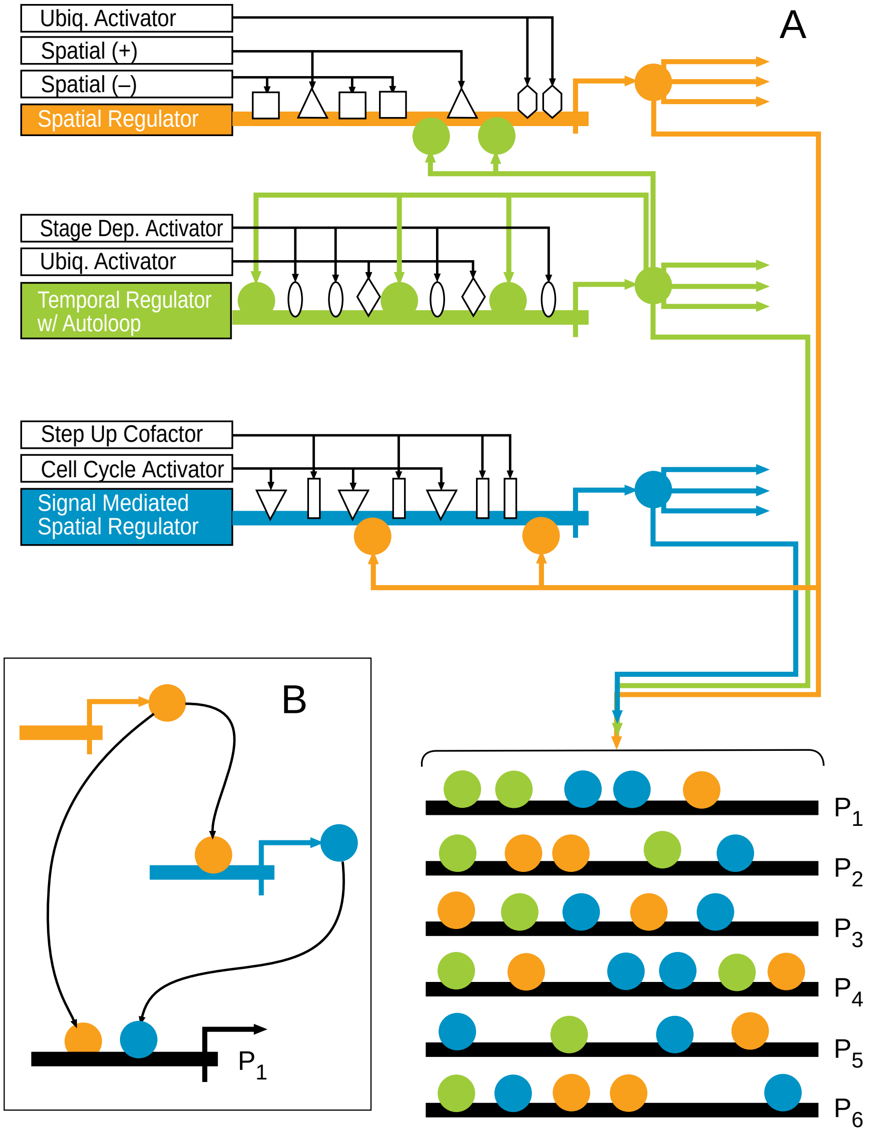


Figure 2.11: Gene regulatory networks. (A) An imaginary GRN taken from [14] with linkages between three transcription factors and a gene battery (P1-P6). The genes in the battery share similar (not identical) *cis*-regulatory modules with binding sites for the three colored TFs. The TFs are themselves regulated by other TFs (open boxes) that can be ubiquitous, stage dependent, or spatial activators, or spatial repressors. Lastly, there are also linkages between the colored factors themselves, for example, green regulates orange and orange regulates blue. (B) A single relationship between the orange TF and two genes it is controlling, namely the blue TF and the peripheral P1 gene. P1 is also directly responsive to the blue TF.

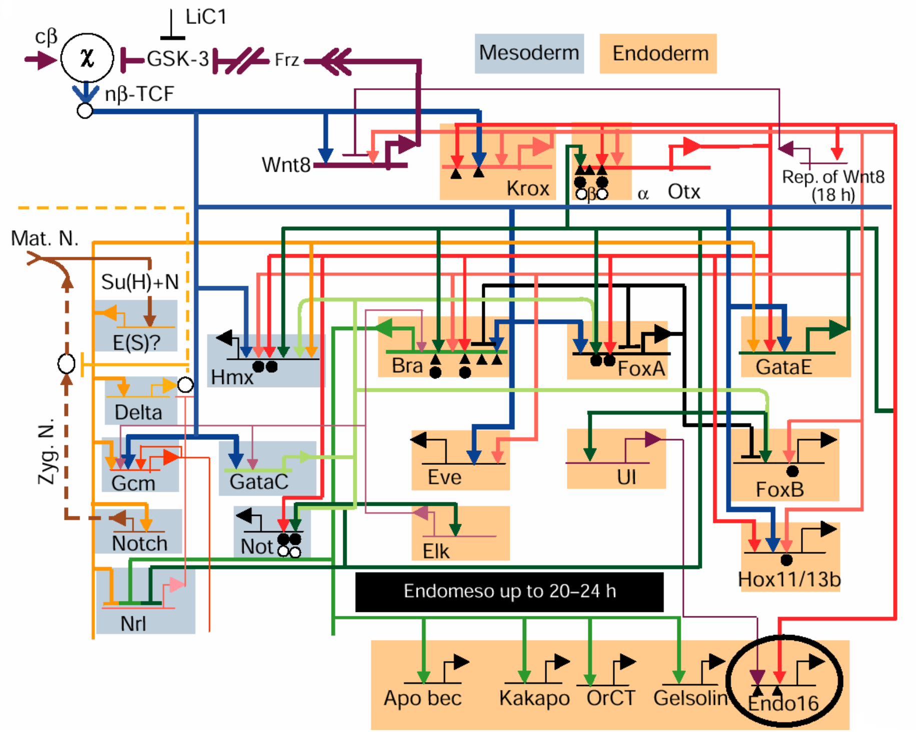


Figure 2.12: Gene regulatory networks. Part of a real gene regulatory network that controls the development of endomesoderm in the sea urchin embryo [78, 147]. Genes are shown as horizontal black lines with an arrow indicating the TSS. Most genes are internal transcription factors regulating each other. Note that most TFs receive inputs from multiple other TFs rather than from a single TF. At the upper left (the WNT pathway), maternal cytoplasmic β-catenin (cβ) is nuclearized (nβ-TCF) by the nuclearization system χ. The five genes at the bottom are peripheral differentiation genes like the encircled Endo16 gene.

2.7 Regulatory evolution

Gene regulatory networks can change rapidly over evolutionary time [147, 59, 325, 78]. Hood [147] illustrates this point by two striking examples. About 550 million years ago in the Cambrian era, changes in GRN's that happened over a short period of time (∼10-30 million years) caused an explosion of metazoan organisms that resulted in the immense morphological diversity in animal species that we see today. A second example involves the divergence of human from its common ancestor with chimpanzees (about 6 million years ago) that may be explained by rapid changes in the regulatory networks that drive brain development.

2.7.1 Mutations in *cis*

Mutations affecting transcription fall into several distinct classes, as stated by Wray et al. [325]. (1) Small-scale, local mutations—single base substitutions, small indels, and changes in repeat number—can modify, eliminate, and generate binding sites and alter their spacing. (2) New regulatory sequences can be inserted in the neighborhood of a gene through transposition. (3) Retroposition may assemble new *cis*-regulatory sequences. (4) Gene duplications may fragment or recombine promoter sequences. Gene duplications that persist are frequently followed by divergence in expression and may be followed by loss of complementary regulatory modules. (5) Gene conversion can spread regulatory elements within a gene family (e.g., beta and gamma globins in human). (6) Sequences that have no prior function in regulating gene expression can become regulatory elements.

2.7.2 Mutations in *trans*

Mutations in *trans*—loci encoding transcription factors that interact with *cis*-regulatory sequences of a gene—can also be the cause of differences in gene expression. These can be regulatory mutations in *cis* of the TFs, mutations that affect the DNA binding domain of the TF, and mutations that affect the protein-protein interaction domain of the TF. For a review see Wray et al. [325].

2.7.3 Phylogenetic footprinting

Although such *cis*-mutations do occur, and although the mutation rate can be higher than in coding sequences, it should certainly be lower than in nonfunctional sequences. Evolution causes nonfunctional DNA sequences to diverge, but functional sequence elements are conserved because of selective pressure. By comparing sequences from different species, the conserved sequences can be detected. This fact has mostly been exploited in gene finding methods, but it is a well-accepted view that also regulatory elements are among the conserved sequences. This is supported by the fact that many experimentally determined regulatory sequences have been shown afterwards to lie in regions that are conserved between species. Only recently has sequence conservation been used as a mean to detect *cis*-regulatory elements, an approach called *phylogenetic footprinting* (PF) [288]. PF is a way to increase the signal to noise ratio in the detection of TFBS, just like the use of coregulation in gene batteries was. An advantage of PF is that it can be applied on a single gene, as long as enough orthologous sequences are available. The use of sequence comparisons for regulatory sequence analysis has been reviewed recently by Hardison [134], Pennacchio and Rubin [230], Ureta-Vidal et al. [304], Zhang and Gerstein [333], and Bulyk [49]. There are two flavors of PF, both are described in the next two sections.

First approach: direct motif detection

One approach is to find conserved motifs within DNA fragments that have been experimentally shown to harbor promoters or modules. Disadvantages of this approach are that the discovery of motifs in unaligned orthologous sequences requires relatively large sets of orthologs to make a clear distinction between conserved and nonconserved elements. Only sufficiently conserved motifs can be discovered and only relatively short regions can be analyzed, as performance of the approach decreases dramatically for longer sequences [22]. Using several tens of kbs around a metazoan gene is not feasible. Although in principle the same motif finding algorithms as for gene batteries could be used for orthologous promoters, PF algorithms can use an extra feature, namely the evolutionary distances or relationships among the involved species. A mutation in a conserved motif can then for example be penalized more if it occurs between human and mouse than if it occurs between human and fly. Such a technique is implemented by Blanchette and Tompa [35, 36] and is called FootPrinter. FootPrinter is a dynamic programming algorithm for the computation of the parsimony score of a fixed set of aligned sequences. The inputs to the algorithm are a number of homologous sequences, the phylogenetic tree relating them, the length k of the motif sought, and the maximum parsimony score allowed. The k-mer with the lowest parsimony score is selected from all 4^k possibilities. It can deal with the absence of a motif in some sequences, and can calculate a statistical significance for each motif. FootPrinter will be used in a case study in Chapter 4 and is available within TOUCAN as described in Chapter 4.

Second approach: conserved non-coding sequences

The second and more general approach of PF is to align large genomic regions around orthologous genes to identify conserved non-coding regions (CNS), for example if over 75% nucleotide identity is observed over more than 100 bp. A CNS can then be assigned a putative regulatory role and can be further examined either computationally (e.g., detecting TFBS) or experimentally or both. A major advantage of this approach is that in principle, only a few orthologous sequences (e.g., human and mouse) are required for the analysis. The evolutionary distance between the two (or more) species that are being compared is important. For too distantly related species, the correct identification of orthologous genes can be difficult; it is difficult or impossible to find an accurate alignment; and the biological roles, binding site organization, and expression patterns are more likely to be altered [317]. Worm–human and fly–human comparisons for example, are generally not fruitful. In too closely related species like human–chimpanzee or mouse–rat, the evolutionary time of divergence has generally been too short for the intergenic sequences to diverge enough so there is too little contrast between functional and non-functional conserved regions, although some researchers are also proposing cross-species comparison between human and other primates, which has been described as phylogenetic shadowing [38]. Human–mouse (diverged ~75 million years ago), the flies *D. melanogaster–D.*

pseudoobscura, and the sea urchins *Strongylocentrotus purpuratus* and *Lytechinus variegatus* are at good evolutionary distances for CNS to module mapping, although a good choice of the species is likely to differ to some extent from gene to gene. Examples where these have been used for single genes, followed by experimental verification, are the human *BTK* gene [222], the human *IL-4 / IL-13 / IL-5* cytokine locus [191], the human *MBP* gene [104], the human *DACH* gene [221], and the human *SCL* gene [126]. Yuh et al. [329] had a 65% success rate when testing all CNSs of the *S. purpuratus otx* gene. Besides single gene analysis, the CNS-approach has also been examined on a larger scale. Wasserman et al. [317] found that 98% of the known experimentally determined TFBSs in a set of 28 skeletal muscle specific genes lie within human-mouse conserved regions. In the same study they detected the binding sites for SRF, MEF2, and MYF factors by applying Gibbs sampling on the conserved regions, while the same method applied to many kb around the genes, detected no meaningful motifs. On the other hand, Emberley et al. [101] found that only 50-75% of 315 binding sites from 30 known modules of *D. melanogaster* reside in CNSs with *D. pseudoobscura*.

Alignment algorithms There are several alignment algorithms that are suited to align large genomic regions and to detect conserved sequence blocks. Alignment programs can be divided into two types: local alignment methods and global alignment methods. The first search for highly similar regions in two sequences, where the regions of similarity are not necessarily conserved in order and orientation. BLAST-like methods like BLASTZ [260] work by first finding very short common segments between the sequences, and then expanding out the matching regions as far as possible. Global alignment on the other hand (e.g., Needleman and Wunsch [218]) work under an extra assumption, namely, that similar regions appear in the same order and orientation in the aligned sequences. This increases the power in finding weakly conserved regions. and the order assumption tends to be satisfied when comparing sequences from related organisms [42]. Global alignment algorithms have generally only been applied to short sequences and were too slow for large genomic regions. This shortcoming has recently been addressed by several new implementations. AVID [42] works by first finding matches using suffix trees, selecting nonoverlapping, noncrossing matches as anchors, and then the regions between the anchors are aligned similarly in a recursive way until all bases have been aligned. The AVID output can be fed into the visualization software VISTA [209], and an example is shown in Figure 2.13.

Other methods with the same goal (with or without the constraint for colinearity) are BLASTZ together with the visualization tool PIPMaker [260], PromoterWise (http://www.ebi.ac.uk/Wise2/promoterwise.html), BBA [335], DBA [157], and LAGAN [48]. CONREAL [22] uses TRANSFAC-PWM matches as anchors. Example of global alignment methods to align more than two large sequences are MAVID [44] that uses AVID in a recursive way and MultiPipMaker [259].

Instead of the *ad hoc* alignment of orthologous sequences, CNSs can also be retrieved from databases with pre-computed alignments like CORG [87].

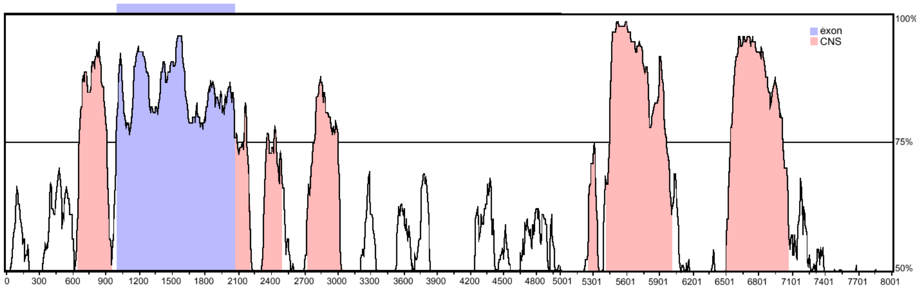


Figure 2.13: Conserved non-coding sequences. The human ATOH1 gene with flanking sequences was aligned with the mouse orthologous gene MATH1 using the AVID alignment algorithm. The plot shows the percent identity (y axis) in sequence windows of minimally 100 bp around each position in the sequence (along the x axis). Such windows with at least 75% identity are filled. The ATOH1 contains only one exon, which is also conserved. There is a conserved non-coding region immediately upstream of the exon (i.e., the proximal promoter), three immediately downstream, and another two larger regions at ~5-6 kb downstream. These two larger regions have been found to be real enhancers in mouse [142].

Chapter 3

Microarray data analysis: a case study in neurobiology

3.1 Introduction

IN this chapter, we consider several techniques to analyze microarray expression data using a case study of gene profiling in mouse hippocampal neurons during development [77]. This analysis is done in collaboration with the Laboratory of Neuronal Cell Biology of the K.U.Leuven. Therefore, more than in the other chapters, the focus is not only on the bioinformatics methods, but also on the biological interpretation of the data.

The progressive differentiation of neuronal precursor cells towards polarized, electrically active, and synaptic transmission competent neurons is a fundamental aspect of brain development. The molecular analysis of this process is difficult because of the anatomical complexity of the developing brain and the multitude of different proteins and metabolic pathways involved in this process. High-density oligonucleotide or cDNA arrays allow for the analysis of this complex process at the RNA expression level. The isolation of sufficient RNA from specific populations of cells at different stages of differentiation from the brain remains however technically challenging. One way to circumvent this problem is to isolate a certain subpopulation of neuronal precursor cells and to let them differentiate *in vitro*. A classical and well-studied example of such a cell culture system is the primary culture of hippocampal neurons as developed by Banker and collaborators [125]. Hippocampal cells are isolated from late stage embryos and can be grown for weeks. One particular advantage of this culture system is that cells at different stages of development in the embryonic hippocampus become apparently resynchronized upon isolation [106]. They then differentiate again *in vitro*, in a quite stereotypical way along five morphologically well-defined stages [90]. The neurons progressively develop neurites that differentiate into dendrites and one axon. In a later stage of the culture, active synapses are generated. The expression and subcellular distribution of proteins or RNA can be studied at any specific stage [125]. A wealth of information on almost every aspect of neuronal differentiation, particularly neurite outgrowth, neuronal polarization, and synapse formation and function, has been obtained using this system [106, 39, 40, 41, 267, 148]. Axonal growth and development of polarity is blocked by inhibitors of RNA or protein synthesis indicating changes in gene expression during the neuronal differentiation process [156]. While a number of genes has been studied in this developing system [41], no attempt has been reported yet to document the global changes in gene expression during neuronal differentiation. We used microarrays containing 21439 cDNA clones to analyze gene expression in primary hippocampal neurons differentiating in culture. In addition to the data and their analysis described in this chapter, we have made our full data set and annotation available at our web site (http://www.esat.kuleuven.ac.be/neurdiff), in a downloadable file format or via a purpose-built web application. The web application implements the functional Gene Ontology (GO) [15] annotation, and allows investigators to select sets of genes of interest and to cluster and visualize expression profiles.

3.2 Neuronal differentiation *in vitro*

For details of the biological methods we refer to [77] and to the supporting methods that are available at http://www.esat.kuleuven.ac.be/neurdiff/. Hippocampal neurons were plated at a cell density of approximately 10000 cells/cm^2, which corresponds to one hippocampus onto one dish. As mentioned above, they pass through the five stages under the right experimental conditions. This was checked by immunofluorescence using monoclonal antibodies against certain stage specific proteins, namely Mapt, Map2, and Syp. Confocal microscope pictures are shown in Figure 3.1.

Cells at 7h in culture displayed initial outgrowth of neurites (Stage 2). The future axon was identifiable in most neurons at 18h in the culture, and we chose 18h as the first of the two time points representing “Stage 3”. Only this neurite (future axon) stained for the Tau (Mapt) protein at 24h (Figure 3.1.B). Around 33h many axonal growth cones reached neighboring cells and we used these cells for the second “Stage 3” time point to document possible changes in gene expression following neurite-target interaction. Future dendrites increased their length from 72h onwards. We took the 72h and 8 days time points as early and late “Stage 4”. It is important to note that, during “Stage 4”, axons continued to grow and branch. They are more difficult to identify morphologically, because of the greater number of outgrowing and branching dendrites. Transition to the mature morphology (Stage 5) occurred gradually and was already visible in some neurons at 8d in the culture. At 9d the dendrites and cell bodies showed punctate synaptophysin (Syp) staining, representing newly formed synaptic contacts (Figure 3.1.B). After 12d in culture practically all neurons displayed mature morphology, with extensively branched dendrites and axons forming a dense mesh on the culture dish (Figure 3.1, 12d). The amount of contaminating glia cells in our cultures was very low, in accordance with previously

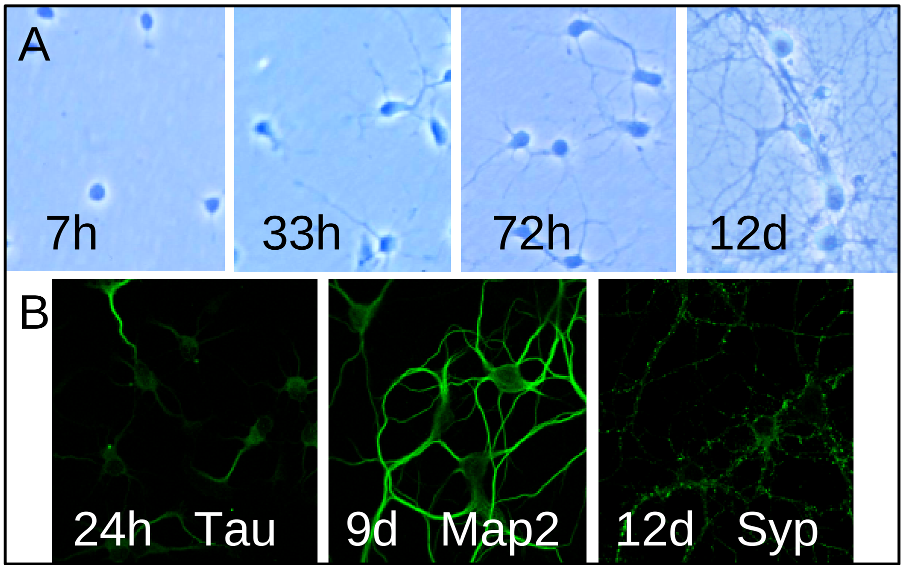


Figure 3.1: Timing of the stages of neuronal differentiation. (A) Contrast phase pictures of neurones at the indicated times in the culture. (B) Confocal pictures of neurones fixed at the indicated times in the culture and stained for the proteins Mapt (Tau), Map2 and Syp (synaptophysin).

published work [90].

3.3 The microarray experiment

Total RNA was isolated at 7h, 18h, 33h, 72h, 8 days, and 12 days from the start of the culture (the culture is started with 17-day old embryos). Total RNA from brains of newborn CD1 mice was used as a common reference for all the 6 time points. Minimum 3 independent cultures, each from one litter consisting of 10-12 embryos, were used for every time point. Typically, we pooled RNA from several neuronal cultures at the same time point of differentiation for a single hybridization. The probes were prepared according to [239]. Briefly, 5 mg of total RNA was reverse transcribed, converted to double-stranded cDNA and amplified by *in vitro* transcription, resulting in the amplified RNA (aRNA). The single stranded fluorescently labeled cDNA probe was prepared from the aRNA by a reverse transcription, in the presence of Cy3-dCTP or Cy5-dCTP.

Hybridization was done on 5 microarray slides containing in total 21,492 cDNA fragments, each spotted at two distant positions. The clone set was composed from the 8K collection of Incyte (Mouse Gem I, Incyte, Palo Alto, USA) and from the 15K collection of National Institute of Aging (NIA, HGMP Resource Centre, Hinxton, UK). The complete set can be found at http://www.microarrays.be. They represent 13606 distinct Unigene clusters (http://www.

ncbi.nlm.nih.gov/UniGene/), or 8984 entries in the Mouse Genome Informatics Database (http://www.informatics.jax.org/mgihome/), or 9502 Locuslink IDs (http://www.ncbi.nlm.nih.gov/LocusLink/).

The probe synthesis and hybridizations were repeated twice, with inversion of the dyes for the experimental samples and the reference. Such an experimental design is called a *dye-swap* experiment. After hybridization the slides are scanned and the raw intensities are extracted computationally from the resulting images (see Section 2.5). Given the duplicate spots this resulted in eight measurements per clone (four test intensities and four reference intensities) for each of the six conditions tested.

3.4 Data preprocessing

It is common practice for cDNA microarrays to represent the expression level of a gene by the relative expression of a clone that represents the gene on the chip, as compared to the expression of the same clone in the reference (whole brain) sample (see Section 2.5). Taking this ratio should remove the systematic spot variation that would otherwise be present because the cDNA concentration is not the same in each spot. For reasons pointed out in Section 2.5 we then take the $\log_2$ of the ratios. Now we have four log transformed ratio measurements for each clone–condition combination.

3.4.1 Normalization

A common way of visualizing these ratios is by the MA-plot [94]. M, plotted on the y axis, is log(test/reference). A is the average of the test and reference intensities in log scale, that is $A = (\log_2(\text{test}) + \log_2(\text{ref}))/2$. The upper plot in Figure 3.2 shows the MA plot for two of the 60 performed hybridizations (for one of the five microarray slides at one of the six time points), one for both possible labels for the test sample (Cy3 in blue dots and Cy5 in green dots; the reference is labeled with the other dye).

In case there would be no systematic dye effects, the MA plots would look like those in the lower part of Figure 3.2. On average, there is no difference between the expression level of the test and reference sample, so $M = 0$ on average. However, we can see two deviations from this “perfect” cloud that is centered around zero. First, the green cloud is shifted upwards (average$>$0) and the blue cloud is shifted downwards (average$<$0). This means that the Cy3 dye (the numerator for the blue cloud) has systematically lower intensities than the Cy5 dye. Second, the clouds are not straight but bended. The dye effect, namely the systematic higher intensities for Cy5, seems to depend on the average intensities A and increases as A decreases. Before we can compare the ratios between different time points, we have to correct all ratios to remove this dye effect and to get a distribution of ratios like in the bottom plot of Figure 3.2. In a certain range of average intensities A, the log ratio M approximates a certain constant level. In this range a constant normalization factor can be used.

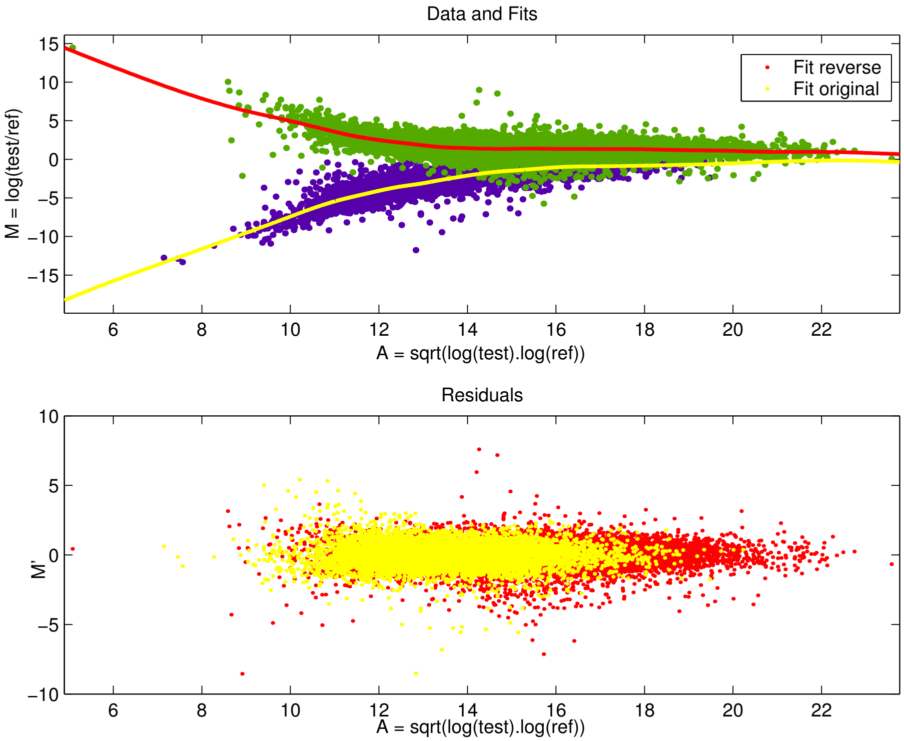


Figure 3.2: Dye-normalization shown with an MA plot. (Above) Log-transformed expression ratios of two hybridizations at the same time point, once with the test sample labeled with Cy3 and the reference with Cy5 (blue dots) and once the other way around (green dots). The y axis is the log ratio and the x axis is the average of the log transformed test and reference intensities. Both clouds show an intensity-dependent dye effect. The green channel (Cy3) has systematically lower intensities than the red channel (Cy5) and this error increases for lower average intensities. A LOWESS fit is shown for both clouds (red and yellow lines through the clouds). (Below) The residuals of the fit are the normalized ratios.

However, as the average expression value A decreases, the log ratio M deviates from a constant level and the use of an intensity-dependent rescaling factor is more appropriate. Therefore a global locally weighted scatterplot smoothing named LOWESS was used [94]. This was done with a smoothing parameter of 0.2 using Matlab's curve fitting toolbox (The Mathworks, MA, US). The results of this fit can be used to simultaneously linearize (i.e., remove nonlinear dye effects) and normalize the data (remove consistent sources of variation caused by within slide dye and condition effects) [327, 94]. Since the M versus A plot was used to fit the data, for each gene novel normalized ratio estimates were calculated. From these normalized ratios, new values for the absolute expression levels can be derived.

Normalization was done using only those ratios with non-zero intensities for test and reference. This is needed because the division by zero values or the log of a zero value result in undefined values. However, the intensities of the test sample that were below the background intensity, still contain information, namely that the gene is turned off in that condition. To keep this information in the data set, the ratios with zero intensities for the test sample are replaced by a small value, namely the 0.025 percentile of all other normalized ratios. The spots with a reference intensity below the background (denominator=0) are considered as missing values in further analyses.

The fluorescent images did not suffer from serious spatial effects since the data measured by all print-tips showed a very similar log expression ratio distribution. Also, the distribution of gene expression ratios from different slides showed similar distributions (data not shown).

3.4.2 Filtering

Spot filtering

Each clone is spotted twice on each microarray slide and the resulting duplicate measurements should optimally be the same. That means that the difference between the log ratios should be zero. A distribution of this value is shown in Figure 3.3.A. High-quality hybridizations show narrowly peaked normal distributions. This was checked for each of the 60 slides as a quality control. The normal distribution can also be used for spot filtering by removing those spots for which the difference between the log ratios of the duplicate measurements is bigger than some standard deviations (SD) from the mean. The latter is implemented in the NEURODIFF web application (see further).

Clone filtering by pairwise correlation

The log-transformed dye-normalized values were used to calculate the Pearson correlation coefficient for all six pairwise combinations of the four individual profiles that are available for each clone:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.1}$$

If two genes are perfectly correlated, the correlation coefficient r is 1. The sign of the coefficient depends on a positive or negative correlation. If the correlation is imperfect, r is less than 1 and a value of 0 indicates that there is no relationship. For four profiles, there are six possible values of r. Clones were filtered using a lower threshold of 0.6 for the minimum of the six correlation coefficients.

Clone filtering by gene-wise ANOVA

An independent filtering was applied on the 4 profiles of each clone to retain only significantly changing genes by ANOVA analysis on each clone with one factor (time) and six treatments, using as H_0-hypothesis that the means at all time points are equal.

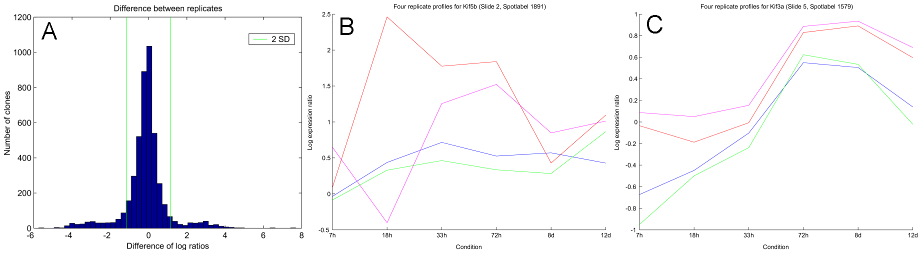


Figure 3.3: Examples of filtering results. (A) Distribution of the differences between the log ratios of duplicate measurements of the same clone on the same slide. A standard deviation cut-off can be used to filter out bad quality spots. (B-C) Two examples of the four replicate profiles for one clone. (B) Kif5b (pval=0.4064 and mincorr=-0.2267) is removed by the clone filtering and (C) Kif3a (pval=7.06062E-005, mincorr=0.8966) is retained. See text for an explanation of the clone filtering.

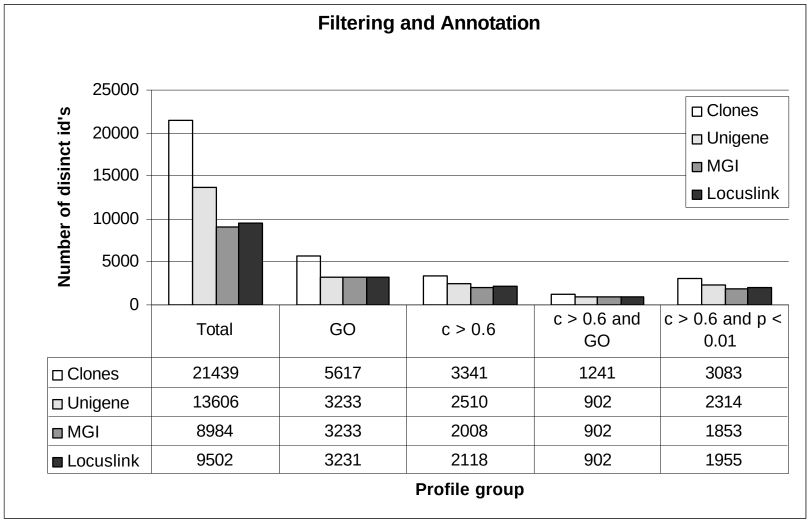
The result of the clone filtering is a selection of reliable (reproducible for the four replicate profiles) average profiles. The correlation filter retained 3,341 clones (15.5 % of all) representing 2,510 distinct Unigene clusters in the differentiating neurons (see 3.4). ANOVA tests showed that almost all (92%) of these 3,341 reliable profiles showed highly significant ($p<0.01$) changes in expression during the time course. In other words, there are very few genes with a reproducible “flat” or unchanging expression profile. For the filtered set of high quality clones, the four profiles are averaged to give one time dependent profile for each clone. These average profiles were stored in a MySQL database together with the clone annotations, and they will be used for further analysis.

Clone annotation

For many practical purposes it is important that a cDNA clone is assigned to a known entry in external databases, describing transcript (Unigene), gene (MGI), genomic location (LocusLink), or function (GO). We reannotated our clone set and made the annotations easy to use via our web application. Figure 3.4 summarizes the annotation.

3.5 Clustering

Several clustering runs were performed using our implementation of the K-means algorithm [300, 292] with different values for the parameter K, and using the Adaptive Quality Based Clustering (AQBC) algorithm [81] with different setting for the threshold. Both algorithms group expression profiles that are similar using the Euclidian distances between the profiles:



	Total	GO	c > 0.6	c > 0.6 and GO	c > 0.6 and p < 0.01
□ Clones	21439	5617	3341	1241	3083
□ Unigene	13606	3233	2510	902	2314
▨ MGI	8984	3233	2008	902	1853
■ Locuslink	9502	3231	2118	902	1955

Figure 3.4: Filtering and annotation of our data set. The columns represent: the whole clone set (Total), its subset annotated with GO, the clones that pass the correlation filter (reliable profile) ($c > 0.6$), the intersection of the previous two subsets ($c > 0.6$ and GO), and the clones with reliable profile and significant change by ANOVA ($c > 0.6$ and $p < 0.01$). The rows give numbers of distinct identifiers from the specified transcript (Unigene) and gene (MGI, LocusLink) databases.

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

K-means groups the profiles in an iterative way into a predefined number of clusters. AQBC differs from K-means in three aspects: (1) the number of clusters is estimated by the algorithm itself; (2) an EM optimization step is used to determine how large each cluster should be; and (3) if a profile does not belong to one of the clusters it is removed. The profiles that were used for clustering were the averages of the four replicate profiles of each clone that passed the correlation filter described above, in log scale and normalized by subtracting the mean and dividing by the standard deviation of the six measurements. Figure 3.5 shows the clustering results for a K-means clustering with K=20.

Several of the clusters are enriched for genes involved in similar functions (as compared to the other clusters). Based on these functions, four larger groups of clusters were identified, indicated in Figure 3.5. Group A contains Clusters 1 and 15, both strongly down-regulated during neuronal differentiation (yellow box in Figure 3.5). This group contains many genes involved in the cell cycle: DNA replication (*Top2b*, *Prim1*, *Prim2*, *Lig1*, *Rev3l*), chromatin assembly (*H1f0*, *H3f3a*, *H3f3b*, *Chaf1a*, *Nasp*), cell-cycle regulators (*Ccnb1-*

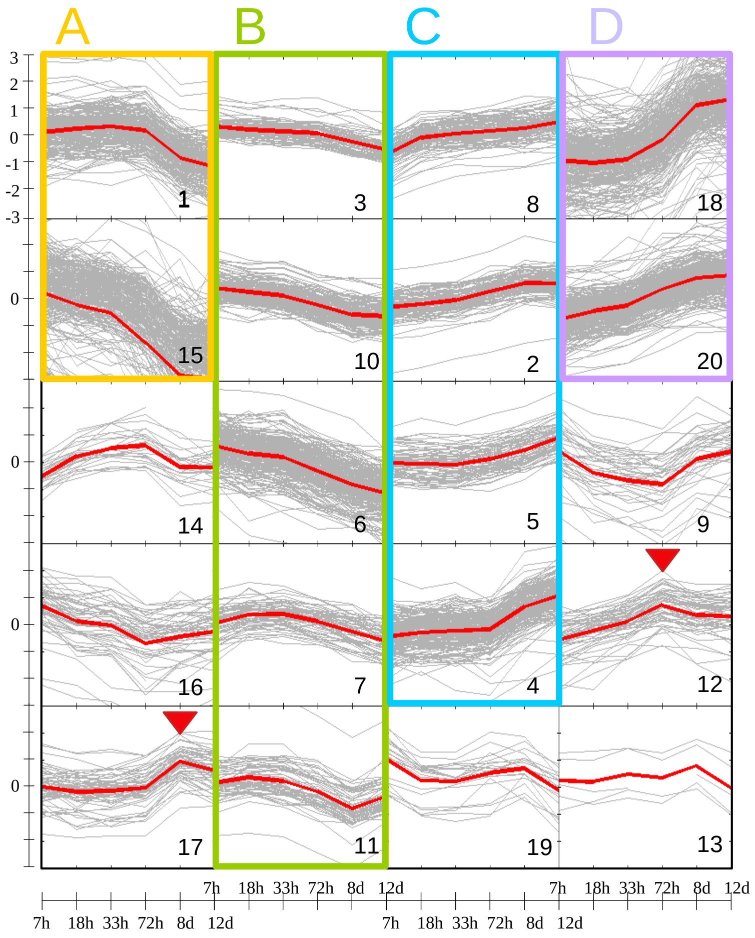


Figure 3.5: Results of a K-means clustering for K (the number of clusters) equal to 20. Each of the 20 squares represents one cluster. On the x axis are the six time points (7 h, 18 h, 33 h, 72 h, 8 days, 12 days) and on the y axis the log-transformed filtered and normalized expression ratios. Each grey line represents a profile for a single clone and the average profile of a cluster is shown as a bold red line. Each cluster or a part of it can be visualized separately as a profile chart or as a heatmap using our web application. The red triangles mark the maxima at 72 h and 8 days. The clusters in which we found an over-representation of genes with stated functions were grouped together and boxed in color. The clusters are not numbered consecutively to maintain the correspondence between the cluster numbers in this figure and in the original clustering results on our web site. The functions over-represented in the boxed groups of clusters: (A) DNA replication, chromatin assembly; (B) ribosomal proteins, RNA binding, translational regulation; (C) Golgi/ER/lysosome, protein traffic, energy, vesicular transport; (D) energy, synaptic, *App*-related.

rs, *Ccna2*, *E2fb*, *Bub1b*). Group A also contained several components of the actin cytoskeleton (*Actb*, *Myo1b*, *Vil2*) indicating that also these genes become down-regulated as a part of the reorganization of the cytoskeleton during the neuronal differentiation. Group B, encompassing Clusters 3, 6, 7, 10, and 11 (green in Figure 3.5), contains moderately down-regulated genes. This group is highly enriched in genes involved in RNA metabolism: ribosomal proteins (at least 29, a full lists of genes in each cluster can be found at our web site), RNA binding proteins (*Pcbp2*, *Refbp1*, *Nsap1*), nuclear ribonucleoproteins (*Snrpa1*, *Snrpd1*, *Snrbp*, *Hnrpa1*, *Hnrpab*, *Hnrpk*, *Hmrph1*, *Hnrpa2b1*), RNA splicing factors (*Sfrs2*, *Sfrs3*, *Sf3b1*), and regulators of translation (*Eif1b2*, *Eif4a1*, *Eif4g2*, *Eef1a1*, *Naca*). Overall these changes indicate a progressive switch from biosynthetic activity towards more functional activity in the developing neurons (also reflected in the upregulated genes, see further). Group B also contains several genes of the TGF-β signaling pathway (*Bmp1*, *Tgfb2*, *Tgfbr1*, *Madh4*, *Sin3b*). Notch1 can be found in Cluster 6 and the kinase Rock1 in Cluster 10, both proteins involved in axonal and dendritic outgrowth [23, 245, 140]. Group C, composed of Clusters 2, 4, 5, and 8 (blue in Figure 3.5), contains all genes that were moderately up-regulated. This group is enriched for genes playing a role in the secretory and endocytic pathways and it includes genes for ER localized proteins (*Hspa5*, *Erp29*, *P4ha2*, *Noe1-pending*, *Aldh3a*, *Sec61a2-pending*), Golgi proteins (*Grs2-pending*, *Vcp*, *Siat6*, *Cope*, *Ap1b1*), lysosomal proteins (*Ptp*, *Tpp2*, *Smpd1*, *Lyst*), a series of intracellular protein trafficking regulators (*Rab12*, *Cop7a*, *Cop9*, *Arl6ip*, *Prnp*, and *Ly6e*), and vacuolar sorting proteins (*Vps29*, *Vps41*, *Vps45*). A second function clearly overrepresented in Group C is energy metabolism, as exemplified by the presence of ATP synthases, cytochrome subunits, and soluble mitochondrial enzymes. Also several synaptic function-related genes are represented in Group C, including *Syt11*, *Synj2bp*, and *Gad1*. Group D, comprising the strongly upregulated Clusters 18 and 20 (purple in Figure 3.5), contains many genes encoding proteins involved in neuron-specific functions and energy metabolism. The latter includes ATP synthase genes (*Atp61*, *Atp6a2*, *Atp6b2*, *Atp6s1*) and glycolysis enzymes (*Aldo1*, *Gpi1*, *Ldh1*, *Pgk1*, *Ckmt1*). Neuron-specific functions are represented by genes of the synaptic vesicle cycle (*Syngr3*, *Vamp2*, *Nsf*, *Syn2*, *Arf2*), a GABA receptor Gabrg2, ion channels (*Kcnd2*, *Clcn3*), and cytoskeletal genes (*Tuba4*, *Nfl*, *Kifc2*, *Kifc3*, *Kifap3*). Interestingly in the light of ongoing research on the function of the amyloid beta precursor protein (*App*) and its role in Alzheimer's disease, the majority of Alzheimer's disease related genes also belongs to this group. This includes *App*, *Aplp2*, *Icam5*, *Adam9*, *Psen2*, and *Snca*. In the remaining clusters we were not able to identify a clearly overrepresented function, possibly because they contain a large proportion of ESTs representing unknown genes. Nevertheless, these clusters are of interest. The genes in Cluster 12 for example have a maximum expression at 72h in culture, correlating with the beginning of dendritic outgrowth. Members of this cluster include *Rac3*, *Kif3*, and *Apc*. Cluster 17, with maximum of expression at 8d in culture (which correlates with the early stage of full polarization of the neurons), contains genes with diverse functions.

3.6 Analysis by gene function: the synaptic vesicle cycle genes

While the clustering approach classified genes according to their expression profiles, followed by a functional interpretation of the clusters, an opposite approach can also be taken to analyze the expression data, namely, by investigating shifts in expression within functional classes of genes. Such gene sets can be created by an expert in a certain field, from the literature, from databases of biological pathways (e.g., Kyoto Encyclopedia of Genes and Genomes (KEGG) [287]), or from databases of functional gene annotations like Gene Ontology (see further).

Here we illustrate this second approach for a group of genes important for the synaptic vesicle cycle, created by an expert in the field [77]. The regulated release of neurotransmitters is a key function of the mature neuron. Many proteins have been implicated in the process of targeting and docking of the synaptic vesicle at the presynaptic membrane, the priming of the vesicle, and its fusion [185, 188]. Many of these proteins are also involved in membrane trafficking outside the synapse [64, 226]. Conversely, for many members of the same protein families, the subcellular localization and function is not yet known. We decided to select all the genes in a gene family (e.g., *Vamp*) if at least one member of this family (e.g., *Vamp2*) is known to play a role at a particular stage of the synaptic vesicle cycle, essentially following [188]. During differentiation a gradual shift in expression of members of the different families was observed (see Figure 3.6).

The most strongly regulated gene in the membrane fusion group was *Nsf*, involved at the disassembly step [299]. The early expression pattern of *Vamp3*, *Snap23*, and *Syt4* coincided temporarily with the outgrowth of the early axon (Stage 3). These findings are in good agreement with published data showing that *Vamp3* is not found in synaptic vesicles [65] and that *Snap23* is almost undetectable in adult brain [324]. *Vamp2* in contrast had a late expression pattern, with a maximum at 12d (Stage 5) in the culture. *Vamp2* is involved in the synaptic vesicle fusion step as part of the SNARE complex and is essential for secretion [8, 213]. It has been postulated that the general exocytosis machinery [155] is used in the outgrowth of both axons and dendrites [207, 290]. A high level of expression of the ubiquitously expressed *Scamp2* coincided with axonal outgrowth. *Syn1* was expressed at a high level until 72h in culture (early Stage 4), and then became strongly down-regulated. The *Syn1* protein (together with *Syp*) is known to preferentially localize to the distal axon and the growth cones at the stage of early axonal outgrowth (before cell-cell contact) [107], which is consistent with its early expression pattern in our experiment. Of the two transcripts of *Syn1* described in rat cerebellum, the longer one is expressed only until P7 [129] (P is postnatal day), similarly to the profile we observed. Expression of *Syn2* and *Syngr3* [284] reached their maximum at 12d in the culture (Stage 5), suggesting a role in mature synapses. Exocytotic events in the nerve terminals are compensated by endocytosis [84, 158]. In keeping with this, the genes in the endocytosis group were also up-regulated in the late Stages

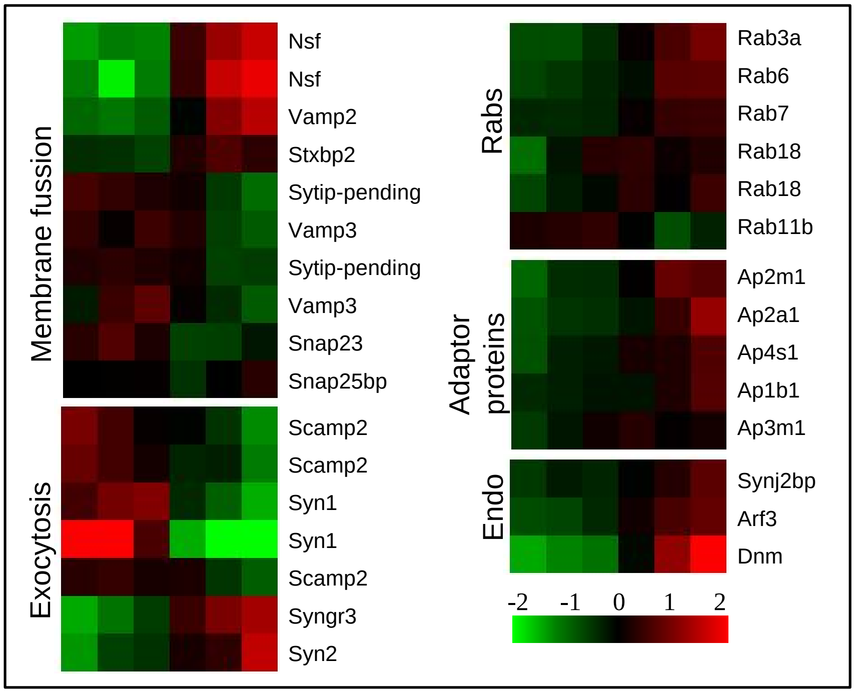


Figure 3.6: Expression of the synaptic vesicle cycle genes. The genes were chosen as described in the main text. Within each subgroup the similar profiles were grouped together by hierarchical clustering (implemented at our web site). The colors indicate level of expression at a given time-point, relative to the average expression of this gene over all time-points, in $\log_2$ scale, with red for expression higher and green for lower than the gene average.

4, 5 of the culture, with a maximum at 12d in culture when synapses have been generated (Figure 3.1). In developing rat brain the expression of *Dnm* starts to increase from P7 and reaches adult levels at P23 [103]. Also in the chick embryo's retino-tectal system dynamin is up-regulated only after synapse formation [25]. In rat brain the mRNA expression of *Arf3* increases postnatally, from P2 to P27 [301]. The clustering analysis discussed above revealed co-expression of many synaptic and mitochondrial genes. Therefore it is interesting to note that *Synj2bp* recruits synaptojanin to mitochondria, which may affect their intracellular distribution [219]. The early-expressed *Rab11b* has a role in the apical membrane recycling systems in polarized epithelial cells and in growth cone mobility [60, 174]. The late-expressed *Rab3a* is up-regulated in development only after synapse formation. Expression of this gene in the developing barrel field occurs later than that of *Sv2a*, *Syn1*, and *Syp*, and coincides with the onset of adult-like physiological activity [274]. The majority of the genes encoding adaptor proteins were expressed late in our culture system, including *Ap2a1*. Similarly to what is seen in *Drosophila* development [89], the expression patterns of *Ap2a1* and *Dnm* were highly similar. Adaptor proteins and *Rab* proteins are important at several stages of intracellular vesicular transport, and their predominant late-expression pattern is likely to reflect increased transport needs in mature neurons. The observed upregulation of *Dyn*, *Rab3a*, *Arf3*, and *Ap2a1* at 12d in the culture is in agreement with the fact that by Stage 5 the neurons had functioning synapses with active secretion and compensatory endocytosis.

3.7 Functional exploration using Gene Ontology

It is obvious that no investigator can have such thorough knowledge of many biological functions or pathways. Therefore, an automated retrieval or construction of functional groups would be valuable to facilitate this process. To this end we will use the functional annotations of genes by Gene Ontology terms. The Gene Ontology (GO) Consortium (http://www.geneontology.org/) [135] has developed a set of controlled, structured vocabularies—known as ontologies—to describe key domains of molecular biology, including gene product attributes and biological sequences. There are three ontologies that describe three non-overlapping domains of molecular biology, namely Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) containing respectively 7267, 8114, and 1378 terms (as of January 5, 2004). GO is structured as a directed acyclic graph (DAG) where the nodes are the GO terms and the arcs between the nodes are parent-child relationships ("is-a" or "part-of"). The GO terms all have a unique GO identifier and these are applied in the annotation of sequences, genes, or gene products and are stored in biological databases by collaborating model organism databases like FlyBase, Saccharomyces Genome Database (SGD), Mouse Genome Informatics (MGI), and Ensembl (the latter is done by GOA or GO Annotation@EBI). An example of a GO term is "motor activity" (the GO identifier for this term is GO:0003774), which is defined in GO as "catalysis of movement along a polymeric molecule such as a microfilament or microtubule, coupled to the hydrolysis of a nucleoside triphosphate". The parent term of "motor activity" is "molecular function" (GO:0003674) and the child terms are "microfilament motor activity" (GO:0000146) and "microtubule motor activity" (GO:0003777). All terms and all term-term relationships in the GO database are stored in our MySQL database together with all available GO annotations for the 3233 MGI IDs in our data set (using the annotation as of October 2002).

GO classes of genes can be constructed by grouping all genes that are annotated with a certain GO term or with one of the child terms along all the possible paths that lead to a leaf node in the DAG downstream from that term. This can be done interactively in the purpose-built web application.

A web application for functional gene expression analysis

The normalized expression ratios (for all replicates and also averaged per clone), the different gene identifiers and descriptions, the clustering results, and the GO annotations are all stored in a MySQL database with a purpose-built data model for fast and flexible data retrieval (not shown). On top of this database, a web application called NEURODIFF was developed and is available at the following URL: http://www.esat.kuleuven.ac.be/neurdiff/.

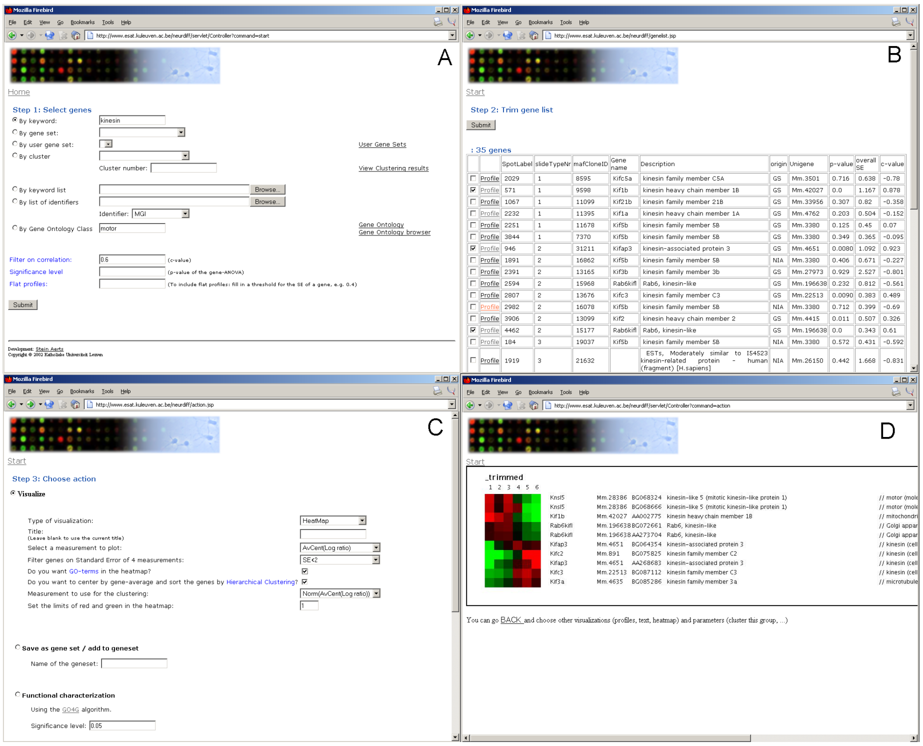
The web application together with the structured data in the database can be regarded as an "intelligent microarray data set". The following functions can be performed:

1. Construct a gene group by
    - Gene id (Accession number, MGI, Unigene)
    - Search for genes using keywords that are present in a gene's description
    - Gene Ontology term or ID
    - Selection of a cluster from one of 15 clusterings performed off-line
2. Visualize all replicate measurements for one clone and apply a filtering on the genes within a group based on their internal correlation or p-value from the ANOVA analysis
3. Perform a hierarchical clustering within the gene set
4. Visualize the time profiles of the gene set in a heat map or in a profile chart using different representations of the measurements (ratio, normalized ratio, mean centered ratio; all in either log or linear scale)
5. Save the gene group as a gene set for later investigation.

The application architecture consists of three tiers: (1) a data layer, namely the MySQL database; (2) a business layer consisting of several Java classes and Java Beans controlled by a central Java servlet (http://java.sun.com/products/servlet/) that run in the Apache Tomcat servlet container (http://jakarta.apache.org/tomcat/), and of C executables for calculations (e.g., hierarchical clustering); and (3) a visualization layer using JavaServer Pages (JSP) and HTML.

3.8 Comparing two microarray data sets

Mody at al. [214] reported a data set consisting of 1,926 significantly changed expression profiles in developing mouse hippocampus *in vivo* (H). The 3,341 expression profiles in the neurons differentiating *in vitro* (N), resulting from our experiment, provided us with an exciting opportunity to compare the expression of genes in both systems. We downloaded the H data set from http:


A
B
C
D
Mozilla Firebird
File Edit View Go Bookmarks Tools Help
http://www.esat.kuleuven.ac.be/neurdiff/servlet/Controller?command=start
http://www.esat.kuleuven.ac.be/neurdiff/genelist.jsp
http://www.esat.kuleuven.ac.be/neurdiff/action.jsp
http://www.esat.kuleuven.ac.be/neurdiff/servlet/Controller?command=action
Home
Start
Submit
Browse...
User Gene Sets
View Clustering results
Gene Ontology
Gene Ontology browser
1 2 3 4 5 6

Step 1: Select genes

● By keyword: kinesin
○ By gene set:
○ By user gene set:
○ By cluster
Cluster number:
○ By keyword list
○ By list of identifiers
Identifier: MGI
○ By Gene Ontology Class: motor

Filter on correlation: 0.6 (c-value)
Significance level (p-value of the gene-ANOVA)
Flat profiles: (To include flat profiles: fill in a threshold for the SE of a gene, e.g. 0.4)

Development: Stein Aerts
Copyright © 2002 Katholieke Universiteit Leuven

Step 2: Trim gene list

: 35 genes

		SpotLabel	slideTypeNr	mafCloneID	Gene name	Description	origin	Unigene	p-value	overall SE	c-value
☐	Profile	2029	1	8595	Kifc5a	kinesin family member C5A	GS	Mm.3501	0.716	0.638	-0.78
☑	Profile	571	1	9598	Kif1b	kinesin heavy chain member 1B	GS	Mm.42027	0.0	1.167	0.878
☐	Profile	1067	1	11099	Kif21b	kinesin family member 21B	GS	Mm.33956	0.307	0.82	-0.358
☐	Profile	2232	1	11395	Kif1a	kinesin heavy chain member 1A	GS	Mm.4762	0.203	0.504	-0.152
☐	Profile	2251	1	11678	Kif5b	kinesin family member 5B	GS	Mm.3380	0.125	0.45	0.07
☐	Profile	3844	1	7370	Kif5b	kinesin family member 5B	GS	Mm.3380	0.349	0.365	-0.095
☑	Profile	946	2	31211	Kifap3	kinesin-associated protein 3	GS	Mm.4651	0.0080	1.092	0.923
☐	Profile	1891	2	16862	Kif5b	kinesin family member 5B	NIA	Mm.3380	0.406	0.671	-0.227
☐	Profile	2391	2	13165	Kif3b	kinesin family member 3b	GS	Mm.27973	0.929	2.527	-0.801
☐	Profile	2594	2	15968	Rab6kifl	Rab6, kinesin-like	GS	Mm.196638	0.232	0.812	-0.561
☐	Profile	2807	2	13676	Kifc3	kinesin family member C3	GS	Mm.22513	0.0090	0.383	0.489
☐	Profile	2982	2	16078	Kif5b	kinesin family member 5B	NIA	Mm.3380	0.712	0.399	-0.69
☐	Profile	3906	2	13099	Kif2	kinesin heavy chain member 2	GS	Mm.4415	0.011	0.507	0.326
☑	Profile	4462	2	15177	Rab6kifl	Rab6, kinesin-like	GS	Mm.196638	0.0	0.343	0.61
☐	Profile	184	3	19037	Kif5b	kinesin family member 5B	NIA	Mm.3380	0.572	0.431	-0.592
☐	Profile	1919	3	21632		ESTs, Moderately similar to I54523 kinesin-related protein - human (fragment) [H.sapiens]	NIA	Mm.26150	0.442	1.668	-0.831

Step 3: Choose action

● **Visualize**

Type of visualization: HeatMap
Title:
(Leave blank to use the current title)
Select a measurement to plot: AvCent(Log ratio)
Filter genes on Standard Error of 4 measurements: SE<2
Do you want GO-terms in the heatmap? ☑
Do you want to center by gene-average and sort the genes by Hierarchical Clustering? ☑
Measurement to use for the clustering: Norm(AvCent(Log ratio))
Set the limits of red and green in the heatmap: 1

○ **Save as gene set / add to geneset**

Name of the geneset:

○ **Functional characterization**

Using the GO4G algorithm.

Significance level: 0.05

_trimmed

Knsl5	Mm.28386	BG068324	kinesin-like 5 (mitotic kinesin-like protein 1)	// motor (mole
Knsl5	Mm.28386	BG068666	kinesin-like 5 (mitotic kinesin-like protein 1)	// motor (mole
Kif1b	Mm.42027	AA002775	kinesin heavy chain member 1B	// mitochondri
Rab6kifl	Mm.196638	BG072661	Rab6, kinesin-like	// Golgi appar
Rab6kifl	Mm.196638	AA273704	Rab6, kinesin-like	// Golgi appar
Kifap3	Mm.4651	BG064354	kinesin-associated protein 3	// kinesin (cell
Kifc2	Mm.891	BG075825	kinesin family member C2	// kinesin (cell
Kifap3	Mm.4651	AA268683	kinesin-associated protein 3	// kinesin (cell
Kifc3	Mm.22513	BG087112	kinesin family member C3	// kinesin (cell
Kif3a	Mm.4635	BG085286	kinesin family member 3a	// microtubule

You can go BACK and choose other visualizations (profiles, text, heatmap) and parameters (cluster this group, ...)

Figure 3.7: Screenshots of the NEURODIFF application. (A) Selection of genes to construct a gene set. (B) Clone in the gene set that pass the chosen filter are automatically ticked. The four replicate profiles of a clone can be visualized to double check the filter. (C) Choose a visualization (profile chart or heatmap), whether to cluster the gene set hierarchically, or choose to save or to characterize the gene set with GO4G (see Chapter 5).

//braingenomics.princeton.edu. There were 475 distinct genes (defined here as distinct Unigene IDs) represented in both data sets. Because some of the genes were represented more than once, the number of possible comparisons (686) was higher than the number of common genes.

N contains six time points and H contains five slightly different time points. To make a comparison we had to find the optimal mapping of five consecutive time points between the two experiments. This was done using two approaches. In the first approach, we calculated the Pearson correlation coefficients (see Equation 3.1) for the six possible order-preserving mappings of five time points between N and H, and identified the mapping for which the median correlation between the two data sets was highest.

In the second approach, we used a published implementation of time warping called *genewarp* [1]. Similar to algorithms used for sequence alignment, time warping aligns two time series against each other. Whereas sequence alignment algorithms consider the similarity of pairs of single bases or residues taken one from each sequence, time warping considers the similarity of pairs of vectors taken from a common k-dimensional space (feature space) taken one from each time series. Here the feature space comprises vectors of RNA expression levels from a common set of k genes. Dynamic programming is used to find the (many-to-many) mapping between the time points of the two series that minimizes a weighted sum of the k-space distances between the corresponding alignments as paths through the grid cells, and finding the path with minimum accumulated weighted distance score. Horizontal or vertical segments of the optimal path identify places where multiple time points of one series correspond to a single time point of the other. Where measurement time intervals are comparable between the series, these may represent situations in which the instance of the biological process measured by one series moves quickly through a phase of the process relative to the instance measured by the other series. Such situations are called compexps (compression/expansions) and they are analogous to the indels (insertion/deletions) considered in sequence alignment algorithms [1]. The result of *genewarp* is shown in Figure 3.8, and the Pearson correlation coefficients between expression profiles *in vivo* and *in vitro* calculated corresponding to this mapping are plotted in Figure 3.9.

The mean and median correlation coefficient are 0.646 and 0.787 respectively. To illustrate the significance of these values, we constructed permuted data sets by random permutation of the time points of each gene profile separately. The median correlation coefficient for the "optimal" mapping between the same data sets with permuted time profiles was only 0.0394. The high correlation obtained for the shown mapping, between an experiment *in vitro* lasting 12 days and an experiment *in vivo* spanning 34 days, suggests that the program of gene expression was accelerated *in vitro* when compared to the situation *in vivo*. This is clearly the case for the highly correlated genes, which are illustrated in Figure 3.10 where profiles are plotted on the same time scale.

The link P30–12d contributes highly to the overall high similarity between the two expression profiles. When we calculated for instance the correlation coefficients for all the possible mappings of four consecutive time points (thus removing in some mappings the P30) we found that the six highest ranking mappings included the link P30-12d (data not shown), indicating that high similarity extends to the latest points in the two experiments.

3.9 Discussion

We provide here the first genome-wide analysis of changes in gene expression accompanying the differentiation of hippocampal neurons *in vitro*. This culture system has been used extensively in the past and is considered as an excellent model to study neuronal cell biology [108, 39, 40, 267, 148]. We extend this claim to the transcriptome level (Figure 3.9). We have demonstrated that neuronal differentiation is characterized by changes in the expression of genes from many different functional families. At least 2,314 genes show a change in expres-

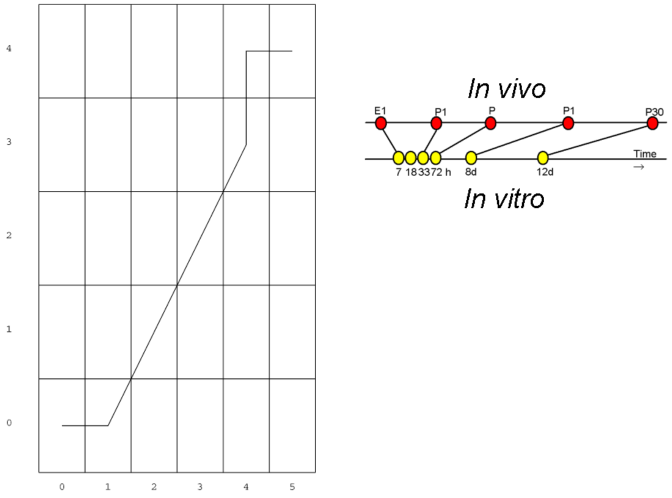


Figure 3.8: Comparison of gene expression in the developing hippocampus and in the neurones differentiating *in vitro*. Experiments and time points used for the comparison of gene expression in the developing mouse hippocampus (data set from Mody et al. [214]) and in the mouse hippocampal neurones differentiating in culture (our data). (Left) Output of the genewarp program showing the optimal path that corresponds to the optimal mapping between the conditions of the two experiments under comparison. The time points of the *in vitro* experiment are on the x axis, those for the *in vivo* experiment on the y axis. For example, time points 0 and 1 *in vitro* both map to time point 0 *in vivo*. (Right) Alternative representation of the optimal mapping but without mapping two points to a single corresponding point.

sion with a statistical certainty (p <0.01). This indicates that the rebuilding of the rounded, unpolarized cell observed at Stages 1 and 2 (Figure 3.1) into the highly complicated polarized and electrically active neuron in Stage 5 requires an orchestrated change in expression of thousands of genes. This change is remarkably smooth as the dominant pattern in the gene expression profiles is a gradual up- or downregulation of gene expression over several stages of differentiation. This pattern was seen at the global level of analysis, but also within most functional groups (e.g., synaptic vesicle cycle in Figure 3.6), resulting in replacement of early genes by the late ones with (seemingly) similar function. The change in expression was significant at the p-level 0.01 for 2314 genes out of 2510 genes that pass the correlation (reproducibility) filter. Apparently, filtering for reproducible profiles results in a bias towards the genes with a change in expression. We hypothesize that the genes that change expression during differentiation are regulated in a more robust way (resulting in a higher biological reproducibility) or are expressed at higher absolute levels (which leads to more reproducible measurements). One word of caution is indicated. When interpreting our data set, it is important to take into account that the less-than-perfect developmental-phase coherence among neurons could

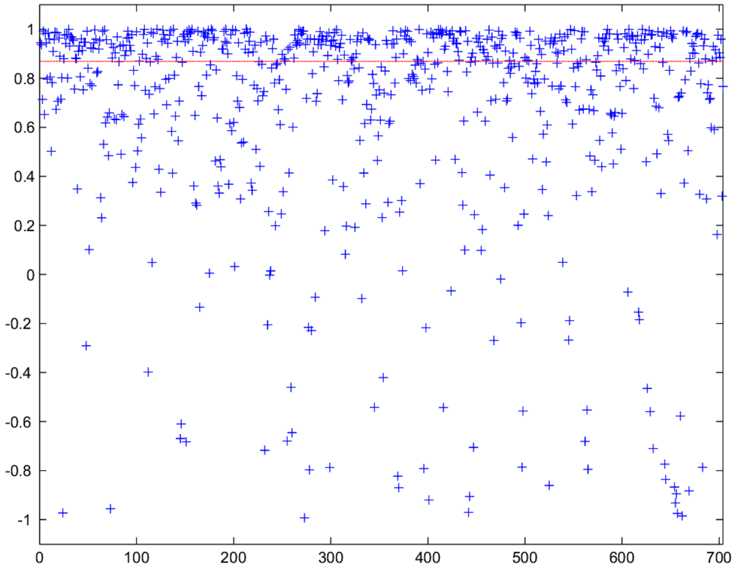


Figure 3.9: Microarray data comparison. Scatter plot of the Pearson correlation coefficients, between the log-transformed normalized profiles composed of the five time points indicated in Figure 3.8, for each distinct pair of profiles representing the same transcript (Unigene ID) in both data sets.

partially smoothen the slope of gene profile curves. On the other hand, the neurons differentiate in a remarkably synchronized way (see Figure 3.1) and data were sampled in duplicate at six different time points in two independent experiments, providing a quite high level of resolution and reliability. The global picture of gene expression patterns during neuronal differentiation as it emerges from the clustering analysis makes remarkable teleological sense. In a first phase of the culture (Stages 2, 3, early Stage 4), a high level of expression of genes characteristic for DNA and protein synthesis is observed, which then becomes progressively down-regulated (Figure 3.53, Group A, Clusters 1 and 15). The later Stages 4 and 5 of differentiation are characterized by a strong enhancement of protein transport (Figure 3.5, Group C: Clusters 2 and 5) and energy generating systems (Figure 3.5, Group D: Clusters 18 and 20) and the turning on of specific neuronal functions, such as synaptic vesicle cycling (Figure 3.5, Group D: Clusters 18 and 20, see also Figure 3.6). The high morphological resolution of the *in vitro* system permitted us to identify gene expression patterns characteristic for the axonal (Stage 3) and dendritic phase (Stage 4) of neuronal differentiation. Therefore we were able to resolve the gene expression patterns described in [214] as “differentiation and synapse formation”, into two very different patterns: early - characteristic for axonal outgrowth and late - characteristic for dendritic outgrowth and maturation. The difference can be appreciated, for example, by choosing “cytoskeleton” as the GO group at our

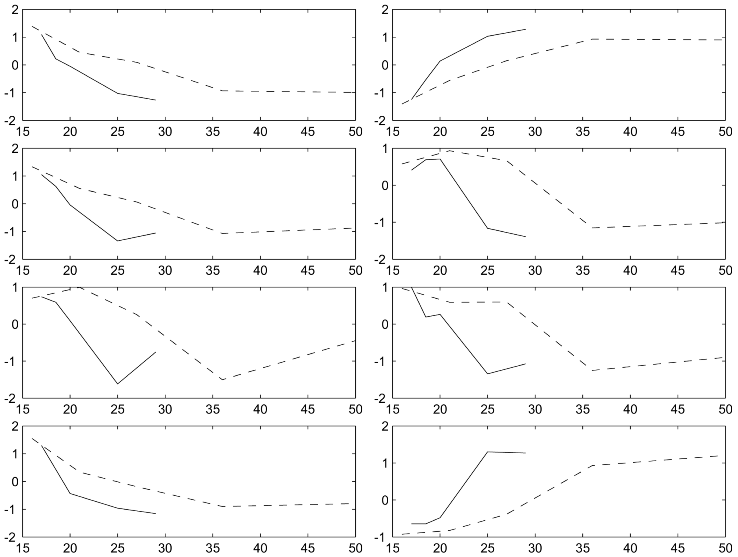


Figure 3.10: Microarray data comparison. Representative profile charts for eight genes with a high correlation coefficient showing the profiles composed of the five time points indicated in Figure 3.8. The *in vitro* and *in vivo* profiles are plotted on the same time scale. The x axis represents time in days from the conception and the y axis the expression values.

web site. Classification of the approximately 1,000 genes in our data set for which no function is known into these two classes may be a useful first step towards the further elucidation of their function in neurons. The high similarity of the expression profiles for the 475 common genes, in our data set and in the data set published by the group of Joe Z. Tsien [214], has several important consequences. First, each data set can be considered as an independent confirmation of the other one. Second, we demonstrate that the results obtained with two different experimental platforms can yield a good agreement. Assuming that the measurements from both platforms are reliable, the effect of which platform is used should be small and this is essentially what we have found. In a recent comparison [171] of previously published data sets obtained with cDNA microarrays [250] and with Affymetrix oligonucleotide chips [56] the mean correlation between the two platforms was 0.278. The authors' conclusion [171] was that "corresponding measurements from the two platforms showed poor correlation [. . .] implying a poor prognosis for a broad utilization of gene expression measurements across platforms". The correlations we report here between the measurements obtained with cDNA microarrays (our data set) and with oligonucleotide chips [214] are much higher than those reported by [171]. Several factors contribute to this difference: (1) differences in preprocessing (Lowess fit vs. constant dye normalization), (2) use of averages across replicates, and (3) filtering of reliable expression profiles. In our case, the use of (1) and (2), prior to comparison with the data from [214], resulted in the mean and median correlations of 0.385 and 0.542. The use of only the filtered profiles increased the mean to 0.646 and the median to 0.787. Stringent filtering has therefore a major contribution in improving the quality of the data. This improvement comes at the cost of reducing the number of measurements available (stringent filtering selects 3,341 clones out of 21,439) but this situation is similar to what happens with the Affymetrix platform where the use of replicates allowed the selection of 1,926 clones out of about 11,000. Third, and most important, the high overall similarity (median correlation 0.787) obtained between expression profiles *in vivo* and *in vitro* demonstrates that expression profiles of at least 50% of genes during neuronal development *in vivo* and *in vitro* were remarkably similar, most likely reflecting the same genetic program of neuronal differentiation. Diaz and colleagues [85] showed also recently that a group of genes in isolated cultured granule cells exhibited very similar temporal expression patterns as those observed in the cerebellum *in vivo* between P6 and P20. Apparently once the cells have taken a neuronal fate, the further program of gene expression is, for a period of time, largely independent of histological or anatomical context. Interpretation of our results shown in Figure 3.9 has to take into account the differences in timing of the two compared experiments shown in Figure 3.8. From the results for the optimal mapping of five time points shown in Figure 3.9, and also from the results of mapping of four consecutive time points (data not shown), we conclude that not only was the program of gene expression *in vivo* and *in vitro* largely the same, but also that its execution *in vitro* was faster than *in vivo*. It is clear that culture conditions (cell density, contact with of glia) can affect the rate of neuronal differentiation (e.g., duration of the initial outgrowth of an axon (before it contacts a target cell), or the balance between dendritic elongation and branching [23, 245]). It is therefore understandable that the transition from the tissue to the cell culture conditions may have a strong effect on the rate of differentiation. We do not know what causes the observed acceleration. We are tempted to speculate that the network of connections between neurons in culture is much less elaborate than in the brain [233] and thus it may take much less time *in vitro* to complete the “wiring” and to establish active synapses. It is also clear that the similarity between the situation *in vivo* and *in vitro* breaks at some point past Stage 5, as the neurons in the culture ultimately die. Taken together our findings clearly demonstrate that hippocampal neurons *in vitro* are a remarkable relevant biological system to investigate hippocampal neuronal differentiation.

3.10 Perspectives on *cis*-regulatory sequence analysis

For this dissertation, the microarray data set described in this chapter has been used to make a comparison between *in vitro* and *in vivo* neuronal differentiation and to investigate the expression *in vitro* for particular gene sets or gene functions. The web-based interrogation of the data is useful for investigators, for example to infer gene function by looking for unknown genes that have similar expression profiles (e.g., in the same cluster) as genes with a known function during neuronal differentiation. It is also useful to analyse a group of known genes to see whether genes with similar biological functions are expressed at different timings (by clustering within a GO group). Another application however where this data set can be useful is to discover genetic linkages in a gene regulatory network that regulates certain aspects of neuronal differentiation. This can either be done by direct inference from the expression data (e.g., by Bayesian structure learning) or by combining microarray data with motif and module detection. Preliminary analyses to this end have been done [76] and further work is currently in progress.

3.11 Perspectives on the comparison of microarray data

In the past few years, a myriad of microarray experiments has been produced, overwhelming the research community with a wealth of potentially valuable data. Efficient access to this data and, in particular, efficient comparison and integration of data obtained in related biological systems provide researchers with an opportunity to address complex questions in an effective way. Tellingly, larger microarray projects are scaling up towards the generation of large compendia of gene expression. These will provide a comprehensive view of the transcriptome in different organisms at different stages of development [13] or under different environmental [118] or genetic [151] conditions and of the changes in gene expression that are associated with a diverse series of human pathologies [243]. We envisage a radical change in microarray studies comparable to what happened in sequence analysis with the advent of the genome projects where a division of labor takes place between a few large consortium-based projects on the one hand and the many smaller investigation-specific projects on the other hand. The compendium projects will chart large areas of the transcriptome whereas smaller-scale projects will refine the details, starting from a careful analysis of publicly available microarray (and sequence) data to design experiments that validate and refine primary hypotheses. But what are the barriers to this bonanza of information and how can they be overcome?

3.11.1 Data access and exchange

Until now, most of the publicly available microarray data have been scattered around the internet, often as supplementary data to a published article. Consequently, it has been difficult for investigators to know where the relevant data are available. This problem has been addressed in several databases by making it possible to search for published microarray data that has undergone uniform processing and filtering and by providing links to the original publications for more detailed information. These databases have diverse purposes and are either: (1) platform specific (e.g., the Stanford Microarray Database; http://genome-www5.stanford.edu/MicroArray/ SMD [124]); (2) organism specific (e.g., yeast Microarray Global Viewer; http://www.transcriptome.ens.fr/ymgv [201]); or (3) project specific (e.g., the Lifecycle database on *Drosophila* development (http://genome.med.yale.edu/Lifecycle [13]), our own NEURODIFF database on neuronal differentiation in mouse (http://www.esat.kuleuven.ac.be/neurdiff [77]), or the HugeIndex database on normal expression in human tissues (http://zlab.bu.edu/HugeSearch [139]). Although supplements and microarray databases on the internet provide access to many data sets, they have some drawbacks. (1) They lack direct access to the experimental information that is needed to judge the quality of the data, to repeat a study or to re-analyze the data. (2) A standard format for microarray data and experiment description is not used. These drawbacks make identifying, collecting and analyzing publicly available data sets a cumbersome and error-prone process.

3.11.2 Microarray standards and repositories

The Microarray Gene Expression Data (MGED) Society (http://www.mged.org) provides guidelines, formats and tools to overcome these two drawbacks. The Minimum Information About a Microarray Experiment (MIAME) specification [45] is a checklist that guides the investigator in the annotation of microarray experiments. Because numerous biological and experimental factors influence gene expression measurements (e.g., lighting conditions in plant experiments, the exact histopathology of a tumor, the difference in specificity of different reporter sequences for the same gene, the particularities of a single batch of slides or the laser intensity at which a slide is scanned), this MIAME specification includes the experimental design, array design, details of the samples and any treatments, hybridization conditions, measurements and normalization controls. Furthermore, the MGED ontology [275] provides a framework of microarray concepts for this annotation and the MicroArray Gene Expression Object Model (MAGE-OM) and Markup Language (MAGE-ML) conceptualize MIAME for data storage and exchange [269]. In practice, a local MIAME-supportive database will allow gradual recording of the information generated in the laboratory. Upon publication of a study, the database can directly export the data to a public repository. For a compendium project (e.g., the Compendium of Arabidopsis Gene Expression which will contain 4000 full-genome Arabidopsis microarrays (http://www.psb.rug.ac.be/CAGE)), the data can be first transferred to a consortium database and later to a repository [275]. Currently, the only fully MIAME-supportive database is the ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress) [47], although other microarray databases are being developed so that they will eventually support MIAME [117, 88]. Some journals already require publication of MIAME-compliant data to one of the two current repositories, ArrayExpress or Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) [97]. Although at this early stage observance of the MIAME guidelines has yet to demonstrate improvements in the comparability of microarray experiments, it is clear that without this information meaningful comparison and integration of data generated in different laboratories or on different platforms will be impaired and errors or misunderstandings could go undetected (e.g., if we had not done time warping in our analysis). Even if these data conforms to MIAME standards, however, comparison will remain difficult because many variables are involved and new flexible statistical procedures will be needed that make the most of this information.

We have submitted our microarray data of mouse hippocampal gene profiling to ArrayExpress in MIAME format.

3.11.3 Microarray analysis in the era of repositories and compendia

A new era is dawning on microarray analysis with large public resources of microarray data easily available for retrieval and integrated analysis across platforms. But what are the obstacles lying ahead? And can we expect more benefits than just the improved statistical efficiency offered by meta-analysis? At the technological level, trade-offs in costs and available expertise probably mean that several platforms will coexist for at least several years. However, sequence identity error in cDNA clones (at least in higher organisms) is worryingly high and sequence specificity is not optimal. Therefore, we can expect spotted cDNA arrays to be progressively replaced by spotted arrays of long oligonucleotides or other methodologies that improve sequence identity and specificity [73]. For compendium projects on two-channel platforms, where the use of a common reference is standard practice, using a specific and calibrated reference (e.g., an equimolar mixture of PCR products or oligonucleotides complementary to all array features [93, 273] or external normalization spikes [305]) could greatly improve precision and accuracy–and might even allow recovering absolute measurements. At the methodological level, there is now enough evidence to suggest that replicates of microarray experiments are essential if the data are to be of any value [171]. It must become standard practice to require sufficient biological replication before lending any credence to results based on microarray data. At the practical level, we should not underestimate the burden placed on investigators to keep the annotation and data of each experiment MIAME compliant. This burden will be lessened if good software tools are developed. At the infrastructure level, we can expect many new powerful features (much beyond simple storage and query). For example, data alerts could be automatically generated when a new data set relevant to your research is deposited just like MEDLINE can generate publication alerts based on keywords. Extensive gene-centric views of the transcriptome could be made available for each gene, with a virtual expression profile summarizing all the available expression data [86, 149]. Even automatic discovery alerts might be possible, after semi-automated data collection, by repeatedly performing a standard analysis script as new data becomes available and dispatching each incremental discovery to the investigator just like automatic daily BLASTing of a sequence of interest for homolog detection. At the data analysis level, the ideas are broadly applicable, for example meta-analysis can improve the detection of differential expression, clustering of gene expression profiles across multiple data sets, and classification methods could benefit from similar treatments. In fact, because reliable statistics is the basis of serious data mining, an improved statistical treatment of microarray data across platforms probably means that most data mining techniques applied to microarray results will eventually be able to deal with multiple data sets. If we address fully these real challenges and pursue these exciting opportunities, in the next decade exploring transcriptomes should become almost as natural as exploring genomes.

Chapter 4

Detecting transcription factor binding sites in metazoan genes

4.1 Introduction

In the previous chapter we have dealt with measuring the output of a gene regulatory network, namely a certain amount of mRNA molecules for the genes of the network. From now on we will shift our attention towards the underlying sequence elements that process all the information of a GRN and activate or deactivate the transcription process by communicating with the basal transcription apparatus at the core promoter. The sequence elements we are looking for are binding sites for transcription factors. Because of their limited length (6-12 bp), because of their redundancy, and because of the enormously large intergenic regions where they are hidden, they are difficult to find and the predictions of their location result in many false positives. In this chapter we discuss a newly developed strategy that serves two goals in regulatory sequence analysis: (1) to minimize the number of false positive TFBS predictions for *known* transcription factors in *metazoan* sequences and (2) to do it *fast and easy.* The latter is useful for large scale analyses, for example if many putative gene batteries obtained by clustering algorithms from microarray experiments have to be analyzed. The method integrates three aspects: a sensible way to predict instances of position specific frequency matrices (PSFM, or motif models), the over-representation of PSFM instances in gene batteries, and phylogenetic footprinting. The strategy is implemented in a software platform called TOUCAN and is available freely for researchers of academic institutions.

We will first describe the distinct parts of the strategy together with the TOUCAN tool, next we will discuss several biological analyses that validate the strategy, and also give an example how this method can be used to detect TF cooperativity in modules. Lastly, we will briefly discuss a case study illustrating the use of phylogenetic footprinting alone to find binding sites in a single gene when co-expressed genes are not available.

4.2 Constructing sets of putative regulatory sequences

From a given set of clone or gene identifiers in a gene battery we first have to construct a corresponding set of DNA sequences that potentially harbor the TFBSs needed to confer the gene battery specific expression pattern. This can be the region 5' upstream of the transcription start site (i.e., the proximal promoter or proximal module) or it can be one or more regulatory regions that lie several kilobases further upstream or downstream of the gene sequence or within introns. The optimal sequence set that we want to approximate is the set of all the functional regulatory regions in Figure 2.2.B labeled as "modules".

4.2.1 Proximal promoters

As opposed to prokaryotes or lower eukaryotes like yeast, the proximal promoter of metazoan genes is not located just upstream of the ATG translation start codon. Instead there can be a long 5' untranslated region (5'UTR) between the TSS and the beginning of the coding sequence. One way to find the proximal promoter is by using promoter prediction algorithms (see Section 2.4.2), but because of (1) the low sensitivity and specificity of these algorithms (certainly at the beginning of this work in 2001; confer reviews by Fickett et. al. [105] and Pedersen et. al. [229]) and (2) the species specificity of most of these algorithms (the prediction of mammalian promoters, other vertebrate promoters, or invertebrate promoters is different), we decided to follow another route.

We investigated whether the annotation of the gene start in genome databases like Ensembl are accurate enough to be used as TSS. The start annotations are derived from the mapping of available cDNA transcripts to the genomic sequence. If for a significant number of genes the start of the Exon 1 annotation (this is the gene start) in Ensembl coincides with or lies close to the TSS of the gene, then the sequence directly upstream of Exon 1 would contain the promoter-proximal sequence that we are interested in. Instead of a statistical test, we used the following observation as an indication that for many genes this assumption is valid.

We retrieved 2000 bp upstream of Exon 1 for 4,000 randomly selected genes from the human genome. First, we calculated the percentages of A, C, G, and T at each position in this stretch of DNA (see Figure 4.1.A).

The G/C content rises sharply when approaching position 1 of Exon 1 and drops again after this position. We cannot think of any other DNA signal with such an impact on the GCAT content other than the TSS. A similar finding can be observed in the regions upstream of the ATG start codon in yeast [46]. Since this figure is intriguing, we decided to analyse it further from a more

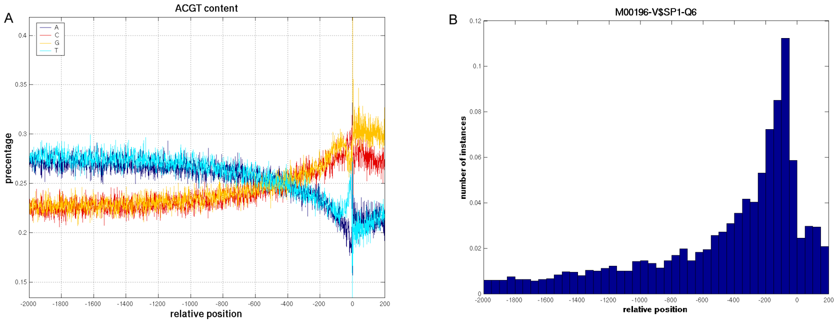


Figure 4.1: Proximal sequence detection. Representation of the genomic region 2,000 base pairs upstream of Exon 1 annotation in Ensembl and 200 base pairs after the start of Exon 1, taken from 4,000 randomly selected genes from the human genome (Ensembl release 8). The relative position of 0 on the x axis is the start of Exon 1. (A) Percentages of A, C, G, and T at each position. The average base composition changes dramatically in the sequence region upstream of the annotated gene start. The GC content increases towards the TSS and around the TSS nucleotide skews are observable (explained in more detail in Chapter 7). (B) The number of instances of SP1 binding sites increases towards the annotated gene start. Both observations indicate that, on average, the annotated gene start position (position 0 on the x axis) corresponds to the TSS.

general perspective of genomic nucleotide composition, and to place it in perspective with other findings that appeared in the literature, in Chapter 7. A second related observation to confirm the gene start statement is the rise in the number of putative SP1 binding sites that occur within these 4,000 regions (see Figure 4.1.B). Since SP1 is known to be a proximal *cis*-acting factor [69], this analysis shows that it is likely that the first 500 bp upstream of the TSS are predominantly promoters. Note however that these two observations pose a chicken-and-egg problem: the G/C rise can cause the presence of more Sp1 binding sites (they are G/C rich), or the presence of Sp1 binding sites can cause the G/C rise. In Chapter 7 we will show that the G/C rise is actually caused by the concentration of CpG dinucleotides around the TSS.

Because the goal of the analysis is to find over-represented motifs in sets of genes, and not in individual genes, it is still acceptable that for some genes in the set we would not have the correct promoter-proximal sequences if for these the start of Exon 1 would not be the TSS (e.g., if longer and yet unknown transcripts exist). This would also be the case when using promoter prediction algorithms. Unlike the latter, this is a very general approach that can be applied to every sequenced organism that is available in the Ensembl database and we can expect that the gene start annotations will only improve during the next years when more cDNAs, ESTs, and transcript sequences determined by 5' capping technique [287] will be mapped to the genomic sequence. In conclusion, for our *fast* strategy we will use the Ensembl gene starts as TSS and use ~400-

800 bp upstream of this site for those analyses where we wish to find TFBSs in the proximal promoter. In more detailed analysis, we will always be able to check the promoters using promoter predictions or the prediction of CpG islands. In practice however, we will focus more on the distal regulatory regions since it is expected that these harbor more process-specific (development, tissue, etc.) binding sites than the proximal promoter.

Note that recently, specialized databases like DBTSS [287], have appeared that contain the experimentally determined TSS for 8,793 genes (as of January 2004), as determined by sequencing with the oligo-capping method. Suzuki and collaborators [287] found that 34.2 % of RefSeq sequences should be extended towards the 5' end. The availability of such high quality data does not make the retrieval of promoter sequences from Ensembl superseded since it is expected that the DBTSS data will be used in the annotation process of Ensembl. Currently, DBTSS only contains TSS annotations for human genes, and not for all genes, and is biased towards moderately to highly expressed genes (see Chapter 7). We will compare the nucleotide frequencies between sequences with DBTSS-defined TSSs and Ensembl annotated gene starts in Chapter 7.

4.2.2 Distal regulatory sequences

To predict other putative regulatory regions that lie more distal from the TSS, the phylogenetic footprinting part comes in. Genomic sequences of the genes and their flanking regions are aligned with the same regions of orthologous genes, and the conserved non-coding sequences (CNS) are then considered to have a putative regulatory function as discussed in Section 2.7.3. Thus out of the hundreds of kb non-coding sequence around and within a gene, we will only use the CNSs in the search for distal binding sites. This selection reduces the search space for a single gene greatly (e.g., a 100 times reduction for a gene with 5 modules of 200 bp within 100,000 bp of intergenic sequence). Although this procedure causes the potential loss of certain regulatory regions, it increases the signal-to-noise ratio to acceptable levels as the case studies below will show. Note that in case the upstream sequence of two orthologous genes are aligned, the proximal promoter can as well be a CNS so using CNSs as putative regulatory regions is not restricted to distal regions only.

To detect CNSs, we used the specialized alignment algorithm AVID [42], together with its visualization tool VISTA [92, 209]. A Perl script takes two FastA[1] formatted sequences as input, executes AVID, passes the output to VISTA, and translates the VISTA output to GFF (General Feature Format, see http://www.sanger.ac.uk/Software/formats/GFF). Figure 4.2 shows an example of the GFF output of the Perl script for the alignment of the *ATOH1* gene for which the VISTA plot was shown in Figure 4.8.

---

[1] The FastA sequence format consists of a sequence name and description on a single line starting with the greater than symbol ">", followed by the sequence.

```
ENSG00000172238      VISTA   misc_feature    9640    9942    86.3    .   .   .
ENSG00000172238      VISTA   misc_feature    9943    11198   83.6    .   .   .
ENSG00000172238      VISTA   misc_feature    11301   11486   76.6    .   .   .
ENSG00000172238      VISTA   misc_feature    11723   12045   79.8    .   .   .
ENSG00000172238      VISTA   misc_feature    14361   15012   87.5    .   .   .
ENSG00000172238      VISTA   misc_feature    15494   16054   86.2    .   .   .
ENSMUSG00000046241   VISTA   misc_feature    9793    10094   86.3    .   .   .
ENSMUSG00000046241   VISTA   misc_feature    10120   11359   83.6    .   .   .
ENSMUSG00000046241   VISTA   misc_feature    11388   11570   76.6    .   .   .
ENSMUSG00000046241   VISTA   misc_feature    11775   12102   79.8    .   .   .
ENSMUSG00000046241   VISTA   misc_feature    14183   14812   87.5    .   .   .
ENSMUSG00000046241   VISTA   misc_feature    15170   15730   86.2    .   .   .
```

Figure 4.2: Conserved non-coding sequences. Example of the GFF output of the alignment of the human ATOH1 gene and the mouse MATH1 gene using the AVID algorithm and the VISTA interpretation software. The columns in the GFF file are: <seqname> - <source> - <feature> - <start> - <end> - <score> - <strand> - <frame> - <attribute>. In TOUCAN the fastA formatted sequences are sent to the alignment service and this GFF file is sent back to the client and is annotated on the active sequence set.

4.3 Detection of transcription factor binding sites

After the construction of a set of CNSs from the set of gene identifiers of a gene battery, we wish to detect the TFBSs that could confer the battery-specific expression pattern. We have chosen to restrict ourselves to TFs for which the binding sites have been determined experimentally. As explained in section 2.4.3, they can be modeled with position specific frequency matrices (PSFM). In a first step we will score all sequences in the set with all available PSFMs and annotate all PSFM instances as putative TFBSs. To this end we will use either of two methods described further in this section, the MotifLocator (see Section 4.3.3) and the MotifScanner (see Section 4.3.4) that both use higher-order background models instead of single nucleotide frequency models that are used in PWMs. An alternative approach—not discussed here but also integrated within the TOUCAN framework—is the *de novo* discovery of motifs from the sequence set as over-represented DNA words using the Gibbs sampling implementation named MotifSampler [298].

4.3.1 PSFM databases

We have transformed third-party and publicly available collections of count matrices to PSFMs (with pseudocounts, see Section 2.4.3) in our proprietary format to be used by the MotifScanner and MotifLocator scoring algorithms. The most important collection is TRANSFAC [323], for which the professional release 7.3 (Sep 2003) contains 674 count matrices and the public release 6.0 contains 336 matrices (Jan 2001). In the case studies we have used the professional releases, but external TOUCAN users have only access to the public release because of licensing restrictions. Recently, another high quality and non-redundant (the TRANSFAC collection is redundant) collection of count matrices has become available, namely the JASPAR database [252] and it will be used in Chapter 6.

4.3.2 Higher-order background models

Higher-order nucleotide frequency models have been used in gene prediction (e.g., higher-order HMM in GeneMark.hmm [195]) and in *de novo* motif detection algorithms [297, 204] to model the noisy sequence background in which the motifs are hidden. We have adopted this approach for the prediction of instances of PSFMs, as a first subtle way to reduce the number of false positive predictions [296].

When referring to higher-order background models, we start from the basic assumption that a DNA sequence can be generated with a Markov model of order m. This means that the probability of observing a certain nucleotide in a sequence depends on the m previous nucleotides in the sequence. The likelihood of a sequence S being generated with a higher-order background model of order m can then be written as

$$P(S|\mathcal{B}_m) = p(b_1, b_2, \ldots, b_m) \prod_{l=m+1}^{L} p(b_l|b_{l-1}, \ldots, b_{l-m}), \tag{4.1}$$

where $\mathcal{B}_m$ represents the parameters of the higher-order background model. These parameters are $p(b_1, b_2, \ldots, b_m)$, the probability of finding a specific m-mer, and $p(b_l|b_{l-1}, \ldots, b_{l-m})$, the probability of finding the base b_l given the m previous bases in the sequence. The latter is stored in the transition matrix of the Markov model.

To construct a transition matrix for a background model of order m, we count all oligonucleotides of length $m+1$ in a reference data set. We rearrange the counts in a matrix of dimension $m \times 4$, such that each row has the same first m bases while each column corresponds to the last base in the oligonucleotides. Next, a pseudocount is added and each row is normalized to one so that they represent probabilities.

Table 4.1 gives an example of a transition matrix constructed from *H. sapiens* conserved non-coding sequences with *M. musculus*. Each entry in the matrix represents the probability of finding the respective nucleotide given the two preceding nucleotides in the sequence. For comparison, we also included the single nucleotide frequency (SNF) in this table. Most rows are significantly different from the SNF. This has a profound impact on the computed probabilities especially if the sequences become longer.

As an example, we look at the probability of the sequences `AAAAAAA` and `CGCGCGC` being generated by this second-order background model using Equation 4.1.

$$\begin{aligned} P(\texttt{AAAAAAA}|\mathcal{B}_m) &= P(\texttt{AA})P(\texttt{A}|\texttt{AA})P(\texttt{A}|\texttt{AA})P(\texttt{A}|\texttt{AA})P(\texttt{A}|\texttt{AA})P(\texttt{A}|\texttt{AA}) \\ &= 0.068 \times 0.343 \times 0.343 \times 0.343 \times 0.343 \times 0.343 \\ &= 3.23e-4 \qquad (\text{SNF} = 4.31e-5) \end{aligned}$$

Table 4.1: Second-order background model from human-mouse conserved non-coding sequences (CNS).

	$P(bb)$[a]	$P(b\|bb)$[b]			
		`A`	`C`	`G`	`T`
`AA`	0.068	0.3436	0.1689	0.2611	0.2262
`AC`	0.046	0.3329	0.2807	0.1101	0.2761
`AG`	0.073	0.2566	0.2612	0.3079	0.1741
`AT`	0.049	0.2106	0.2124	0.2654	0.3114
`CA`	0.069	**0.2147**[c]	0.2277	0.3623	0.1951
`CC`	0.083	0.2538	0.3189	0.1613	0.2658
`CG`	0.037	0.1433	0.3534	0.3651	0.1380
`CT`	0.072	0.1212	0.2929	0.3353	0.2503
`GA`	0.059	0.2936	0.1842	0.3469	0.1752
`GC`	0.072	0.2347	0.3210	0.1882	0.2560
`GG`	0.083	0.2393	0.2823	0.3226	0.1556
`GT`	0.046	0.1674	0.2517	0.3260	0.2546
`TA`	0.039	0.3219	0.1921	0.2300	0.2558
`TC`	0.061	0.2635	0.3377	0.0904	0.3081
`TG`	0.067	0.2302	0.2420	0.3028	0.2249
`TT`	0.068	0.1860	0.2571	0.2190	0.3377
	SNF :	0.2379	0.2631	0.2619	0.2369

[a] Probability of finding dimer in the CNSs. [b] Representation of second-order transition matrix. [c] An example of how this table is read: the number 0.2147 in bold is the probability of observing an A (the column) after a CA (the row).

$$\begin{aligned}
P(\texttt{CGCGCGC}|\mathcal{B}_m) &= P(\texttt{CG})P(\texttt{C}|\texttt{CG})P(\texttt{G}|\texttt{GC})P(\texttt{C}|\texttt{CG})P(\texttt{G}|\texttt{GC})P(\texttt{C}|\texttt{CG}) \\
&= 0.037 \times 0.353 \times 0.188 \times 0.353 \times 0.188 \times 0.353 \\
&= 5.75e-6 \qquad (\text{SNF} = 8.61e-5)
\end{aligned}$$

This example illustrates that there are differences between the scores from the higher-order background model and the single nucleotide frequency model. For the shown examples, these differences make sense because both poly-A oligonucleotides and GC doublet containing sequences have a functional meaning in the genome and are either over-represented (poly-A) or under-represented (GC) in the genome.

The classical method to find instances of a known motif model is to transform the matrix of counts to a position-specific weight matrix (PWM) where single nucleotide frequencies are used to calculate the weights (see Section 2.4.3). The introduction of a higher-order background model into the scoring can be done at the level of the PSFM. Namely, given the PSFM, $\mathbf{\Theta}$, and the background model, $\mathcal{B}_m$, we can compute the score of the segment being generated by the motif model and compare this with the score of the segment being generated by the background model. For each segment $\mathbf{x}$ of length W in the sequence S, we compute the corresponding score as

$$W(\mathbf{x}) = \log\left(\frac{P(\mathbf{x}|\mathbf{\Theta})}{P(\mathbf{x}|S,\mathcal{B}_m)}\right) = \sum_{j=1}^{W}[\log(\boldsymbol{\theta}_j^{b_j}) - \log(P(b_j|S,\mathcal{B}_m))],$$

where $\boldsymbol{\theta}_j^{b_j}$ is the probability of observing base b_j at position j in segment x of length W.

4.3.3 MotifLocator

As is done in the classical detection using PSFMs, we can apply a threshold to these scores and select those segments with a score above the threshold as putative TFBSs. To define a reliable threshold over different motif models, we need to normalize the scores. The preferred method is to rescale the scores such that they have values between 0 and 1. First we compute the minimal and maximal value of $W(\mathbf{x})$ over all possible segments $\mathbf{x}$ as

$$\begin{aligned} W_{\min} &= \min_{\mathbf{x}} W(\mathbf{x}) \\ W_{\max} &= \max_{\mathbf{x}} W(\mathbf{x}) \end{aligned}$$

Once minimum and maximum are found, the scores $W(\mathbf{x})$ are rescaled as

$$\bar{W}(\mathbf{x}) = \frac{W(\mathbf{x}) - W_{\min}}{W_{\max} - W_{\min}}. \tag{4.2}$$

This results in a distribution of scores over the full sequence set, with scores between 0 and 1. On these scores we can impose a threshold and select all instances with a score higher than this threshold.

4.3.4 MotifScanner

A more sophisticated way of selecting PSFM instances instead of an arbitrary threshold, is implemented in the MotifScanner algorithm [296]. A probabilistic sequence model is used to estimate the number of instances Q of a motif model $\mathbf{\Theta}$ in a sequence S given the background model $\mathcal{B}_m$. The expected number of instances can be computed as

$$E_{S,\mathbf{\Theta},\mathcal{B}_m}(Q) = \sum_{c=0}^{\infty} c \times P(Q = c|S,\mathbf{\Theta},\mathcal{B}_m). \tag{4.3}$$

To compute Equation 4.3 we need to estimate the probability $P(Q = c|S,\mathbf{\Theta},\mathcal{B}_m)$ of finding c instances of the motif in the noisy background sequence. Applying Bayes' rule to this probability leads to

$$P(Q = c|S,\mathbf{\Theta},\mathcal{B}_m) = \frac{P(S|Q = c,\mathbf{\Theta},\mathcal{B}_m)P(Q = c|\mathbf{\Theta},\mathcal{B}_m)}{P(S|\mathbf{\Theta},\mathcal{B}_m)}. \tag{4.4}$$

We can distinguish three different parts in Equation 4.4. The denominator $P(S|\mathbf{\Theta}, \mathcal{B}_m)$ serves as the normalization factor. The first factor $P(S|Q = c, \mathbf{\Theta}, \mathcal{B}_m)$ of the numerator is the probability that the sequence is generated by the motif model $\mathbf{\Theta}$, the background model $\mathcal{B}_m$, and contains c motif instances. This probability can be calculated in linear time by summing over all possible non-overlapping combinations of c motifs in sequence S,

$$P(S|Q = c, \mathbf{\Theta}, \mathcal{B}_m) = \sum_{a_1} \cdots \sum_{a_c} \Big(P(S|\mathcal{A}_c, Q = c, \mathbf{\Theta}, \mathcal{B}_m) P(\mathcal{A}_c|Q = c, \mathbf{\Theta}, \mathcal{B}_m) \Big), \tag{4.5}$$

with $\mathcal{A}_c$ the set of c start positions $a_1, \ldots, a_c$. Assuming that each position is equally probable, the factor $P(\mathcal{A}_c|Q = c, \mathbf{\Theta}, \mathcal{B}_m)$ is replaced by a constant inversely proportional to the number of possible combinations of c motif instances in a sequence of length L. Within this model, we see the motif instances in the context of the noisy background sequence. This implies that the longer the sequence is the harder it is to find an instance within this noise. Therefore in a long sequence only those instances that have a very high score with the motif model do rise above the noise level and can be selected.

The second factor $P(Q = c|\mathbf{\Theta}, \mathcal{B}_m)$ in the numerator is the prior probability of finding c instances given the motif model and the background model. Let us define $P(Q = c|\mathbf{\Theta}, \mathcal{B}_m)$ as $\gamma(c)$. Since the complete prior distribution is not known, we propose one. There are two conditions to construct this distribution: (1) $\sum_{c=0}^{\infty} \gamma(c)$ should be equal to 1; (2) for all $c > 1$, $\gamma(c + 1)$ is smaller than $\gamma(c)$. The user should define only $\gamma(1)$, a value between 0 and 1, the probability of finding 1 instance. Initially $\gamma(0)$ is set to $1 - \gamma(1)$ and the remainder of the distribution $\gamma(c)$ is set to $\kappa\gamma(c - 1)$ and the distribution is then normalized. We use $\kappa = 0.25$. The effect of lowering the prior is that $E[Q]$ decreases and that fewer instances will be selected. Normally, we should compute the sum in Equation 4.3 for c from 0 to ∞. Since this is unpractical, we propose to compute the next term in the distribution $P(Q = c|S, \mathbf{\Theta}, \mathcal{B}_m)$ as long as the previous value is larger than a predefined small value ϵ (e.g., 0.0001).

Given the previously defined formulas we can find the number of instances of $\mathbf{\Theta}$ in sequence S using the procedure in Program 1. The algorithm only needs one parameter to be set: the prior $\gamma(1)$. The effect of this parameter on the performance of the algorithm, the complexity of the algorithm, and a comparison with the MotifLocator can be found in [296, pp. 119-144].

4.3.5 Discussion

We use this probabilistic model to estimate the number of instances of a motif in a specific sequence given the background model and the motif model, instead of using a predefined threshold that is independent of the sequence being scored. The advantages of this method can be summarized as follows. Firstly, by choosing an appropriate background model for the sequences to be scored we can reduce the number of false positive hits [204]. For example, when scoring human promoter sequences using a 3rd-order background model that is calcu-

Program 1 The MotifScanner algorithm.

```
• Input: sequence in FastA format
• Initialization of the prior distribution $[[1 - \gamma(1), \gamma(1)]]$ and $i$=0
• Algorithm
    1. Score each segment $\mathbf{x}$ in $S$ with the motif model $\boldsymbol{\Theta}$.
    2. Score each segment $\mathbf{x}$ in $S$ with background model $\mathcal{B}_m$.
    3. Compute $P(Q = 0|S, \boldsymbol{\Theta}, \mathcal{B}_m)$ and $P(Q = 1|S, \boldsymbol{\Theta}, \mathcal{B}_m)$.
    4. While $P(Q = i|S, \boldsymbol{\Theta}, \mathcal{B}_m) > \epsilon$
        (a) Increment $i$
        (b) Update $P(Q = c|S, \boldsymbol{\Theta}, \mathcal{B}_m)$ for $c = 0, \ldots, i$.
    5. Compute the expected number of instances as $E_{(S,\boldsymbol{\Theta},\mathcal{B}_m)}[Q]$ according
       to equation 4.4.
    6. Select the $Q$ best scoring positions as motif instances.
• Output: start and stop positions, and score of each motif instance, in GFF
  format.
```

lated from a large set of human promoters, a putative motif instance would need a higher resemblance to the PSFM to be a positive hit than when using a background model of mouse sequences to score the same human promoters. Another example can be given by the fact that a A/T rich motif scores higher with MotifScanner in a G/C rich context than in a A/T rich context. Because the presence of a motif in a dissimilar context can imply a functional conservation we believe that its detection should indeed be promoted. Secondly, by estimating the number of motif instances instead of using a threshold, only the best matching instances are considered as hits instead of all the instances that score above a certain threshold. This approach reflects the situation in a cell where a transcription factor is bound more often to the stronger sites than to the weaker sites. However, if one also wishes to select the weaker sites then the prior parameter can be increased, the sequences trimmed (to remove noise), or the MotifLocator can be used.

The disadvantages of the MotifScanner algorithms are (1) the prior is difficult to interpret and (2) which and how many PSFM instances are retained as hits depends on the length of the sequence. To control the latter, a good choice of the prior is necessary. For ∼500 bp sequences like proximal promoters, a prior of 0.2-0.5 is suitable, while for 200 bp sequences like CNSs, a prior of 0.05-0.2 is better. Although these parameter settings have resulted from thorough testing [296], a possible improvement of the MotifScanner could be the automatic determination of the prior parameter from the sequence length.

In this chapter we have only used the MotifScanner, not the MotifLocator, to detect TFBSs. We have only used the MotifLocator briefly in Chapter 5 in one of the validation procedures.

4.4 Statistical test for over-representation

A binomial distribution model is used to correlate all PSFMs that have at least one instance in the sequence set—as determined with the MotifScanner—with a p value and a significance score based on the number of PSFM instances in the sequence set relative to the expected number of instances that is based on the expected frequency of the PSFM. This statistical test with the calculation of a p-value and a significance score for each motif was done as described in [307], where it was developed to detect over-represented hexanucleotides within the upstream regions of families of coregulated genes in yeast.

The expected frequencies for all PSFMs are calculated by scoring large reference sequence sets with all available PSFMs using the MotifScanner. The reference sets can for example be all experimentally defined promoter sequences of the Eukaryotic Promoter Database (EPD, see http://www.epd.isb-sib.ch/), or the 500 bp 5' upstream sequence of all genes in Ensembl, or all (or a set of randomly selected) CNSs in all (or a set of randomly selected) genes in Ensembl, etc. The number of occurrences in the reference set divided by the number of base pairs where an occurrence can begin (i.e., almost the total number of bp in the reference set) is used as the expected frequency for each motif m, $F_e\{m\}$. The expected frequencies are used to calculate the expected number of occurrences for each motif in the set of regulatory regions under analysis:

$$E(\mathrm{occ}\{m\}) = F_e\{m\} \times 2 \times \sum_{i=1}^{N_S}(L_i - w + 1) = F_e\{m\} \times T,$$

where T is (by definition) the number of possible start positions, L_i is the length of the i^{th} sequence, N_S is the number of sequences in the set, and w is the length of the motif. The probability to observe exactly n occurrences of the motif m is estimated by the binomial formula:

$$P(\mathrm{occ}\{m\} = n) = \frac{T!}{(T-n)! \times n!} \times (F_e\{m\})^n \times (1 - F_e\{m\})^{T-n}.$$

The probability to observe n or more occurrences of the motif m is:

$$P(\mathrm{occ}\{m\} \geq n) = \sum_{j=n}^{T} P(\mathrm{occ}\{m\} = n).$$

A significance coefficient *sig* is used to select the most over-represented patterns:

$$\mathrm{sig} = -\log_{10}[P(\mathrm{occ}\{m\} \geq n) \times D]$$

where D is the number of distinct motifs that are used. The highest values for this parameter correspond to the most over-represented patterns. When selecting only the patterns for which sig > 0, one expects less than one pattern to occur at random within each possible sequence set. Each increment of 1 for the significance coefficient represents a drop of a factor of 10 for the occurrence probability. In other words, one expects to find at random one pattern with sig > 1 every ten sequence sets, one with sig > 2 every 100 sets, and one with sig $> s$ every 10^s sets [307].

The choice of the reference set to calculate the expected frequencies is important, both functionally (e.g., proximal versus distal) and technically. The latter refers to the dependency of the MotifScanner hits on the length of the sequence being scored. To get unbiased p-values, the sequences in the reference set should be of the same length as the sequences in the set under study. At the ftp site ftp://ftp.esat.kuleuven.ac.be/pub/sista/aerts/software/freqfiles/ we provide several files with expected frequencies calculated from EPD or from the upstream regions or CNSs of random gene subsets from complete genomes.

4.5 TOUCAN

The software tool TOUCAN that was fully developed in this work is the embodiment of the above mentioned strategy that integrates the methods for sequence selection, TFBS detection, and TFBS over-representation scoring. Figure 4.3 shows local and remote components and the relationships with local and external databases.

A generalized regulatory sequence analysis in Toucan can be described as a sequence of the following steps:

1. Loading a local sequence file (in fastA, EMBL or GenBank format) or sequence retrieval from Ensembl using gene identifiers (Ensembl stable gene ID, HUGO gene name, LocusLink ID, or any other identifier that is mapped to an Ensembl gene in the Ensembl database), and specifying (a) the species, (b) the sequence wanted (upstream of Exon 2, upstream of the CDS, or whole gene with flanking sequences), and (c) the species for which the orthologous sequence should be retrieved if it is known;

2. CpG island prediction[2] in proximal sequences if required;

3. Alignment of orthologous sequences to annotate CNSs;

4. Sequence manipulation, for example selecting a number of proximal regions or CNSs into a new sequence set;

5. TFBS prediction on this set using MotifScanner, specifying which PSFM collection to use, the prior, and the background model;

---

[2] A CpG island is a region of at least 200 bp with a G/C content over 60% and a CG doublet frequency that is at least 1.667 times the expected genomic CG frequency.

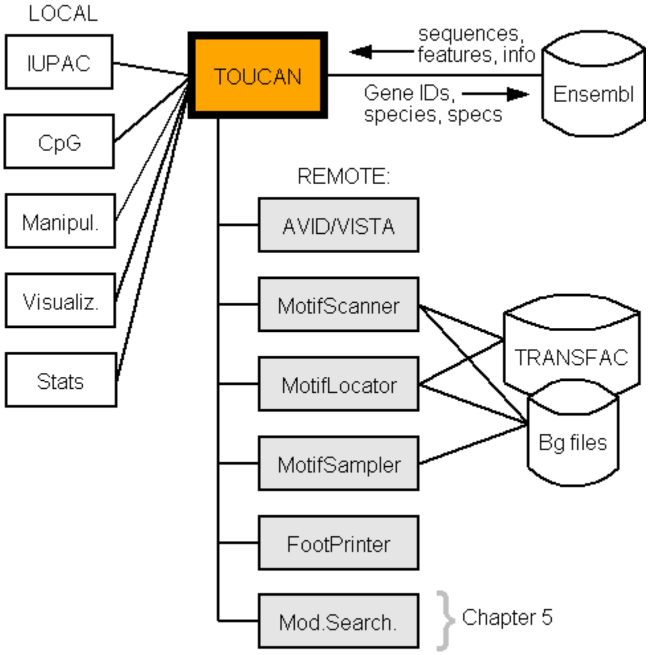


Figure 4.3: Overview of the functional components of TOUCAN. URLs for the remote tools can be found in appendix B.

6. Binomial analysis and sorting of all PSFMs according to their *sig* value[3]. An alternative analysis step instead of the combination of (5) and (6) can be the detection of over-represented sequence motifs using the MotifSampler.

In Figure 4.4 several screenshots are shown that capture some of these analysis steps in action.

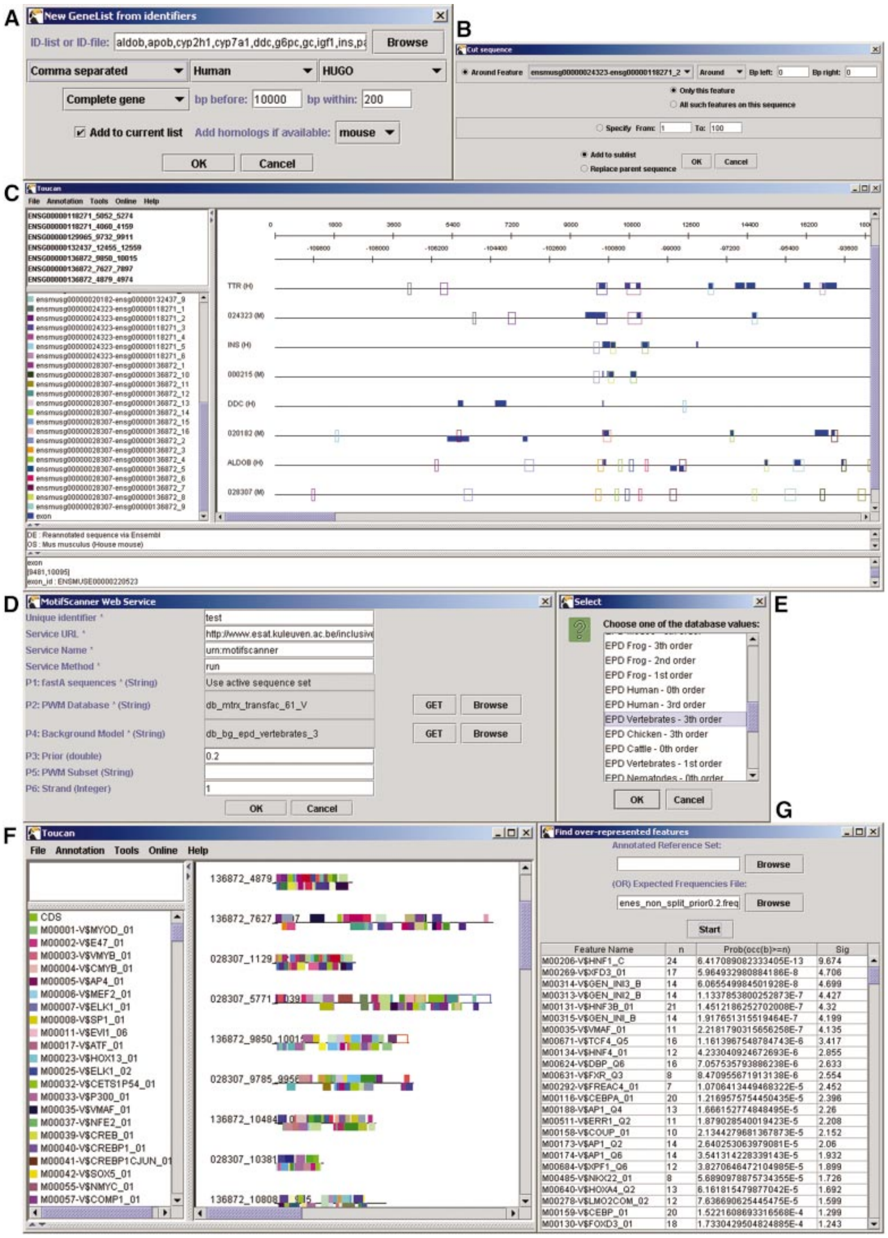
TOUCAN technicalities

Figure 4.5 summarizes the computational system from an IT perspective.

TOUCAN has a Graphical User Interface (GUI) implemented in Java (Sun Microsystems). The application can be started directly from our web site using Java Web Start (JWS, http://java.sun.com/products/javawebstart/) and has been tested under the Windows, Linux, and MacOS operating systems. The layer of “business logic” is implemented in the Java package `sista.sequence.*` and uses the `BioJava` package (http://www.biojava.org) for most sequence handling tasks, and the `ensj_core` package of Ensembl to access genomic sequences and annotations in the Ensembl database.

---

[3] In Chapter 5 the binomial analysis is replaced by searching for the optimal *combination* of TFBS predictions using the ModuleSearcher algorithm.



New GeneList from identifiers

ID-list or ID-file: aldob,apob,cyp2h1,cyp7a1,ddc,g6pc,gc,igf1,ins,pa

Comma separated | Human | HUGO

Complete gene | bp before: 10000 | bp within: 200

☑ Add to current list | Add homologs if available: mouse

Cut sequence

● Around Feature ensmusg00000024323-ensg00000118271_2 | Around | Bp left: 0 | Bp right: 0

● Only this feature
○ All such features on this sequence

○ Specify From: 1 To: 100

● Add to sublist
○ Replace parent sequence

Toucan

ENSG00000118271_5052_5274
ENSG00000118271_4060_4159
ENSG00000129965_9732_9911
ENSG00000132437_12455_12559
ENSG00000136872_9850_10015
ENSG00000136872_7627_7897
ENSG00000136872_4879_4974

ensmusg00000020182-ensg00000132437_9
ensmusg00000024323-ensg00000118271_1
ensmusg00000024323-ensg00000118271_2
ensmusg00000024323-ensg00000118271_3
ensmusg00000024323-ensg00000118271_4
ensmusg00000024323-ensg00000118271_5
ensmusg00000024323-ensg00000118271_6
ensmusg00000028307-ensg00000136872_1
ensmusg00000028307-ensg00000136872_10
ensmusg00000028307-ensg00000136872_11
ensmusg00000028307-ensg00000136872_12
ensmusg00000028307-ensg00000136872_13
ensmusg00000028307-ensg00000136872_14
ensmusg00000028307-ensg00000136872_15
ensmusg00000028307-ensg00000136872_16
ensmusg00000028307-ensg00000136872_2
ensmusg00000028307-ensg00000136872_3
ensmusg00000028307-ensg00000136872_4
ensmusg00000028307-ensg00000136872_5
ensmusg00000028307-ensg00000136872_6
ensmusg00000028307-ensg00000136872_7
ensmusg00000028307-ensg00000136872_8
ensmusg00000028307-ensg00000136872_9
exon

DE : Reannotated sequence via Ensembl
OS : Mus musculus (House mouse)

exon
[9481,10095]
exon_id : ENSMUSE00000220523

MotifScanner Web Service

Unique identifier *	test
Service URL *	http://www.esat.kuleuven.ac.be/inclusiv
Service Name *	urn:motifscanner
Service Method *	run
P1: fastA sequences * (String)	Use active sequence set
P2: PWM Database * (String)	db_mtrx_transfac_61_V
P4: Background Model * (String)	db_bg_epd_vertebrates_3
P3: Prior (double)	0.2
P5: PWM Subset (String)	
P6: Strand (Integer)	1

Select

Choose one of the database values:

EPD Frog - 3th order
EPD Frog - 2nd order
EPD Frog - 1st order
EPD Human - 0th order
EPD Human - 3rd order
EPD Vertebrates - 3th order
EPD Chicken - 3th order
EPD Cattle - 0th order
EPD Vertebrates - 1st order
EPD Nematodes - 0th order

Toucan

CDS
M00001-V$MYOD_01
M00002-V$E47_01
M00003-V$VMYB_01
M00004-V$CMYB_01
M00005-V$AP4_01
M00006-V$MEF2_01
M00007-V$ELK1_01
M00008-V$SP1_01
M00011-V$EVI1_06
M00017-V$ATF_01
M00023-V$HOX13_01
M00025-V$ELK1_02
M00032-V$CETS1P54_01
M00033-V$P300_01
M00035-V$VMAF_01
M00037-V$NFE2_01
M00039-V$CREB_01
M00040-V$CREBP1_01
M00041-V$CREBP1CJUN_01
M00042-V$SOX5_01
M00055-V$NMYC_01
M00057-V$COMP1_01

Find over-represented features

Annotated Reference Set:

(OR) Expected Frequencies File: enes_non_split_prior0.2.freq

Feature Name	n	Prob(occ(b)>=n)	Sig
M00206-V$HNF1_C	24	6.417089082333405E-13	9.674
M00269-V$XFD3_01	17	5.964932980884186E-8	4.706
M00314-V$GEN_INI3_B	14	6.065549984501928E-8	4.699
M00313-V$GEN_INI2_B	14	1.1337853800252873E-7	4.427
M00131-V$HNF3B_01	21	1.4512186252702008E-7	4.32
M00315-V$GEN_INI_B	14	1.917651315519464E-7	4.199
M00035-V$VMAF_01	11	2.2181790315656258E-7	4.135
M00671-V$TCF4_Q5	16	1.1613967548784743E-6	3.417
M00134-V$HNF4_01	12	4.233040924672693E-6	2.855
M00624-V$DBP_Q6	16	7.057535793886238E-6	2.633
M00631-V$FXR_Q3	8	8.470955671913138E-6	2.554
M00292-V$FREAC4_01	7	1.0706413449468322E-5	2.452
M00116-V$CEBPA_01	20	1.2169575754450435E-5	2.396
M00188-V$AP1_Q4	13	1.666152774848495E-5	2.26
M00511-V$ERR1_Q2	11	1.879028540019423E-5	2.208
M00158-V$COUP_01	10	2.1344279681367873E-5	2.152
M00173-V$AP1_Q2	14	2.640253063979081E-5	2.06
M00174-V$AP1_Q6	14	3.541314228339143E-5	1.932
M00684-V$XPF1_Q6	12	3.8270646472104985E-5	1.899
M00485-V$NKX22_01	8	5.689097887573435E-5	1.726
M00640-V$HOXA4_Q2	13	6.161815479877042E-5	1.692
M00278-V$LMO2COM_02	12	7.636690625445475E-5	1.599
M00159-V$CEBP_01	20	1.5221608693316568E-4	1.299
M00130-V$FOXD3_01	18	1.7330429504824885E-4	1.243

Figure 4.4: Description on next page.

The algorithms in the grey boxes in Figure 4.3 are used through web services using the Apache implementation of SOAP (Simple Object Access Protocol), Version 2.3. The services that reside on a Tomcat server, start the actual programs on a Linux "numbercruncher" using Java Remote Method Invocation (RMI). The fact that these functions work remotely is transparent for the user. Generally, fastA formatted sequences together with the required parameters are sent to the service, and GFF formatted output is sent back to the GUI where the features are directly annotated and visualized on the corresponding active sequences. This setup ensures the advantages of a local installation (e.g., a higher user interactivity) and the advantages of distributed computing using web services [272].

4.6 Case studies

We have performed several analyses on human gene sets. The first two case studies serve to validate the proposed strategies and their implementations: (1) the automated retrieval of proximal promoter sequences from the Ensembl database, and the subsequent detection of over-represented TFBSs is tested on

---

Figure 4.4 (Previous page) Screenshots of TOUCAN during the analysis fof liver-specific genes. (A) Dialog where all gene names (HUGO symbols) are entered as a comma separated list. In the second drop-down box "Human" is selected to search for and retrieve human genes. All organisms that are available in Ensembl (see http://www.ensembl.org) can be chosen from this list, and in the "Preferences" menu the user can update these settings if Ensembl were to add new organisms. Depending on which organism is chosen, the third drop-down box shows all available external database identifiers that can be mapped to a stable Ensembl gene. The fourth drop-down box allows to choose between "complete gene", "upstream of CDS", and "upstream of Exon 1". The latter corresponds in most cases to the region upstream of the TSS. The text boxes labeled with "bp before" and "bp within" specify how many base pairs should be retrieved as flanking sequence upstream or around the specified region. In the last drop-down menu "mouse" is selected to retrieve also the mouse orthologous sequences for each human gene in the list. (B) Every region that seems likely to contain putative regulatory modules (e.g., because it is conserved between species or because it contains a CpG island), can be selected and added to a sequence sublist. (C) Feature map. All open boxes represent human-mouse aligned regions that are at least 75% identical, resulting from the AVID/VISTA web service. (D) Matrices, background model, and all other parameters are set to run the MotifScanner. (E) Dialog showing the background models on our server. The values are retrieved transparently through the web service when the user presses the "GET" button. (F) The results of the MotifScanner that can either be saved or can be automatically added as features on the currently active sequence set. (G) Results of the binomial formula to detect over-represented motifs, where n is the number of occurrences of a binding site within this set, the third column is the p-value for this motif, and the fourth column is the *sig* value. The top scoring motifs for the human-mouse conserved regions in 10kb upstream sequence of liver-specific genes are shown.

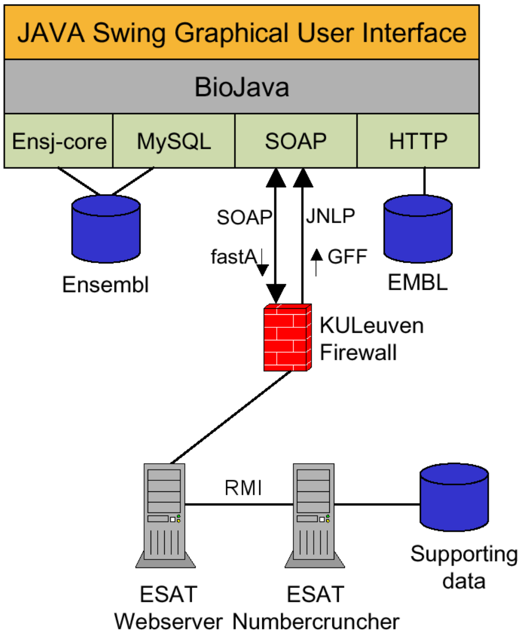


Figure 4.5: The TOUCAN software environment. The TOUCAN software tool is started by a user from his/her web browser using Java Web Start. The Graphical User Interface (GUI) that becomes visible is programmed to allow for flexible and user-friendly sequence manipulations and visualizations. The business logic (second tier) uses the BioJava open source library, which contains for example functions for feature annotations. The data access layer (third tier) uses the ensj-core library from Ensembl for the retrieval of sequences and annotation from the Ensembl database. The Simple Object Access Protocol (SOAP) is used to send XML messages to remote services on an ESAT server that start algorithms (see Figure 4.3) on another ESAT server. Communication between both servers is done with Remote Method Invocation (RMI).

a set of E2F target genes, for which the E2F binding sites are often located proximal to the TSS (deduced from TFBSs locations in TRANSFAC); and (2) the automated retrieval of long upstream sequences of all human-mouse orthologous gene pairs in a co-regulated gene set, followed by the selection of conserved non-coding sequences (CNS), and the detection of over-represented TFBSs is tested on two benchmark data sets of muscle and liver specific genes. The usage of CNSs is compared with known distal enhancers and with proximal promoters.

The third and fourth case study are performed in collaboration with molecular biology groups, and illustrate alternative usages of the TOUCAN software system and its algorithms.

4.6.1 E2F target genes

In this example, we investigated eight human genes of which the E2F complex is a known regulating transcription factor: *CAV1*, *CDC6*, *MYC*, *DHFR*, *E2F2*, *RBL1*, *TK1*, and *RB1*. Since E2F mostly binds to the proximal promoter of its target genes, a region of 500bp upstream of the putative TSS (start of Exon 1) was obtained using the direct Ensembl access within TOUCAN.

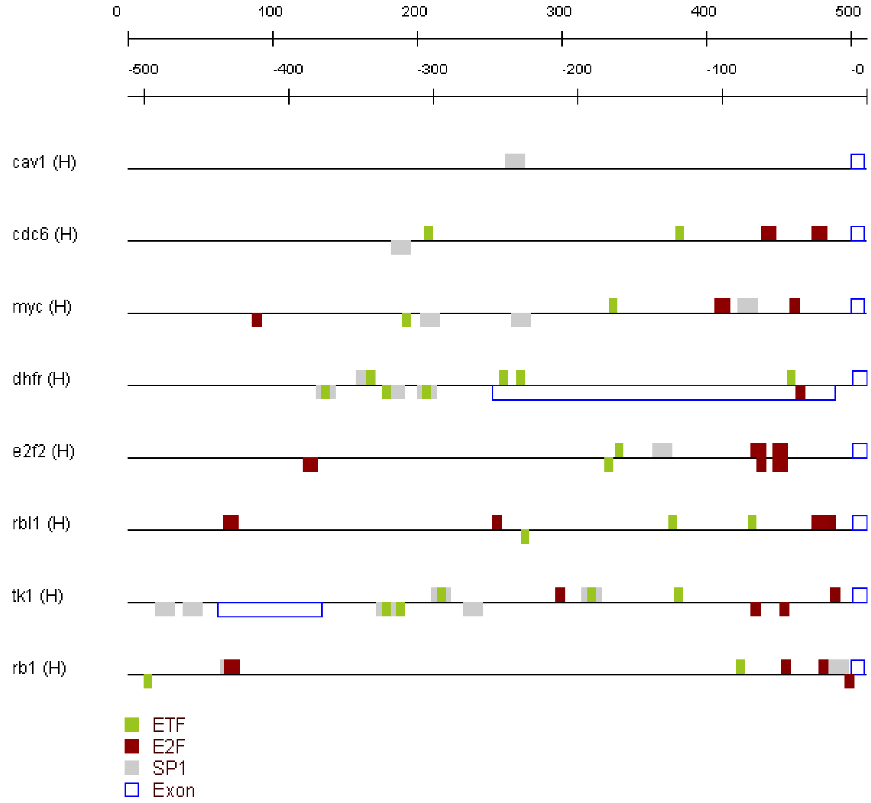


Figure 4.6: Promoter regions of eight E2F target genes with the over-represented TFBSs. The sequences were retrieved from Ensembl starting from a comma separated list of HUGO symbols and choosing "upstream of Exon 1", 500 "bp before", and 10 "bp within".

All retrieved sequences are visualized in a sequence feature map. Next we have scored these sequences with PSFMs that reside on our server by using the MotifScanner web service. Although a low prior (0.2) was used, most of the sequences are packed with putative binding sites. Running the binomial analysis we could select the significantly over-represented motifs. The expected frequencies needed for this statistic were calculated by scoring the same matrices on all human sequences in EPD. The presence of E2F, ETF, and SP1 was significant (sig ≥ 2). Figure 4.6 shows the sequence set with the instances of these motifs annotated. The presence of two to three putative ETF binding sites in almost all E2F target genes is interesting since this is also the case in the mouse P53 promoter, which is bound by E2F and ETF upon adenovirus infection in the presence of the Early 1a protein [130].

4.6.2 Liver and muscle genes

Wasserman and Fickett [316] and Krivan and Wasserman [168] have compiled and analyzed respectively muscle-specific and liver-specific regulatory regions that are experimentally verified. They found a significant occurrence of specific binding site clusters within these regions. We have tested the MotifScanner and the over-representation statistic three times: (1) on their training sets of known enhancers, (2) on sets of proximal promoters of the human genes represented in their training sets, and (3) on sets of CNSs of the same genes.

Known enhancers

The fastA formatted sequence files were downloaded from http://bio.cse.psu.edu/mousegroup/Reg_annotations/ and loaded straight into TOUCAN (after removing blanks within the sequences). We used the MotifScanner with the TRANSFAC collection of vertebrate matrices, a prior of 0.2, and a background model of vertebrate sequences of EPD.

Among the over-represented motifs, some are known to be muscle specific: SRF, myogenin, MYOD, MEF-2, MZF, MINI, and MEF-3; so their presence in these sequences is not surprising. The only muscle-specific factor that was used in [316] that we could not confirm with sig ≥ 2 is TEF. Some others can interact with muscle-specific factors: E12 (dimerizes with MYOD and myogenin of the Myf family) and HEB (interacts with E12 and myogenin). The finding that their actual binding sites are significantly present in these sequences is new. The detection of SP1 is not surprising since it is a general promoter element. Some of the remaining factors may not be muscle-specific but they may play a role in transcriptional regulation in certain circumstances. VDR (vitamin D receptor) for example is involved in the genomic response of avian embryonic skeletal muscle cells to vitamin D(3) [57], LMO2 (LIM-only protein) may play a role in differentiation and myofibrillogenesis of heart [184], and LBP-1 (UBP-1) binds at the promoter of skeletal troponin I [220]. For the remaining factors we could not find any references that point to regulation of muscle genes. These are MAZ (Pur-1, Zif87), MAZR (MAZ related factor), ZIC, and RREB (Ras-responsive element binding protein).

An analogous analysis on the set of liver-specific regions shows similar results, although fewer factors have over-represented sites. HNF1 and C/EBP were also used by Krivan and Wasserman [168] and are known to be liver-specific. Other significant factors include COUP, which may antagonize with HNF4 [211], and

IPF (Insulin Promoter Factor). Mutations in IPF or HNF both result in a common progression of maturity-onset diabetes of the young (MODY) [280]. They can therefore be interesting hypotheses. The last one is AP1, a general regulatory factor.

Proximal promoters

We used the same genes that were represented in the set of known regulatory sequences used above: for the muscle set these are *CHRM2, CHRM3, ACTC, CKM, DES, MYF6, MYOG, MYL1, MYLA, TNNI3, MYHCA, ACTA1, DMD, ANF,* and *ALDOA*; and for liver set these are *ALDOB, APOB, CYP2H1, CYP7A1, DDC, G6PC, GC, IGF1, INS, PAH, PROC, SLCA2, SULT2A2, SULT2A1, TTR, UGT1A1.* When using only 400 bp upstream of Exon 1 like in the E2F analysis, fewer elements were detected both for the muscle and for the liver genes (see Figure 4.7). For muscle, the highly significant elements are SRF and MAZR, and for liver HNF-1 and FOX (previously called HNF-3/forkhead transcription factors).

Conserved non-coding sequences

If we look at the location of the known regulatory regions relative to the TSS, we see that most of the regions are actually enhancers that lie further upstream, or even downstream of the TSS. We therefore retrieved, in a new analysis, 10 kilobases of sequence upstream of the translation start (start of CDS annotation) together with the same part of the mouse ortholog when such a correspondence was available. Figure 4.4 shows screenshots of several steps performed in TOUCAN of the analysis for the liver genes. For each pair of orthologous sequences we used AVID and VISTA to detect regions having minimal 75% of base identity in a sliding window of 100 bp. The regions located 5' upstream of the TSS or in the 5' UTR were selected and scored with the TRANSFAC collection of vertebrate matrices using the same parameter settings as before. The statistical analysis performed thereafter showed over-representation of HEB, LBP and MEF-2 in the muscle regions and HNF-3, HNF-4, C/EBP, COUP, and AP1 in the liver regions.

These are probably factors that bind to sites in distal modules rather than in the region just upstream of the TSS. There are also factors that were present in neither of the two other analyses: for muscle RSRFC4 (SRF-related), STAT6 (involved in hypercontractility of smooth muscle cells) and others without established muscle relatedness; for liver TCF4 (tumors arising in the liver can be caused by a complex of TCF4 and mutated beta-catenin), DBP (a member of the C/EBP family that is enriched in liver) and others without established liver relatedness. This shows that putative regulatory motifs can be detected computationally that have not been detected experimentally yet, which might be caused by the difficulty of mimicking every developmental and metabolic condition in the cell. The presence of factors without a direct link with the experimental setup can sometimes be due to the fact that they recognize sequences that are related to the sites of other factors. This is probably the case for v-MAF, which binds to AP-1 sites since v-MAF forms heterodimers with Fos and Jun (the consensus binding site of v-MAF is TGCTGACTCAGCA and the consensus site of AP-1 is GVTGACTCA, so they are very similar).

Conclusions for the muscle and liver genes

It is shown that the retrieval of orthologous sequences (here human and mouse) enables the selection of putative regulatory regions through comparative sequence analysis. Starting from upstream sequences tens of kilobases long, this selection narrows down the search region for regulatory modules to a couple hundred base pairs—this length restriction is essential for the detection of over-represented motifs (if not, the over-representation statistic is buried by the sequence noise). Reasonable results can be obtained by the detection of over-represented instances of available PSFMs: most of the TFs for which there are over-represented binding sites are related to muscle or liver specific gene expression respectively. Furthermore it is clear that some of the expected TFs can be found back in CNSs and not in proximal promoters, and the other way around (see Figure 4.7), and that some "statistically significant" motifs are probably still false positive predictions.

There are also limitations: (1) treating all CNSs of a single gene independently creates a lot of noise in the data set (this will be dealt with in the next Chapter); (2) the reference sequence set that is used to calculate the expected frequencies influences the results of the statistical analysis because of the length dependency of the MotifScanner (we did not show different results using different frequency files here); (3) the choice of which sequences to use is not always clear (e.g., use only human CNSs or human and mouse CNSs); (4) the collection of matrices is limited, the quality of the matrices is uncertain, and it is not certain whether a matrix model from TRANSFAC is able to generalize; and (5) there can be a significant redundancy in the PSFMs that are over-represented.

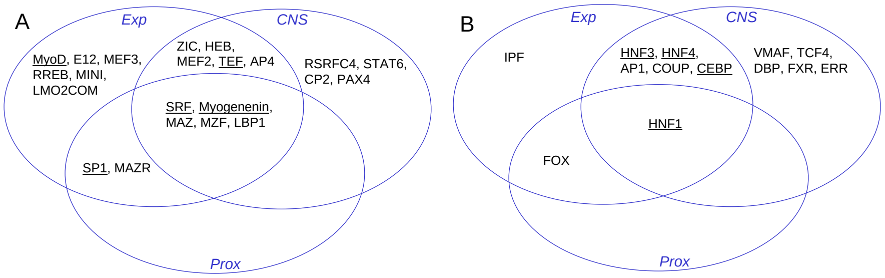


Figure 4.7: TFBS over-representation. Transcription factors for which the TFBSs are over-represented in the three examined sequence sets of muscle genes (A) and liver genes (B). *Exp* = experimentally determined enhancers; *CNS* = computationally detected conserverd non-coding sequences between 10 kb of the human and mouse orthologous upstream sequences; *Prox* = 400 bp sequence 5' upstream of the gene start as annotated in Ensembl.

4.6.3 TCF3-β-catenin target genes

In collaboration with the Center for Human Genetics we have analyzed two sets of genes that are differentially regulated in desmoid tumors (aggressive fibromatosis, locally invasive soft tissue tumors) as compared to normal fascia tissue. Sporadic desmoids harbor somatic mutations in either the *APC* gene or in the β-catenin gene—key components of the WNT signalling pathway—resulting in β-catenin protein stabilization and nuclear accumulation. Here it interacts with members of the TCF/Lef family of transcription factors to modulate transcription of target genes. In colorectal cancer it is the TCF4 member of this family that forms a complex with β-catenin, while in desmoids it is TCF3 [293]. The consensus binding site for the TCFs is WWCAAWG.

The selection of the genes that are differentially expressed was done using Affymetrix DNA chips [83]. For a gene to be selected as differentially expressed, it had to be expressed at least 2.5 fold higher or lower in the desmoid samples compared to the fascia samples and with a minimum difference in hybridization signal of 200. Where expression was below baseline, it was determined to be absent and set at 50, the background level. This way 33 genes were found to be up-regulated and 36 genes down-regulated. Of the down-regulated genes, IGFBP6 was shown to be a direct target of the TCF4–β-catenin [83].

If the TCF–β-catenin complex can regulate different target genes in colorectal cancer and desmoids, and if furthermore the same complex can up-regulate certain genes and down-regulate other genes, the *cis*-regulatory system involved has to be more complex than only the TCF binding site. As a first test we have checked whether the copy number of TCF binding sites in the proximal promoters (2000 bp upstream of Exon 1) of the two gene sets is different from the genomic frequency. The IUPAC annotation functionality (i.e., regular expression matching) of the TOUCAN framework was used to annotate instances of the WWCAAWG motif, in all 2000 bp upstream sequences of the genome and in the sets of differentially expressed genes. These expected and observed frequencies were used to calculate a p value with the binomial formula. Neither for the up-regulated gene set, nor for the down-regulated gene set, the copy number of motif instances was statistically over-represented.

Next we have tested whether flanking base pairs around putative TCF binding sites are conserved within these sets. In TOUCAN we annotated the proximal promoter sequences with the IUPAC annotation functionality and selected the TCF3 instances most proximal to the gene with 5 bp flanking sequences using the cut functionality. A sequence logo of the resulting set did not reveal any conserved flanking base pairs (not shown, see [82]). This however could also be due to the fact that not all genes in the sets are direct targets of TCF–β-catenin. Last, we hypothesized that other TFs could be involved that work together with the TCF–β-catenin complex. In TOUCAN we now selected the two WWCAAWG occurrences most proximal to the TSS with 100 bp flanking sequence on both sides. This set was scored with the MotifScanner and the complete TRANSFAC collection of PSFMs. The over-represented TFBS (sig >1)—as determined with the binomial analysis—were CDXA, OCT1, GATA2, OCT1, EN1, and STAT5A for the up-regulated genes and GEN-INI2, GEN-INI, IK-2, STAT5A, GEN-INI3 and ISRE for the down-regulated genes.

Although this analysis is an illustration of the usage of TOUCAN to detect TF cooperativity, the biological feasibility of this study is not so clear. It is not certain for example whether the gene sets contain enough direct targets to find over-represented motifs in the proximity of a TCF binding site and neither is it certain that among all the selected WWCAAWG occurrences there are enough functional binding sites. Therefore we can conclude that although some binding sites in the final sequence sets are *statistically* over-represented, their biological function is uncertain. The same analysis should optimally be performed on gene sets with more verified direct targets with verified TCF3 binding sites, which are currently not available.

4.6.4 Binding site detection without gene batteries: a case study in neurogenesis

In the previous examples we have used gene batteries to find over-represented TFBSs. In case we wish to find TFBSs in a single gene for which there are no coregulated or co-expressed genes known or available, we can fall back on pure phylogenetic footprinting (PF). A recently published method for PF is FootPrinter [35, 36], as described in Section 2.7.3. A limitation to this method however is that only small (the authors mention ∼1000 bp) sequences can be used. This poses a problem for metazoan sequences for which we know that several tens of kb can harbor the regulatory elements.

To solve this problem we have combined the two approaches of PF, namely the CNS detection and the motif discovery in the analysis of the regulation of the gene *atonal* in *Drosophila melanogaster*, in collaboration with the Laboratory for Neurogenetics of the K.U.Leuven. *atonal* gets its name[4] from the disruptive effects the gene's mutation has on chordotonal neuron differentiation[5]. Figure 4.8 shows all CNSs as open rectangles obtained by pairwise alignments between *D. melanogaster* (*Dm*) and *D. pseudoobscura* (*Dp*) and between *Homo sapiens* (*Hs*) and *Mus musculus* (*Mm*) and between *H. sapiens* and *Rattus norvegicus* (*Rn*). All sequences except the *Dp* sequence were obtained from Ensembl using the gene names (*ato* for *Dm* and *ATOH1* for *Hs*) and the ortholog mappings. The *Dp* contig sequence was found by BLAST on the pre-assembled *Dp* genome at the Human Genome Sequencing Center (HGSC), Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu).

We have concatenated all CNSs of a single gene into one sequence. These sequences of ∼2-4 kb were submitted to the FootPrinter web service in TOUCAN (see Section 2.7.3 and Figure 4.3) with the following parameter settings: phylogenetic tree = ((*Hs,(Mm,Rn)), (Dm,Dp)*); Motif size = 10; Maximum parsimony score = 2; Maximum number of mutations per branch = 1. This resulted

---

[4] The system of gene nomenclature used by *Drosophila* workers is largely based on mutant phenotypes.

[5] For a biological overview of *atonal* see http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly/neural/atonal.htm.

in two conserved motifs over all five species, namely TTTTATTTTG and CYTAATTARA. These were annotated in TOUCAN on the original sequences, as shown in Figure 4.8. In flies, both motifs are found in the 5' upstream region of the *atonal* CDS, while in mammals they are found in the 3' downstream region. The difference between 5' and 3' can also be found in the existing knowledge of the regulation of *atonal*, namely the two known enhancers of *ATOH1* (see Figure 2.13 and [142]) are located 3' of the gene and the known regulatory regions of *Ato* in *D. melanogaster* for its expression in R8s, antenna, Leg/Wing, and embryo are all located 5' of the gene [285]. Autoregulation is known for *atonal* so the location of one of these new motifs in a CNS that also contains an Atonal binding site (see Figure 4.8) could be an interesting finding.

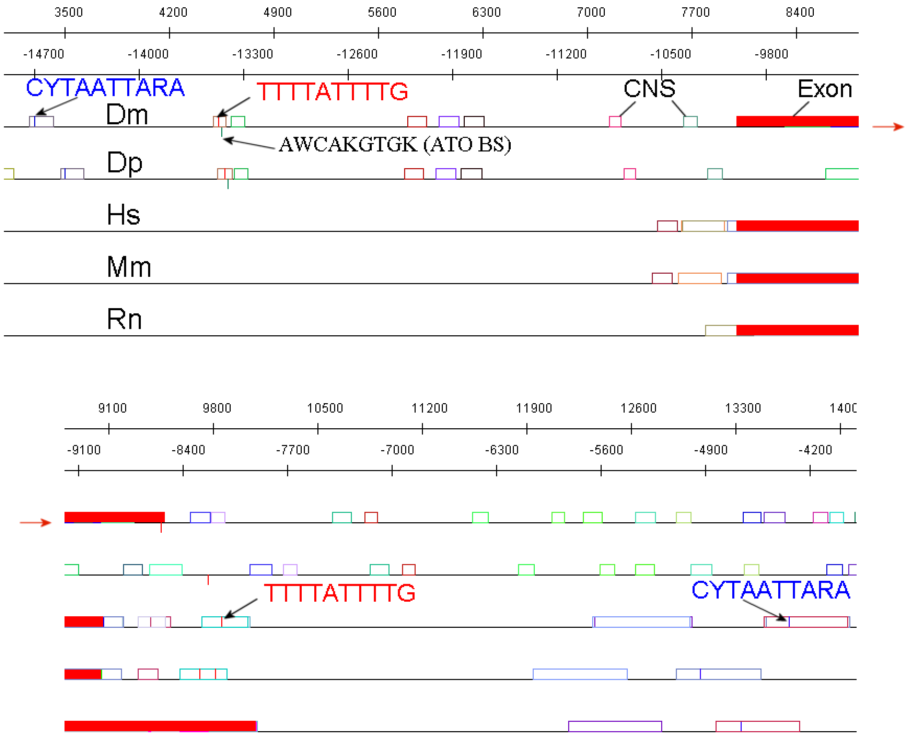


Figure 4.8: Phylogenetic footprinting. Sequence of *atonal* and of four *atonal* orthologs with upstream and downstream flanking sequences. The open boxes are CNSs, either between *Dm* and *Dp* or between *Hs* and *Mm* or *Rn*. The black arrows point at conserved motifs found with the FootPrinter algorithm [35] in a set of concatenated CNSs. The red arrows indicate that the sequences in the upper part of the figure continue in the lower part.

4.7 Related work

TOUCAN was designed with the analysis of regulation in gene sets of higher eukaryotes as its primary goal. It is in this setting that it provides the user with the most added value as compared to existing tools. TOUCAN can be considered an 'all-in-one' application, allowing for multiple approaches in regulatory sequence analysis, making it sometimes difficult to compare with other tools. Other multi-purpose tools are the web-based RSAT (Regulatory Sequence Analysis Tools [308]) and Genomatix (http://www.genomatix.de) suites. RSAT however is better suited for the analysis of prokaryotic and yeast sequences and Genomatix is a commercial package with different functionalities.

A brief list of the possible approaches for TFBS detection is given below together with some of the available tools and algorithms. They are classified according to the number of species used (one or multiple, the latter is phylogenetic footprinting), the number of genes used (one or more, the latter is a battery of co-regulated genes with shared motifs), and the technique used (sequence scoring with PSFMs or *de novo* motif discovery). In fact most of the existing tools for regulatory sequence analysis are compatible with TOUCAN as long as their output is formatted as, or can be converted to GFF that can be imported into TOUCAN. The detection of *cis*-regulatory modules or combinations of TFBSs is not included in this list, it is discussed in Section 2.4.6 and in Chapter 5.

1. *Single gene - single species - PSFM*: predict PSFM instances or occurrences of a consensus sequence, for example with RSAT or TOUCAN (IUPAC annotation or MotifScanner). The disadvantage is the high number of false positive predictions and for that reason all other tools mentioned here have been developed.

2. *Single gene - single species - motif*: not possible unless a regulatory region contains multiple binding sites of the same TF. In that case Gibbs sampling, suffix trees, etc. could be used.

3. *Single gene, phylogenetic footprinting, PSFM*: TraFaC [160], rVISTA [192] and ConSite [179] can be used but are limited to two species. TFBSs are predicted in aligned orthologous sequences and only those TFBSs are retained that are conserved and that have equivalent positions in the aligned sequences. The remainder of the sites, which are not conserved between the two species, are considered to be false positives and are eliminated. rVISTA uses the TRANSFAC database of PSFMs, ConSite uses the JASPAR database. Alternatives for multiple species are: (1) use FootPrinter and search for a corresponding TF in TRANSFAC or use MotifScanner in TOUCAN (TRANFAC, JASPAR, or both) on the same set and compare the motif and PSFM instances; or (2) use TOUCAN with overrepresentation statistics as if it were a gene battery, thereby losing the phylogenetic information.

4. *Single gene, phylogenetic footprinting, motif*: the FootPrinter algorithm [35].

5. *Gene battery, single species, PSFM*: MotifScanner and statistics in TOUCAN (restricted to proximal sequences). This approach has been applied independently in [334] using proximal promoters from EPD.

6. *Gene battery, single species, motif*: Oligo-analysis [307], MotifSampler, MEME, etc. (see 2.5.2).

7. *Gene battery, phylogenetic footprinting, PSFM*: Select CNSs in TOUCAN (proximal and distal but only for two species) followed by MotifScanner and Statistics. Alternative for multiple species: PhyloCon [314] followed by MotifScanner to compare the instances, or search for a corresponding TF in TRANSFAC for the PhyloCon motifs.

8. *Gene battery, phylogenetic footprinting, motif*: the PhyloCon algorithm.

4.8 Conclusions

Toucan provides an efficient and integrated environment for gene regulation bioinformatics. Starting only from gene identifiers, it can retrieve, visualize, annotate, and analyze proximal and distal regulatory sequences of coregulated genes. Because we use web services, we can add more services that work with fastA formatted sequence files and we will be able to link with bioinformatics service registries in the future. This flexibility will help to improve the interoperability among visualization tools, algorithms, and data providers for gene regulation bioinformatics [272]. As discussed under the liver and muscle case studies, the over-representation method has several limitations. We will try to solve some of these in the next Chapter where we describe a method to detect over-represented *combinations* of TFBSs as modules, which should cause a further decrease in the number of false positive TFBS predictions.

Chapter 5

Detecting *cis*-regulatory modules

5.1 Introduction

WORKING with combinations of factors makes it possible for the cell to integrate multiple inputs and this further provides cross-coupling of signal transduction and gene regulatory pathways. This way, a *cis*-regulatory module (CRM) functions as an information processing device (see Chapter 2). It is therefore meaningful in a biological sense to search for the co-occurrence of multiple TFBSs within a confined window of DNA sequence. In a computational sense, the detection of combinations of TFBSs has the advantage of reduced false positive predictions and of specificity. Namely, it will be feasible to search for target genes in the full genome that are putatively controlled by a given combination of TFs while this is not feasible for single TFs.

Here we present a novel approach for finding combinations of TFBSs that occur several times across multiple coregulated human genes. Again we combine coregulation with phylogenetic footprinting by focussing our search to syntenic regions with respective mouse orthologous genes. We apply a score function that combines the scores generated by the MotifScanner (i.e., log likelihood ratios) of individual PSFMs from TRANSFAC. Here, attention is paid to the sensitivity and specificity of the PSFM scoring. Obviously an efficient algorithm is needed to search the enormous set of possible combinations of binding sites[1]. The ModuleSearcher algorithm implements the score function in an A^* tree search and in a faster Genetic Algorithm (GA) version. We show the results of the ModuleSearcher obtained on four artificial data sets and explore the sensitivity and specificity of the algorithm. We justify the methodology and the different thresholds and parameters used along the road by applying the ModuleSearcher on real biological data. For the latter we have chosen a coherent cluster of

---

[1] For example, if we have 400 factors then there are $400^5/5! = 8.10^{10}$ possibilities for a CRM with 5 binding sites.

gene expression profiles, as captured by a microarray study on the cell cycle in a human cancer cell line. The modules we find are proven to contain real regulatory information. To our knowledge, this shows for the first time that module detection in microarray clusters of human genes is feasible, when taking all precautions discussed here to reduce the level of noise into account.

The score function alone is used in the ModuleScanner program to detect genes in the genome that might be controlled by a certain CRM. We have tested this program using the IFN-β enhancer as a model, and using the predicted CRM of the microarray cluster. Predicted targets are validated *in silico* using Gene Ontology annotation.

5.2 Module score function

Analogous with the distinction between a binding site and a motif model (a PSFM is a motif model and is denoted as $\mathbf{\Theta}$), we distinguish modules (or CRM), denoted as $\mathbf{m}$ and module models (CRM models) denoted as $\mathcal{M}$. Modules are clusters of sequence segments $\mathbf{x}$. In our analyses, the $\mathbf{x}$s are predicted binding sites, namely instances of PSFMs on the sequence as predicted by a scoring algorithm like the MotifScanner. Module models are thus sets of motif models. The score of a CRM model $\mathcal{M}$ on a set of sequences $\mathbf{set} = (\mathrm{seq}_1, \ldots, \mathrm{seq}_n)$ is calculated as

$$\mathcal{S}_{\mathcal{M}}(\mathbf{set}) = \sum_{i=1}^{n} \mathcal{S}_{\mathcal{M}}(\mathrm{seq}_i). \tag{5.1}$$

The score of a CRM model $\mathcal{M}$ on one sequence seq is calculated as

$$\mathcal{S}_{\mathcal{M}}(\mathrm{seq}) = \max_{\mathbf{m} \in T} p(\mathbf{m}) \times \sum_{\mathbf{x} \in \mathbf{m}} W(\mathbf{x}). \tag{5.2}$$

The different elements of this formula are the following.

- $W(\mathbf{x}) = \log\left(\frac{P(\mathbf{x}|\mathbf{\Theta})}{P(\mathbf{x}|\mathcal{B}_m)}\right) = \sum_{j=1}^{W}[\log(\boldsymbol{\theta}_j^{b_j}) - \log(P(b_j|\mathcal{B}_m))]$, as described in Chapter 4;

- $\mathbf{m}$ is a module (a collection of $\mathbf{x}$s);

- $\mathbf{x}$ is a short sequence segment that is an instance of a $\mathbf{\Theta}$ and thus a putative TFBS;

- T is the collection of all *valid* $\mathbf{m}$s, or in other words all possible instances of $\mathcal{M}$ on a seq. Whether a module $\mathbf{m}$ is valid is determined by:

  - the $\mathbf{x}$s of $\mathbf{m}$ can only be instances of one of the $\mathbf{\Theta}$s of the module model $\mathcal{M}$=[$\mathbf{\Theta}_1$, …, $\mathbf{\Theta}_{n_{\mathbf{\Theta}}}$], where $\mathcal{M}$ is a parameter of the score function. Instances of other motif models are not taken into account.

  - All $\mathbf{x}$s of $\mathbf{m}$ have to lie together within a sequence window of maximally `maxLength` base pairs. In other words, the largest distance between two $\mathbf{x}$s can maximally be `maxLength` base pairs. `maxLength` is a parameter of the score function for which the default value is 200 bp.
  - Another parameter of the score function is `overlapAllowedYesNo`. If this is set to `true`, then the $\mathbf{x}$s are allowed to overlap.
    If `overlapAllowedYesNo` is `false` then $\mathbf{m}$s with overlapping $\mathbf{x}$s are not part of T.

  T is constructed "on the fly" in the recursive score function
  `computeBestTarget(seq,`$\mathcal{M}$`,maxLength,allowOverlapYesNo,penalizeYesNo)`

- The factor $\mathrm{p}(\mathbf{m})$ functions as a penalization for CRMs that do not contain an instance of one of the $\mathbf{\Theta}$s in $\mathcal{M}$. It is the number of $\mathbf{x}$s in the module $\mathbf{m}$ divided by the number of motif models $n_{\mathbf{\Theta}}$ in $\mathcal{M}$. Penalization of incomplete modules can be enabled or disabled, as required by the user. If it is disabled, $p(\mathbf{m}) = 1$.

This score function does not take the motif order into account, nor the distance between motifs. The only distance constraint is the total length of $\mathbf{m}$ as defined by the window size `maxLength`. The simple score function presented here was satisfactory for our current goals. However, more complicated score functions based on hidden Markov models could be tested in the future, such as COMET [116].

Technically, the score function is implemented in Java and uses the $\mathbf{x}$s and $\mathrm{W}(\mathbf{x})$s from the GFF output of the MotifScanner.

5.3 The A^* search algorithm

Our search for the best CRM model on a set of sequences is handled with an A^* procedure, a branch-and-bound search[2] with a heuristic estimate of the remaining distance to the solution. A^* algorithms are *admissible* [136, 194]. This means that "it is guaranteed to find a minimal path to a solution [in our case this is a solution with the maximal score] whenever such a path exists" [194]. In bioinformatics, the A^* algorithm has already been used for multiple sequence alignment [180]. Each node in the implicit search tree is a CRM model $\mathcal{M}$. Creating child nodes involves adding $\mathbf{\Theta}$s to parent $\mathcal{M}$s. Since we do not consider the order of sites in this step, we have removed redundant nodes by allowing only alphabetically ordered CRM models. A function $\mathcal{G}_{\mathcal{M}} = \mathcal{S}_{\mathcal{M}} + \mathcal{H}_{\mathcal{M}}$

---

[2] Branch-and-bound searches do not consider all possible paths but instead eliminate unnecessary work by only extending, at each iteration, the node with the best score in the queue. The new nodes (new paths) that are generated this way are added to a queue that is sorted according the score (path length). In our algorithm, not the path length is used but the score of a node. This is possible because only alphabetically ordered modules are considered and thus there are no redundant paths that lead to the same node.

is used, where $\mathcal{S}_{\mathcal{M}}$ is the score-function, and $\mathcal{H}_{\mathcal{M}}$ is a heuristic overestimate of the rise in score from $\mathcal{M}$ to the best child $\mathcal{M}_b$. To explain how this heuristic is calculated, consider the following example. For a module $\mathcal{M}$=[$\mathbf{\Theta}_{249}$,$\mathbf{\Theta}_{34}$] ($n_{\mathbf{\Theta}}$=2) somewhere in the search tree, and for $N_{\mathbf{\Theta}}$=5 (the desired number of elements in the module, $N_{\mathbf{\Theta}}$ is a parameter of the algorithm to be specified by the user), $\mathcal{H}_{\mathcal{M}}$ is the sum of the scores of the 3 (5-2=3) best single element modules on the set. This can be calculated off-line before the algorithm starts, by summing the maximal scores (i.e., the best hits) of each $\mathbf{\Theta}_i$ on each sequence. Since these instances do not obey the distance constraints, the heuristic is always an overestimate of the rise in score of this $\mathcal{M}$.

The collection of $\mathbf{\Theta}$s that is used (TRANSFAC professional release 7.x) is redundant, so to avoid adding motif models to a module model that already contains a similar motif model (for which the instances will largely overlap), we have constructed *classes* of motif models as follows. For each pair of motifs $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$, the similarity is measured with the mutual information or Kullback-Leiber distance [170]. The mutual information is computed as

$$\frac{1}{W}\sum_{j=n}^{W}\sum_{b=A}^{T}\mathbf{\Theta}_1(j,b)\log\frac{\mathbf{\Theta}_1(j,b)}{\mathbf{\Theta}_2(j,b)}$$

where $\mathbf{\Theta}_1(j,b)$ is the probability of finding base b at position j in motif model $\mathbf{\Theta}_1$. Since this equation is asymmetric, we take the average between the distance from $\mathbf{\Theta}_1$ to $\mathbf{\Theta}_2$ and from $\mathbf{\Theta}_2$ to $\mathbf{\Theta}_1$. All motif models are considered similar if their distance is lower than, for example, 0.2. The distances are saved in a text file that is passed to the algorithm as a parameter. Thus, different thresholds can be used.

The algorithm, searching for the maximal score, is shown here:

1. Initialization
   (a) `Queue` contains the root node as only element (the empty CRM model).
   (b) `Solution` is null.
   (c) The parameter $N_{\mathbf{\Theta}}$ is set, which is the number of sites a module should contain.
   (d) The parameters of the score function are initialized.
2. While $\mathcal{G}_{\mathcal{M}}(\mathbf{set}) \geq \mathcal{S}_{\texttt{Solution}}(\mathbf{set})$, where $\mathcal{M}$ is the first CRM model in the `Queue` (or while no `Solution` is found yet), do
   (a) Remove the first $\mathcal{M}$ from `Queue`.
   (b) Consider all $\mathbf{\Theta}_i$ for which there is no $\mathbf{\Theta}$ of the same *class* (see above) already present in $\mathcal{M}$. Also consider those $\mathbf{\Theta}_i$ for which *exactly the same* $\mathbf{\Theta}$ is already present in $\mathcal{M}$, but only if the parameter `multipleCopiesAllowedYesNo==true`. For all these $\mathbf{\Theta}_i$ do:

```
        i. Create a new CRM model $\mathcal{M}_{\text{new},i} = [\mathcal{M},\mathbf{\Theta}_i]$ (add $\mathbf{\Theta}_i$ to $\mathcal{M}$).
        ii. If the size of $\mathcal{M}_{\text{new},i}$ is $N_{\mathbf{\Theta}}$, and if $\mathcal{S}_{\mathcal{M}_{\text{new},i}}(\mathbf{set}) > \mathcal{S}_{\texttt{Solution}}(\mathbf{set})$,
            then Solution $= \mathcal{M}_{\text{new},i}$.
        iii. If the size of $\mathcal{M}_{\text{new},i}$ does not equal $N_{\mathbf{\Theta}}$, add $\mathcal{M}_{\text{new},i}$ to Queue.
    (c) Sort the Queue by descending $\mathcal{G}(\mathbf{set})$
3. Solution now contains the optimal $\mathcal{M}$.
```

5.4 Validation of the A^* ModuleSearcher

5.4.1 Semi-artificial sequence sets

A 3rd-order Markov model was calculated from all human-mouse syntenic regions within the 10 kb gene-upstream sequence (i.e., the "syntenic fastA database", see further), representing the base pair composition of conserved regions (see Section 4.3.2). Artificial sequences were generated by sampling symbols from this background model. Transcription factor binding sites were implanted at random locations by sampling a TFBS from position-specific frequency matrices. To reflect a more realistic biological situation, we added artificial sequences without implanted binding sites that represent false positive sequences[3]. The first column of Table 5.1 describes the contents of the four constructed test sets. In Art_4 multiple of these artificial sequences were implanted themselves into larger sequences. Figure 5.3 shows 10 such sequences with four implanted CRMs each, separated by Ns. The blanks between the modules illustrate the fact that we will consider only the syntenic regions, not other intergenic DNA.

Table 5.1 lists the results obtained on semi-artificial data. Analysis of Art_1 shows that the ModuleSearcher is able to detect a module of 5 elements correctly (all 5 elements are found) when it is hidden in 10 sequences of 200 bp and when another 10 random sequences of the same length are added. The results on the Art_2 set show that the ModuleSearcher can detect 2 distinct modules that are hidden in a set of 15 sequences, although some elements were misidentified: 4 out of 5 elements of Module 1 are correct, and 2 out of 3 elements of Module 2 are correct. Figure 5.1 shows Art_2 when scored with the MotifScanner. It can be seen from this figure that many implanted sites are missed in the scoring step, which causes an important limitation on the sensitivity of module detection. In Figure 5.2 the implanted sites are compared with the output of the ModuleSearcher (i.e., the best hit of the found module on each sequence).

We search for a combination of factors that is over-represented in a set; therefore a distinction can be made between treating all syntenic regions of one gene independently (in that case, a set contains all regions of all genes separately, like for the liver and muscle sets in Section 4.6.2) and keeping all regions of a gene together (the set contains all genes, each having one or more regions). The effect is that in the case of combining the syntenic regions of a single gene,

---

[3] A real set of sequences that *all* contain the same CRM can probably never be found and sequence sets could consist of multiple gene batteries each containing another CRM.

a good module will have to be found in at least one of the syntenic regions of most genes, while in the case of treating the regions independently, a good module will have to be found in most of the syntenic regions. To investigate this effect, and more importantly to decide whether to keep the regions in a real biological data set together, we tested both possibilities on semi-artificial data as well. Comparing Art_3 (where all regions are added independently to a set) and Art_4 (where multiple syntenic regions of one gene are kept together, see Figure 5.3) shows that the second approach is advisable, so this will be applied further on the co-expressed gene set.

5.4.2 Sensitivity to PSFM scoring

Because the ModuleSearcher algorithm uses the scores of individual matrix hits, we have compared the effectiveness of the algorithm using different types of scoring. The Art_1 set was scored with the MotifScanner using different values for the prior parameter. When 0.1 or 0.2 were used, the ModuleSearcher found 5 out of 5 correct CRM elements. Using 0.5 as a prior, it found 4 out of 5 elements. The same set was also scored with the MotifLocator, with varying threshold values. The MotifLocator can be compared with other programs that score frequency matrices such as Matinspector [241]. Setting the threshold to 0.75 resulted in 4 out of 5 correct elements, but this threshold yields 12 times as many hits as for the MotifScanner with prior 0.2. A threshold of 0.8 resulted in 3 out of 5 correct elements; 0.85 in 1 out of 5 and 0.9 in 0 out of 5. Taken together, the MotifScanner (with its probabilistic estimation of the number of hits) confers robustness to the ModuleSearcher and will be used in the Syntenic GFF database and in the study of co-expressed genes.

5.5 ModuleSearcher on real gene batteries

Figure 5.4 shows a flow chart that overviews the system for detecting regulatory modules. All human-mouse orthologous pairs were selected from Ensembl Release 9 (19,914 pairs). Ten kilobases of sequence upstream of the coding sequence of the human and mouse gene were selected (18,778 pairs with successful selection). Each 10kb pair was aligned with AVID [42] and the alignment output was parsed using VISTA [209] to select regions with at least 75% identity in windows of 100 bp (10,049 pairs had at least one region; 33,282 regions in total). These regions form the “Syntenic fastA” database. All syntenic regions were scanned to predict transcription factor binding sites (TFBSs) using the MotifScanner algorithm (prior parameter set to 0.2). Frequency matrices were taken from TRANSFAC Professional Release 6.3, which contained 429 vertebrate matrices. All occurrences are stored in GFF format in the “Syntenic GFF” database that is both used for the selection of annotated regions of coregulated genes (to find CRMs) and for “genomic searches” to find genes containing a given CRM. In the current version we have limited the intergenic sequence space to 10kb upstream of the coding sequence, but extensions towards syntenic regions located

Table 5.1: Results of the ModuleSearcher on four artificial sequence sets ([a] Motif belongs to the same class as the implanted motif; [b] Motif that was not implanted).

Data set	Highest scoring module
Art_1: 10 random sequences of 200bp, each with following implants: M00134-V$HNF4_01 M00131-V$HNF3B_01 M00190-V$CEBP_Q2 M00174-V$AP1_Q6 M00206-V$HNF1_C + 10 random sequences of 200bp without implants (i.e., noise)	M00134-V$HNF4_01 M00131-V$HNF3B_01 M00190-V$CEBP_Q2 [a]M00188-V$AP1_Q6 M00206-V$HNF1_C The found module contains all 5 hidden elements.
Art_2: 5 random sequences of 200bp, each with following implants: M00134-V$HNF4_01 M00131-V$HNF3B_01 M00190-V$CEBP_Q2 M00174-V$AP1_Q6 M00206-V$HNF1_C + 5 random sequences of 200bp, each with following implants: M00054-V$NFKAPPAB_01 M00747-V$IRF1_Q6 M00750-V$HMGIY_Q6 + 5 random sequences of 200bp without implants (i.e., noise). See Figure 5.1	First run: M00134-V$HNF4_01 M00131-V$HNF3B_01 M00190-V$CEBP_Q2 [a]M00188-V$AP1_Q6 [b]M00328-V$PAX8_B The first module was found with 4 out of 5 elements correct. Second run: [a]M00052-V$NFKAPPAB65_01 M00750-V$HMGIY_Q6 [b]M00158-V$COUP_01 The second module was found after masking the elements of the first module; 2 out of 3 elements of the second module are correct.
Art_3: 5 random sequences of 200bp, each with following implants: M00134-V$HNF4_01 M00131-V$HNF3B_01 M00190-V$CEBP_Q2 M00174-V$AP1_Q6 M00206-V$HNF1_C + 35 random sequences of 200bp without implants (i.e., noise)	[b]M00446-V$SPZ1_01 [b]M00285-V$TCF11_01 [b]M00748-V$STAT5B_Q6 [b]M00137-V$OCT1_03 [b]M00734-V$CIZ_01 The hidden module is not found when it is present in only 5 out of 40 sequences.
Art_4: 5 genes with 1 module as in Art_1 and 3 empty regions, well separated + 5 genes with 4 empty regions. The empty stretches between the regions are not scored with TRANSFAC. See Figure 5.3	M00134-V$HNF4_01 M00131-V$HNF3B_01 M00190-V$CEBP_Q2 [a]M00188-V$AP1_Q6 M00206-V$HNF1_C When different regions of the same gene are grouped together, the level of noise is reduced and the module can be found, with 5 out of 5 elements correct.

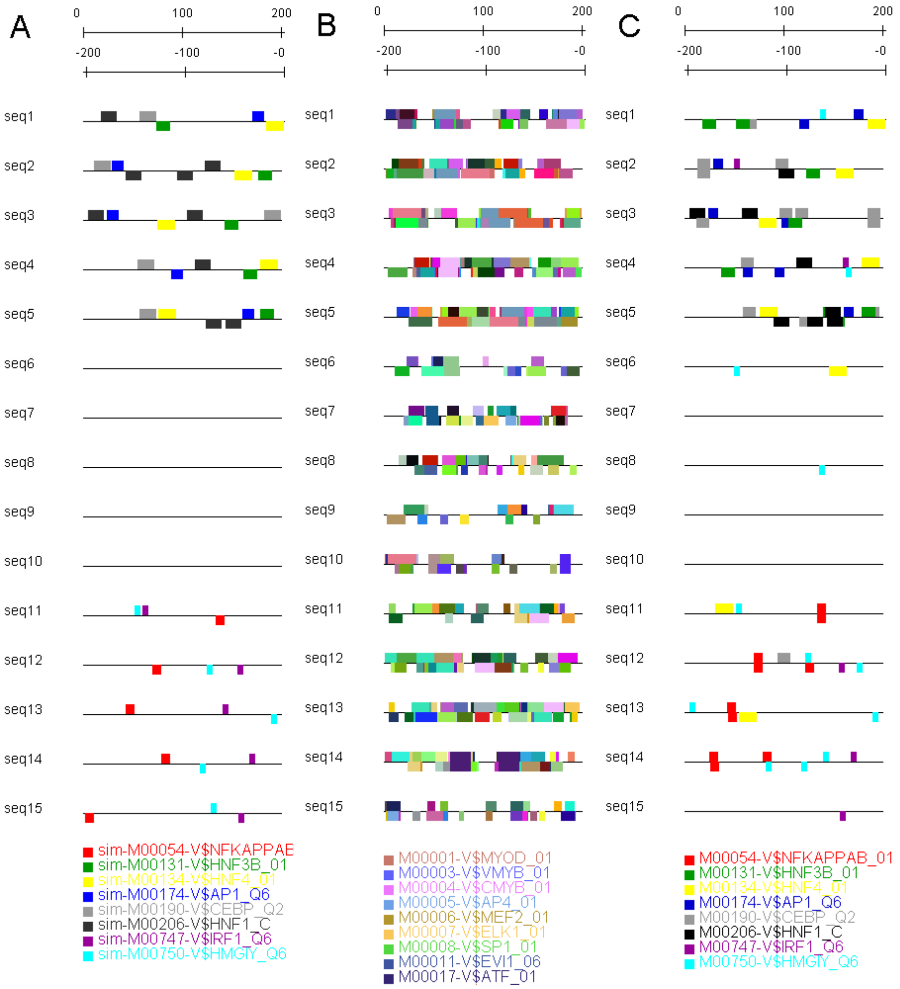


Figure 5.1: Module detection in artificial data sets. (A) Set Art_2 as described in Table 5.1, showing only the implanted binding sites, sampled from the respective matrices from TRANSFAC. (B) The same set, scored with the MotifScanner using all available matrices. This is the actual data in which the ModuleSearcher will search for modules. (C) The same as in (B), but now only displaying the instances of the matrices that were implanted. It is clear that there are many false positives and many true negatives, a fact that obviously hinders module detection.

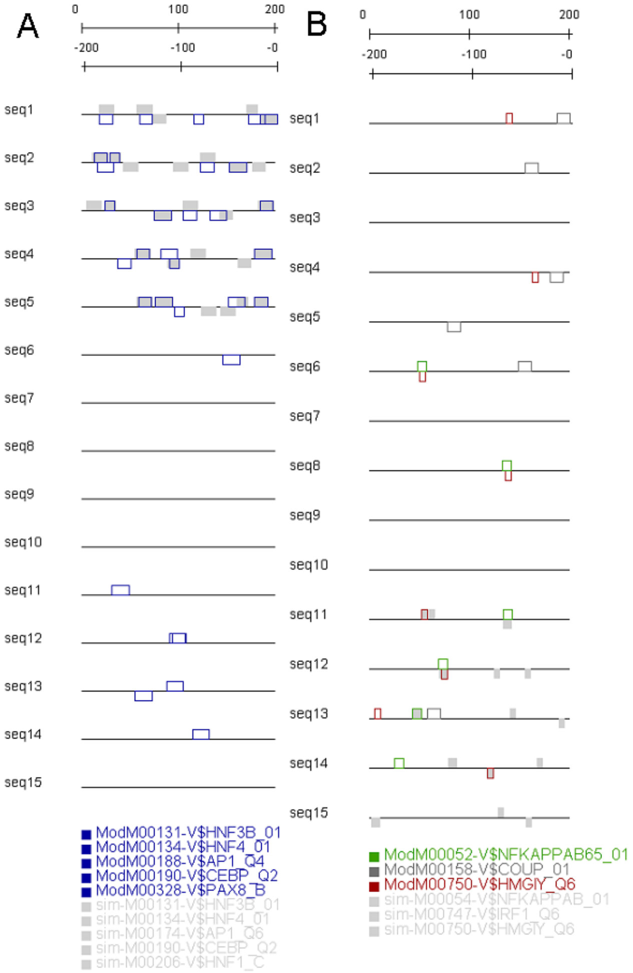


Figure 5.2: Results of the ModuleSearcher on the Art_2 set presented in Figure 5.1 (A) In blue are the results of a first run of the ModuleSearcher and in grey the implanted sites as in Figure 5.1.A. (B) In red and green are two of the three hidden matrices, as detected in a second run on the same set (masking the results of the first run) of the ModuleSearcher.

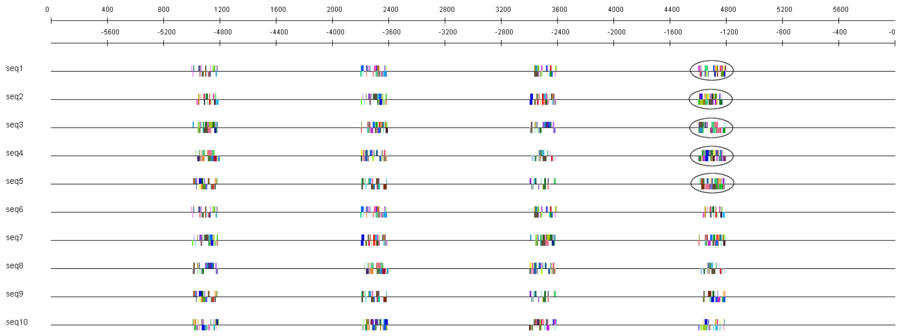


Figure 5.3: Set Art_4 as described in Table 5.1, resembling the biological situation where multiple syntenic regions of one gene belong together. Only the encircled regions have implanted modules (5 out of 40 regions), and these can still be detected.

in introns or downstream of the gene are possible.

5.5.1 Gene Ontology statistics

We have developed a software tool, GO4G, to calculate the functional coherence of a gene set based on Gene Ontology associations. GO4G will be used further below to validate newly found modules by the functional coherence of putative target genes of this module. It is available at http://www.esat.kuleuven.ac.be/~saerts/software/go4g.html and works as follows. All annotated GO terms for a set of genes are retrieved from the GOA annotations of the EBI (http://www.ebi.ac.uk/GOA/). For each term, each path to the root of the GO tree is followed and each encountered term is added to a gene's annotation. For each term, the frequency of this term is then the number of genes that have the term in their extended annotation divided by the total number of genes in the gene set. The binomial formula (see Section 4.4) is then used to calculate p-values for each frequency, where the expected frequencies are calculated from a large reference set, such as the complete human genome. For the analysis described here we have used the set of human genes that have a mouse ortholog. The p-values are then corrected for multiple testing. GO4G can be used for testing the functional coherence of a gene set and is therefore useful for validating predicted target genes.

5.5.2 Genomic searches

Using the ModuleScanner we can score the complete "Syntenic GFF" database to find syntenic regions that potentially contain a CRM. To determine the specificity of target detection, we have compared the scores of the sequences in the Art_1 set (using the best CRM found with the ModuleSearcher in this set) with the scores of the same (artificial) CRM on the database. There are 6 regions

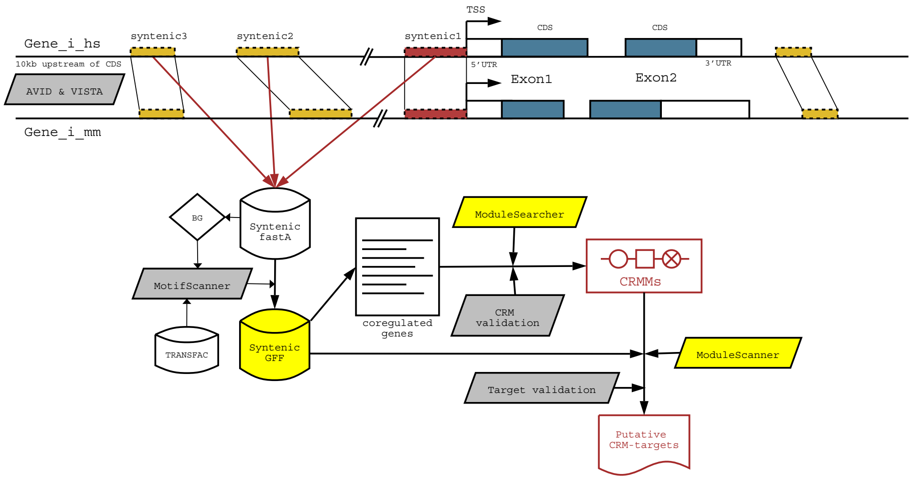


Figure 5.4: Overview of the system to detect regulatory modules. *All* syntenic DNA regions, ranging from 100 to several hundreds of base pairs, resulting from global alignment of *all* human-mouse ortholog pairs are extracted from the genomic sequence and stored in a "Syntenic FastA database". A 3rd order background model is calculated from these sequences. The sequences are scored with all vertebrate matrices of TRANSFAC, and contrasted with this background model in the MotifScanner algorithm. The output of the MotifScanner, in GFF format, is stored in a "Syntenic GFF database". If now, one wishes to find a module in a set of co-regulated genes, the relevant features (i.e., MotifScanner hits) are extracted from the Syntenic GFF database and used as input for the ModuleSearcher. A first "CRM validation" can be the visual inspection of the modules in TOUCAN, although the structure of a module (the ordering and spacing of the elements) is often not conserved throughout a sequence set. A better validation of a newly found module is therefore to scan all syntenic regions in the Syntenic GFF database with this module, using the ModuleScanner algorithm. The list of putative target genes with the highest scores (e.g., the top 20 genes) can be inspected either manually in the literature or automatically with GO4G to check their functional coherence. All of these steps and the used algorithms are explained further in the text.

(out of the 10 regions where we implanted it) that have a higher score than all the regions in the database.

A second test was carried out, this time using a known *cis*-regulatory module, namely the IFN-β enhancer [217]. This module contains, within less than 100 base pairs, functional binding sites for NF-$\kappa\beta$, ATF2/JUN, IRF, and HMGI(Y) (four copies and one overlaps with the NF$\kappa\beta$ site, see Figure 2.7). The TRANSFAC database only contains matrices for HMGI(Y), NF$\kappa\beta$, and IRF-1 so we used these three to specify a module model. The ModuleScanner scored the GFF database with this model, and the top 10 scoring genes were fed into the GO4G program. Table 5.3 shows the significantly over-represented GO terms within these 10 genes, and it can be seen that they are related to the response of a cell to viral infection, the process where the IFN-β enhancer is active. The IFN-β gene itself was found as fourth best scoring gene. Other high scoring genes include: EH-domain containing protein 1 (testilin, *EHD1*) involved in the recycling of major histocompatibility complex class I molecules to the plasma membrane; IL-1 β precursor (catabolin, *IL1B*), an important mediator of the inflammatory response; NF-$\kappa\beta$ inhibitor alpha (*NFKBIA*), involved in apoptosis and possibly pointing at feedback control mechanisms; and semaphorin 3B precursor (*SEMA3B*), involved in cell-cell signaling and possibly coregulated with IFN-β to mediate contacts between dendritic cells and T lymphocytes. By combining transcription factors in modules, the specificity increases to a level where genomic searches become feasible. This result opens the door to the validation of predicted modules, as illustrated in the next paragraph, because a genomic search with a false module will retrieve random top scoring genes that have an extremely low chance of statistical significant functional coherence.

5.5.3 Biological validation of the ModuleSearcher

Sets of co-expressed genes were selected using SOURCE [86]. A typical case of coregulation is the cell cycle and we have queried the SOURCE database for cyclin B2 (CCNB2). In the "expression view" we have chosen the data set of gene expression during the cell cycle in a human cancer cell line (HeLa) [322]. By searching for genes that have a similar profile, using the functionality provided by the application, we selected 44 genes that might share a common *cis*-regulatory element. Of these, 34 had an Ensembl identifier, and in this set we found 13 genes with at least one syntenic region with the respective mouse orthologous gene (32 regions in total).

The selected gene cluster around cyclin B2 is functionally tight: among the highly significantly over-represented Gene Ontology terms are cell cycle (15 genes, p-value = 10^{-14}), M phase (9 genes, p-value = 3.10^{-13}), and microtubule cytoskeleton (9 genes, p-value = 2.10^{-7}). The best module model in the cluster, as selected by the ModuleSearcher (window=100bp and $n_{\boldsymbol{\Theta}}$=4) consisted of NFY, STAF, TCF4, and CEBPA.

Table 5.2: Results of the ModuleSearcher on a set of co-regulated cell cycle genes.

Data set	Highest scoring module
CCNB2_clus: Set of 13 human genes co-expressed with cyclin B2 during the cell cycle in HeLa cells; selected from SOURCE. In total they have 32 conserved sequence blocks within 10kb upstream of the CDS. The blocks of a gene are grouped together as in Art_4.	M00116-V$CEBPA_01 M00264-V$STAF_02 M00287-V$NFY_01 M00671-V$TCF4_Q5 This result was validated by finding target genes of the module using the ModuleScanner, see text.

It has been shown that NFY (nuclear factor Y) regulates genes (e.g., cyclinB1) in a cell type specific and cell-cycle dependent fashion [162]. TCF4 regulates cyclin D1 expression in a complex with β-catenin [294], so its involvement in cell-cycle specific expression of other genes is plausible. CEBPA (CCAAT/enhancer binding protein alpha) overlaps with some of the NFY sites (see Figure 5.5), which could explain its presence in the module. The fourth element, STAF, is a zinc finger protein that is a promiscuous activator for enhanced transcription by RNA polymerases II and III [253].

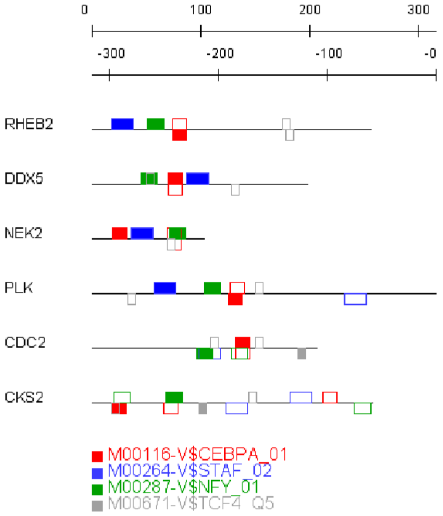


Figure 5.5: Biological validation. Six of the 20 highest scoring syntenic regions with the CEBPA-STAF-NFY-TCF4 model that was found in the cyclin B2 microarray cluster. The closed boxes are the sites of the module and the open boxes are putative sites of the same factors scored with a lower threshold. Taking the open and closed boxes together, each region has at least one instance of each module factor.

Using the [STAF–CEBPA–NFY–TCF4] module in a genomic search with the ModuleScanner shows indeed that this combination contains cell-cycle specific regulatory information, because (1) 30.8% (4 out of 13) of the original cluster is found in the top 100 scoring genes, and (2) the GO4G statistics on the top 20 scoring genes show a significance (corrected p-value smaller than 0.05) for terms like “mitosis”, “regulation of cell cycle”, and “cell proliferation” (see Table 5.3). Figure 5.5 shows the actual modules in some of the top 20 scoring cell cycle genes. Polo-like kinase (PLK) is possibly active in chromosomal segregation, NEK2 is involved in chromosome segregation and centrosome separation. CDC2 (cell division cycle 2) is a catalytic subunit of the highly conserved protein kinase complex known as M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of eukaryotic cell cycle. CKS1B is also known as CDC2 associated protein so its coregulation with CDC2 is plausible.

Table 5.3: Functional coherence. Validating putative target genes found by the ModuleScanner using GO4G. Terms with at least 2 occurrences and corrected p-value smaller than 0.05 are shown. When both parent and child terms were significant, only the child is shown.

Genes	Significant GO terms	Corrected p-value
Top 10 scoring genes for a simplified IFN-β enhancer: $\left\{ \begin{array}{l} \text{M00750-V\$HMGIY\_Q6} \\ \text{M00054-V\$NFKAPPAB\_0} \\ \text{M00747-V\$IRF1\_Q6} \end{array} \right.$	apoptosis	0.002046759
	negative regulation of cell proliferation	0.003660635
	protein amino acid dephosphorylation	0.004239594
	response to pest/pathogen/parasite	0.005201112
	protein phosphatase activity	0.006733237
	innate immune response	0.010171468
	cytokine activity	0.012253669
	response to stress	0.014523294
	phosphoric monoester hydrolase activity	0.015017083
	cell communication	0.031928904
Top 20 scoring genes for a new module found with the ModuleSearcher on a set of cyclin B2 co-expressed genes: $\left\{ \begin{array}{l} \text{M00116-V\$CEBPA\_01} \\ \text{M00264-V\$STAF\_02} \\ \text{M00287-V\$NFY\_01} \\ \text{M00671-V\$TCF4\_Q5} \end{array} \right.$	mitosis	0.000435808
	M phase of mitotic cell cycle	0.000452022
	cytokinesis	0.000468573
	nuclear division	0.001257531
	regulation of cell cycle	0.001395887
	protein serine/threonine kinase activity	0.010734361
	obsolete	0.024402181
	cell proliferation	0.026461498

5.6 Genetic Algorithm version of the Module-Searcher

Although the A^* method guarantees to find the optimal solution, it can be slow for certain parameter settings, for large sequence sets, or for modules that contain many different transcription factors (e.g., more than five). Therefore we have implemented another search algorithm based on Genetic Algorithms (GA) [6], which is faster and more practical. The algorithm, which is summarized in Figure 5.6.A starts with the creation of p random modules. A module is a vector that contains n_Θ position specific probability matrices derived from TRANSFAC [323] or from other matrix collections that are available on our server. The list of modules is sorted according to the score function (see Section 5.2), and the s highest scoring modules are retained for the reproduction step. In the reproduction step the population grows back to size p by successive pairing and mutating of randomly selected modules (see Figure 5.6.B). When two modules are paired, for each position in the vector one element is chosen from either of the two parents, unless this element or a similar element is already present in the child module. Each element of a child module can then be mutated according to a mutation probability ρ. After g generations the "fittest" module is selected as solution.

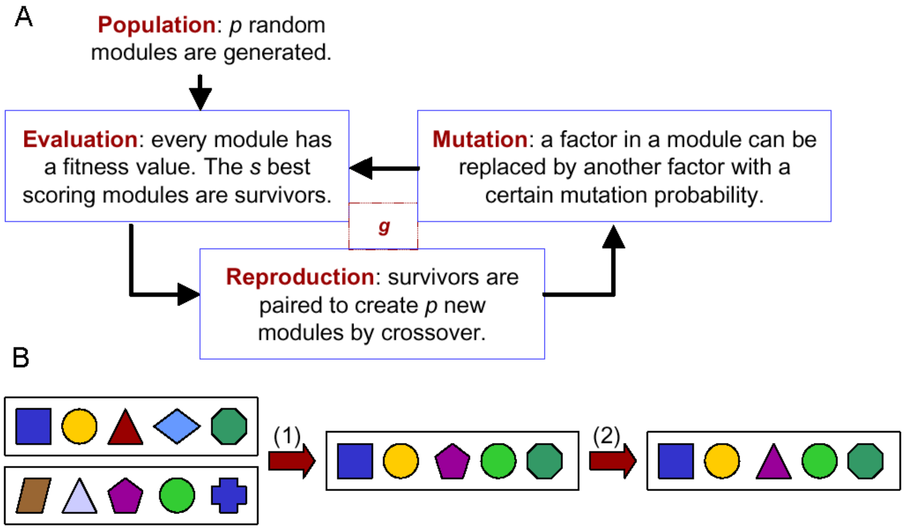

A
Population: p random modules are generated.
Evaluation: every module has a fitness value. The s best scoring modules are survivors.
Mutation: a factor in a module can be replaced by another factor with a certain mutation probability.
g
Reproduction: survivors are paired to create p new modules by crossover.
B
(1)
(2)

Figure 5.6: Genetic Algorithms version of the ModuleSearcher. (A) Procedure of the genetic algorithm; g is the number of generations. (B) Example of the generation of child modules by pairing (1) and mutations (2). Each geometrical figure represents a transcription factor.

For the technical and biological validation of the algorithm we refer to the validation of the A* algorithm. Since the GA does not guarantee optimality the user can perform multiple runs of the GA and select only those modules that are consistently found among different runs. To compare GA with A* in terms of accuracy (i.e., does GA also find the optimal solution that A* finds?) and of speed, we have run the GA and the A* version on the same set of sequences as in Section 5.5.3. The CPU time (on a 1 GHz Pentium III processor running Red Hat Linux) taken by GA, setting L to 100 bp and g to 100 iterations, is about 7, 10, 13, and 18 minutes when $n_{\mathbf{\Theta}}$ is set to 4,5,6, and 7 respectively. The time required for A* increases more dramatically with $n_{\mathbf{\Theta}}$. For $n_{\mathbf{\Theta}} = 4$, A* takes about 30 minutes, and for $n_{\mathbf{\Theta}} = 5$ it takes between five hours and three days depending on the data set and on L. $n_{\mathbf{\Theta}} > 5$ was not feasible for this particular data set, neither in time, nor in memory.

The maximum scores of three GA-runs with 100 iterations is, for $n_{\mathbf{\Theta}}$=3,4,5 exactly the same (and thus the optimal module is found) as in A*. Although we have no results of A* for $n_{\mathbf{\Theta}} > 5$, the results of GA for larger $n_{\mathbf{\Theta}}$s show the same scores in multiple runs of GA (e.g., in two out of three runs), and therefore these can be assumed to be the optimal scores. In conclusion, the GA version of the ModuleSearcher is able to find the optimal combination of binding sites without a limitation of the number of sites, and within a fraction of the time that A* needs.

5.7 Availability within Toucan

The ModuleSearcher is included in Toucan as a web service (see Figure 4.4). The GFF formatted TFBS instances and scores are sent to the service and the best instances of the optimal module are returned in GFF format and are annotated on the active sequence set.

5.8 Discussion

We have first tested the module detection algorithms on artificial data and showed that we could find back the hidden modules with a high sensitivity (i.e., after adding multiple sequences without the module), even if many of the implanted sites are missed by the matrix scoring step. The influence of the latter on the robustness of module finding was also tested and it was shown that our probabilistic estimation of the number of hits is more reliable than traditional log-odds scoring. Another test showed that the signal to noise ratio is much higher when the syntenic regions of a gene are kept together instead of separating them.

Our current program always finds a "best" module model in a set of sequences. Therefore, it is necessary to validate the module. Some possibilities are (1) the ability to retrieve target genes in the genome, (2) functional coherence of predicted target genes, (3) structure conservation of the modules in the training set and in the top-scoring database modules, and (4) phylogenetic footprinting. Structure conservation can imply conserved strand preferences or distances between binding sites. Here we have only used (1) and (2). We tested these approaches using the known IFN-β enhancer model and the results show that real module models are specific enough to find back their instances in the full genome. Lastly, we predicted a module in a set of co-expressed genes and validated the prediction using the same approach. It was shown that module detection can yield valuable hypotheses and these can ultimately help in cracking the complex gene regulatory code.

How exactly the top scoring genes are related to the modules remains to be investigated. We believe however that using the described approaches, the *in silico* generated hypotheses regarding *cis*-regulation should have a higher success rate compared to approaches based on single factors or that do not take cross-species sequence conservation into account.

Related work

Module searching

To our knowledge there is only one algorithm to detect over-represented combinations of TFBSs in sets of co-regulated genes, namely CREME [262], developed independently and published at the same moment as the ModuleSearcher. CREME does not work with all PSFMs but only with those that are individually over-represented (p-value<0.01) and then filters similar PSFMs using a greedy algorithm. Possible combinations of PSFMs are generated with a hashing algorithm and tested on the sequence set. However, not all combinations are tested but those with *consecutive* instances. That is, clusters that contain at least one instance of the combination that is tested and no instances of other PSFMs. Thus, the search space is first reduced drastically and then searched exhaustively. A statistical significance of a cluster is also calculated based on its count in a gene set as compared to a background set using a hypergeometric distribution. Similar clusters are filtered afterwards and module validation is done by measuring the expression profile coherence [232] of a gene set with pairwise similarities.

Module scanning

Several methods have been published recently that take the individual matrices of a module as input and that return putative modules in the genome with a certain statistical significance: COMET [116], MSCAN [161], Stubb [265], and MCAST [18].

Functional coherence of a set of target genes

We believe that a newly found module should be validated *in silico* by screening the full genome of the species that was used. A module that was found in the "training set" by using the ModuleSearcher (either the A* or the GA version) can be retained for experimental validation in case (1) multiple top-scoring genes found in the genome scan overlap with the genes of the training set; and (2) the top-scoring genes are functionally coherent and related to the function of the genes in the training set. Besides our own GO4G, the latter can also be investigated by other tools that similarly compare the over-represented Gene Ontology annotations of both gene sets, like FatiGO (http://fatigo.bioinfo.cnio.es/), GOMiner [331], EASE (http://david.niaid.nih.gov/david/ease.htm). A list of software tools for GO analysis in general can be found at http://www.geneontology.org/GO.tools.html.

Chapter 6

Data integration for module and target validation

6.1 Introduction

GENOME-WIDE searches for *cis*-regulatory modules using combinations of position weight matrices yielded lists of putative target genes in the previous chapter. *In silico* validation of these genes was needed to determine whether the newly found module is a real module, and we proposed the functional coherence of the target genes (based on GO) as a possible validation procedure. In this chapter we will take this validation one step further and propose a framework that allows for the simultaneous validation of a putative module and for the prioritization of the putative targets (or the removal of false positive targets from the list). The latter seems useful because there were only small differences between the scores of the module instances and because there can be false positive module instances. In a molecular biology environment this could be useful before the wet-lab validation of the targets. Instead of using only Gene Ontology annotations, we have developed a data integration strategy based on multiple information sources, namely (1) microarray expression data, (2) EST expression data, (3) Gene Ontology annotation, (4) InterPro protein domain data, (5) KEGG pathway membership, and (6) textual data from LocusLink and Medline [3]. The statistical prioritization is based on order statistics.

The idea is to build a model $\mathcal{M}$ for a set of n training genes denoted as $\mathcal{A}$ (this can be the gene battery or a set of known target genes) and to score a set of m test genes denoted $\mathcal{B}$ (these are the putative target genes as found by genome-wide module detection) with $\mathcal{M}$. The genes of $\mathcal{B}$ are then ranked according to how good they match the training genes, or more generally how good they belong to the biological process that is represented by the training genes.

All genes are represented internally by their Ensembl stable_gene_id [31] and each data source is mapped to this identifier.

6.2 Data sources

All data that can contribute to the characterization of a gene can ultimately be included in the framework. Each data source is used to build an “information submodel” (ISM), the test genes are scored with each ISM separately, and afterwards all ISM scores are combined. Depending on the format of the data, the training and representation of the ISM and the scoring of test genes with the ISM is different. We distinguish two categories of data types: vector data and non-vector data.

6.2.1 Vector data

In the vector model, the data of gene g are represented as an attribute vector $\mathbf{V}_g = [w_1, ..., w_N]$ where each w_k can be any normalized score or weight. The ISM for vector data is simply the average vector over all training genes in $\mathcal{A}$: $\mathbf{V}_\mathcal{A} = [\bar{w}_1, ..., \bar{w}_N]$. A test gene h is scored with an ISM by calculating the cosine similarity between $\mathbf{V}_h$ and $\mathbf{V}_\mathcal{A}$:

$$r(\mathbf{V}_h, \mathbf{V}_\mathcal{A}) = \frac{\mathbf{V}_h \cdot \mathbf{V}_\mathcal{A}}{||\mathbf{V}_h||||\mathbf{V}_\mathcal{A}||} = \frac{\sum_{k=1}^{N} w_{k,\mathbf{V}_\mathcal{A}} \cdot w_{k,\mathbf{V}_h}}{\sqrt{\sum_{k=1}^{N} w_{k,\mathbf{V}_\mathcal{A}}^2} \cdot \sqrt{\sum_{k=1}^{N} w_{k,\mathbf{V}_h}^2}}. \tag{6.1}$$

We used the cosine similarity measure as it is the one primarily used in most Information Retrieval (IR) systems.

Microarray data

In case a module was found in a set of co-regulated genes obtained by clustering microarray data, the same microarray data set (or data sets originating from similar conditions) can be used to train an ISM. The only requirement is that the clones (or oligos) are mapped to Ensembl identifiers, for which the MatchMiner tool [55] or the Ensembl database itself can be used. The Su data set [283] is available by default. It contains expression measurements from 101 different samples taken from 47 different human tissues and cell lines under normal physiological state. Large repositories of microarray data like ArrayExpress or GEO can be queried manually for suitable microarray data to be imported into the framework.

The effect of scoring with a microarray ISM is that genes with similar expression profiles as the average expression profile of the training genes will get high scores (a small angle between the vectors results in a cosine value close to 1).

Textual data

The bioinformatics group at the department of Electrical Engineering ESAT (BioI@SCD) has a record in text mining research [10, 120, 121, 122]. We have integrated our system with the TxtGate framework [122]. TxtGate contains a text vector $\mathbf{V}_g$ for each gene that is present in the LocusLink database. Each element $w_{k,g}$ in $\mathbf{V}_g$ is a weight for term t from the vocabulary of size N. This representation is often referred to as *bag-of-words*. Briefly, the indexing of all textual information is done in the following steps:

1. A vocabulary is built from textual information contained in the Gene Ontology database, namely by extracting all GO terms and the words in the descriptions of the terms.

2. For each gene in a downloaded textual version of LocusLink, the *GeneRIF*[1], *summary*, and *sum_func* fields are selected and the words forming its text are extracted. Further, all the PubMed identifiers in the LocusLink record are used to retrieve the corresponding MEDLINE abstracts of which the words are also extracted.

3. Stemming [110] is applied to reduce the words to their stems by using suffix stripping. These stems are called terms. After stemming, all words that have the same stem are supposed to be reduced to the same item and thus treated as the same. In this manner, the discrimination among the documents is increased. Weights of the terms are calculated based on the frequencies of the terms in the document and the number of documents that contain that particular term over the whole document collection. The weight of a term is calculated by its TF-IDF (Term Frequency Inverse Document Frequency) value:

   $$w_{k,g} = f_{k,g} \times log_2(N_g/f_k) \tag{6.2}$$

   where $w_{k,g}$ is the weight of term k in the concatenated document of a certain gene g, $f_{k,g}$ is the frequency of term k in the text of g (i.e., term frequency), f_k is the number of genes for which the documents contain term t (i.e., document frequency), and N_g is the total number of genes in the collection.

4. Vectors are normalized by their lengths to cope with documents of differing lengths. This is accomplished by computing the length of the vector representing the document and dividing the weights of the terms by this value. The length of vector V_g is

   $$\frac{\mathbf{V}_g}{\sqrt{\sum_k w_{k,\mathbf{V}_g}^2}}. \tag{6.3}$$

$\mathcal{B}$ is scored according to the cosine similarity with the average text profile of $\mathcal{A}$ as mentioned above.

---

[1] GeneRIFs ("Gene Reference into Function") are systematically assigned statements about gene function as described by a publication in MEDLINE [212].

6.2.2 Non-vector data

Gene annotation data that cannot be represented by a vector, or where a vector representation causes the loss of information, is treated as follows.

1. If the annotation is represented as a tree (e.g., EST expression data) or a diacyclic graph (DAG) (e.g., GO), the initial annotation of a gene (i.e., certain leaves or nodes in the tree or DAG) is extended with all nodes on all possible paths to the root. If the annotation is represented as a simple attribute, it is left unchanged.
2. The frequency of the all resulting annotations in the training set is compared with the expected frequency in the genome using the binomial formula (see Section 4.4) and the resulting p-values are corrected for multiple testing. Thus, each annotated attribute is correlated with a p-value.

The p-values in $\mathcal{A}$ are used for the scoring of the $\mathcal{B}$ genes. For each annotated term or parent term of test gene h, the corresponding p-value in $\mathcal{A}$ is used. The p-values of all terms of h are combined with Fisher's Chi-square method:

$$\chi^2_{df=2k} = -2\sum \log(p_i), \tag{6.4}$$

where k is the number of combined p-values. This way, the score of h is actually a new p-value describing the probability of being similar to $\mathcal{A}$.

Data sources that are suitable for this approach are (1) Gene Ontology annotation from Ensembl, (2) EST expression data from Ensembl that represent the anatomical sites in the body where a gene is already found to be expressed, (3) pathway membership data from the KEGG database [287], and (4) InterPro protein domains from the Ensembl database.

6.3 Order statistics and overall ranking

For each gene in $\mathcal{B}$ the scoring with $\mathcal{M}$ gives six different rankings $r_1, r_2, ..., r_6$, one for each of the six data sources that are currently considered. The ranks are divided by the total number of ranked genes to obtain rank ratio's. The joint cumulative distribution of a n-dimensional order statistic is then used to compute a p-value as was also done by Stuart and colleagues [282] (see http://www.math.uah.edu/statold/sample/sample7.html for a description):

$$P(r_1, r_2..., r_n) = n! \int_0^{r_1} \int_0^{r_2} \dots \int_0^{r_n} ds_1 ds_2 ... ds_n \tag{6.5}$$

This can be computed efficiently with the recursive formula:

$$P(r_1, r_2..., r_n) = \sum_{i=1}^{n} (r_{n-i+1} - r_{n-i}) P(r_1, r_2, ..., r_{n-i}, r_{r-i+2}, ..., r_n), \tag{6.6}$$

where r_0=0. Since we included six information sources, we used $n = 6$. This p-value represents the probability of getting the observed rank ratios by chance. These values are corrected for multiple testing and $\mathcal{B}$ is then sorted according to these values. Next to the original six rankings, we now have a seventh which is the combined ranking.

6.4 ENDEAVOUR

The complete software framework that implements the methodology is called ENDEAVOUR and can be used freely by academic investigators. ENDEAVOUR consists of a platform independent Java client that can be started from a web browser using Java Web Start, several SOAP web services and a MySQL database running at our department, and an interface for direct communication with the Ensembl database. Training genes and test genes are loaded into the client by their Ensembl, HUGO, or LocusLink IDs and all other IDs, descriptions, and so on. are loaded automatically from the Ensembl database. Model training is as easy as selecting the desired data sources, and all data fetching is done automatically either from Ensembl or from tables in our database. Figure 6.1 shows the software environment and the interfaces with different databases and in Figure 6.2.A a screenshot of ENDEAVOUR is shown in the "training set view". The scoring of test genes is simply done by pressing a button, after which the genes can be sorted according to all individual and combined rankings. Visualization of the results is done with a "sprintplot" where all rankings can be viewed at once (Figure 6.2.B).

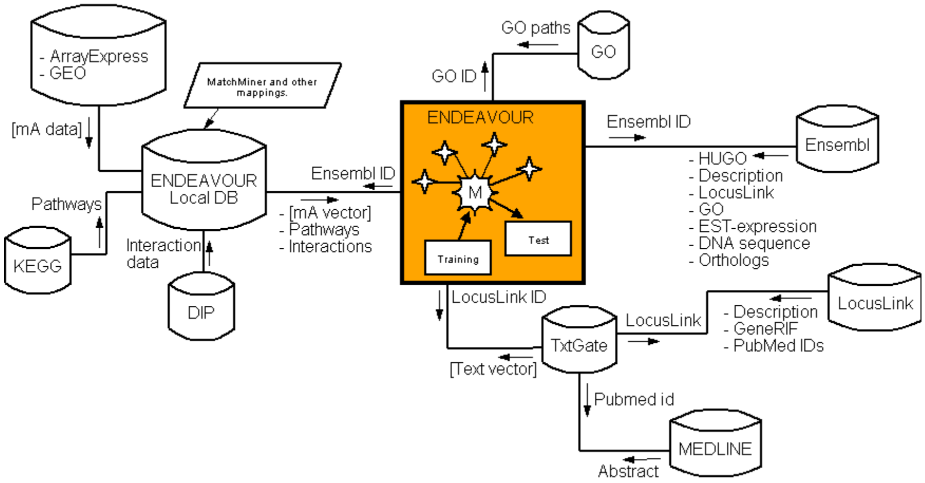


Figure 6.1: Information technological overview of ENDEAVOUR. The M within the main square is the model to be trained, connected with all information submodels (ISM) represented as stars. For each link with an external or local database, the communication details in each direction are given.

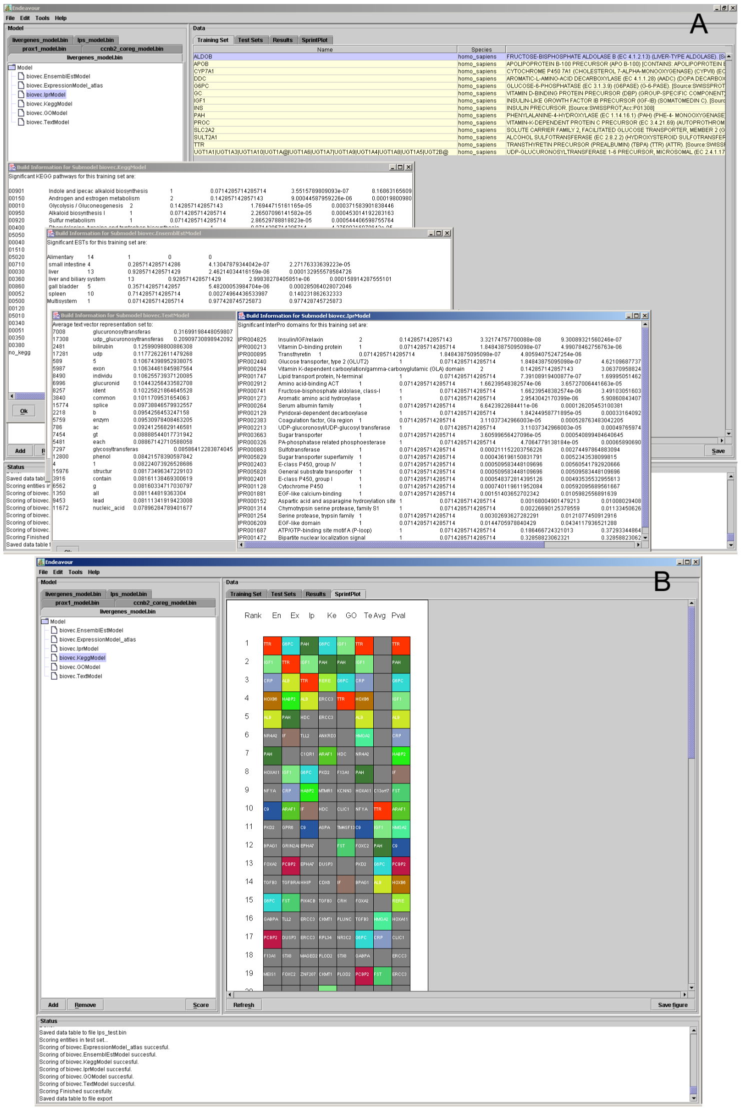
6.5 Cross validation

To validate the proposed methodology, that is to test whether the test genes that are most related to the training genes have the best combined rank and whether they have a significant p-value, we have performed a cross-validation procedure using a set of 14 liver-specific genes [168]. In each of 14 runs, one gene is left out, and a model is trained for the other 13 genes using the six data sources mentioned above. Then 199 randomly selected genes from the human genome together with the left-out gene are scored with the model. We have done this for three different random sets and averaged the rankings. For the combined ranking, the left-out gene was found in the top 3 (of 200) in 12 of the 14 runs (85.7%). The other two left-out genes were ranked 12th and 24th on average. The overall average combined rank was 4.5, and for 13 out of 14 genes the Bonferroni-corrected combined p-value obtained by Equation 6.6 was significant for $\alpha < 0.05$. The rankings (within 200 test genes) according to the individual data sources were on average 1 for the EST expression data (always ranked first), 1.74 for the KEGG model (although not many genes have KEGG annotation), 5.93 for the microarray data, 12.26 for GO, 13.10 for text, and 24.12 for InterPro. This is 9.69 on average, which is slightly worse than the average combined rank of 4.5. In conclusion, these data prove that sorting in ENDEAVOUR, both using the data separately and combined, is biologically meaningful.

6.6 Case studies

The ModuleSearcher was used to detect a module in the same set of liver genes that was used for the cross-validation. This set is the training set $\mathcal{A}$. For 10 out of the 14 genes we found a mouse orthologous gene in Ensembl that has at least one conserved non-coding sequence (CNS) in the 10 kb upstream of the coding sequence as detected with AVID (see Chapters 4 and 5). The ModuleSearcher then finds the optimal combination of PWMs in the set of CNSs using all 86 PWMs in the JASPAR database [252]. For 5 PWMs and a 100 bp window the best module was *mod*=[MA0042/HFH-3, MA0038/Gfi, MA0045/HMG-IY, MA0047/HNF-3beta, MA0046/HNF-1] which seems reasonable for liver specificity. We aligned all human-mouse orthologous gene pairs and scored all CNSs with all PWMs of JASPAR. Then we used our ModuleScanner [7] to score the CNSs with *mod*. The best scoring 200 genes are then used as test set $\mathcal{B}$ in ENDEAVOUR. When they are scored with the model trained on $\mathcal{A}$, the best 4 genes of $\mathcal{B}$ are also part of $\mathcal{A}$. That means that *mod* was specific enough to find four of the ten genes back in the genome. Among the other significant or highly ranked genes are liver related genes like albumin, hyaluronan binding protein 2, follistatin, *ARAF1*, and *RERE*.

As a second example of target gene prioritization after module scanner, we have used the set of genes that are co-regulated with cyclin B2 as described in Section 5.5.3 as training genes and the top 200 genes of the ModuleScan-



Name	Species	
ALDOB	homo_sapiens	FRUCTOSE-BISPHOSPHATE ALDOLASE B (EC 4.1.2.13) (LIVER-TYPE ALDOLASE). [S
APOB	homo_sapiens	APOLIPOPROTEIN B-100 PRECURSOR (APO B-100) [CONTAINS: APOLIPOPROTEIN
CYP7A1	homo_sapiens	CYTOCHROME P450 7A1 (CHOLESTEROL 7-ALPHA-MONOOXYGENASE) (CYPVII) (EC
DDC	homo_sapiens	AROMATIC-L-AMINO-ACID DECARBOXYLASE (EC 4.1.1.28) (AADC) (DOPA DECARBO
G6PC	homo_sapiens	GLUCOSE-6-PHOSPHATASE (EC 3.1.3.9) (G6PASE) (G-6-PASE). [Source:SWISSPROT
GC	homo_sapiens	VITAMIN D-BINDING PROTEIN PRECURSOR (DBP) (GROUP-SPECIFIC COMPONENT
IGF1	homo_sapiens	INSULIN-LIKE GROWTH FACTOR IB PRECURSOR (IGF-IB) (SOMATOMEDIN C). [Sourc
INS	homo_sapiens	INSULIN PRECURSOR. [Source:SWISSPROT;Acc:P01308]
PAH	homo_sapiens	PHENYLALANINE-4-HYDROXYLASE (EC 1.14.16.1) (PAH) (PHE-4- MONOOXYGENASE
PROC	homo_sapiens	VITAMIN-K-DEPENDENT PROTEIN C PRECURSOR (EC 3.4.21.69) (AUTOPROTHROM
SLC2A2	homo_sapiens	SOLUTE CARRIER FAMILY 2, FACILITATED GLUCOSE TRANSPORTER, MEMBER 2 (G
SULT2A1	homo_sapiens	ALCOHOL SULFOTRANSFERASE (EC 2.8.2.2) (HYDROXYSTEROID SULFOTRANSFER
TTR	homo_sapiens	TRANSTHYRETIN PRECURSOR (PREALBUMIN) (TBPA) (TTR) (ATTR). [Source:SWISS
UGT1A1|UGT1A3|UGT1A10|UGT1A@|UGT1A6|UGT1A7|UGT1A9|UGT1A4|UGT1A8|UGT1A5|UGT2B@	homo_sapiens	UDP-GLUCURONOSYLTRANSFERASE 1-6 PRECURSOR, MICROSOMAL (EC 2.4.1.17

Build Information for Submodel biovec.KeggModel

```
Significant KEGG pathways for this training set are:

00901       Indole and ipecac alkaloid biosynthesis      1          0.0714285714285714      3.5515789809093e-07     8.16863165609
00150       Androgen and estrogen metabolism             2          0.142857142857143       9.00044587959226e-06    0.00019800980
00010       Glycolysis / Gluconeogenesis    2      0.142857142857143     1.76944715161165e-05    0.000371583901838446
00950       Alkaloid biosynthesis I         1      0.0714285714285714    2.26507096141582e-05    0.000453014192283163
00920       Sulfur metabolism               1      0.0714285714285714    2.86529788818823e-05    0.000544406598755764
00400
05050
00040
01510
05020
00710
00030
00360
00860
00052
00500
00120
05010
00340
00051
00350
00380
no_kegg
```

Build Information for Submodel biovec.EnsemblEstModel

```
Significant ESTs for this training set are:

Alimentary       14       1                     0                        0
 small intestine 4        0.285714285714286     4.13047879344042e-07     2.27176333639223e-05
 liver           13       0.928571428571429     2.46214034416159e-06     0.000132955578584726
 liver and biliary system      13      0.928571428571429     2.99838278405851e-06     0.000158914287555101
 gall bladder    5        0.357142857142857     5.48200053984704e-06     0.000285064028072046
 spleen          10       0.714285714285714     0.00274964436533987      0.140231862632333
Multisystem      1        0.0714285714285714    0.977428745725873        0.977428745725873
```

Build Information for Submodel biovec.TextModel

```
Average text vector representation set to:
7008        glucuronosyltransferas           0.31699198448059807
17308       udp_glucuronosyltransferas       0.20909730898942092
2481        bilirubin        0.12599098800886308
17281       udp              0.11772622611479268
589         5                0.10674398952938075
5987        exon             0.10634461845987564
8490        individu         0.10625573937120085
6996        glucuronid       0.10443256433582708
8257        ident            0.10225821864645528
3840        common           0.1011709531654063
15774       splice           0.09738846579932557
2218        b                0.0954256453247158
5759        enzym            0.09530978408463205
786         ac               0.09241256829146581
7454        gt               0.08888544017731942
5481        each             0.08867142710588058
7297        glycosyltransferas               0.08586412283874045
12800       phenol           0.08421578390597842
4           1                0.08224073926528686
15976       structur         0.08173496347229103
3916        contain          0.08161138469300619
6562        g                0.08160334717030797
1350        all              0.081144819363304
9453        lead             0.08111341919423008
11672       nucleic_acid     0.07896284789401677
```

Build Information for Submodel biovec.IprModel

```
Significant InterPro domains for this training set are:

IPR004825   Insulin/IGF/relaxin                2           0.142857142857143      3.32174757700088e-08    9.30089321560246e-07
IPR000213   Vitamin D-binding protein          1           0.0714285714285714     1.84843875095098e-07    4.99078462756763e-06
IPR000895   Transthyretin    1      0.0714285714285714      1.84843875095098e-07    4.80594075247254e-06
IPR002440   Glucose transporter, type 2 (GLUT2)            1      0.0714285714285714     1.84843875095098e-07    4.62109687737
IPR000294   Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain     2      0.142857142857143      3.06370958824
IPR001747   Lipid transport protein, N-terminal            1      0.0714285714285714     7.39108919400877e-07    1.69995051462
IPR002912   Amino acid-binding ACT             1           0.0714285714285714     1.66239548382574e-06    3.65727006441663e-05
IPR000741   Fructose-bisphosphate aldolase, class-I        1      0.0714285714285714     1.66239548382574e-06    3.49103051603
IPR001273   Aromatic amino acid hydroxylase                1      0.0714285714285714     2.9543042170399e-06     5.90860843407
IPR000264   Serum albumin family               1           0.0714285714285714     6.6423922684411e-06     0.000126205453100381
IPR002129   Pyridoxal-dependent decarboxylase              1      0.0714285714285714     1.84244958771895e-05    0.00033164092
IPR002383   Coagulation factor, Gla region     1           0.0714285714285714     3.11037342966003e-05    0.000528763483042205
IPR002213   UDP-glucoronosyl/UDP-glucosyl transferase      1      0.0714285714285714     3.11037342966003e-05    0.00049765974
IPR003663   Sugar transporter                  1           0.0714285714285714     3.60599656427096e-05    0.000540899484640645
IPR000326   PA-phosphatase related phosphoesterase         1      0.0714285714285714     4.7064779138184e-05     0.00065890690
IPR000863   Sulfotransferase                   1           0.0714285714285714     0.000211152203756226    0.00274497864883094
IPR005829   Sugar transporter superfamily      1           0.0714285714285714     0.000436196150831791    0.00523435380099815
IPR002403   E-class P450, group IV             1           0.0714285714285714     0.000509583448109696    0.00560541792920666
IPR005828   General substrate transporter      1           0.0714285714285714     0.000509583448109696    0.00509583448109696
IPR002401   E-class P450, group I              1           0.0714285714285714     0.000548372814395126    0.00493535532955613
IPR001128   Cytochrome P450                    1           0.0714285714285714     0.000740119611952084    0.00592095689561667
IPR001881   EGF-like calcium-binding           1           0.0714285714285714     0.00151403652702342     0.0105982556891639
IPR000152   Aspartic acid and asparagine hydroxylation site  1      0.0714285714285714     0.00168004901479213     0.01008029408
IPR001314   Chymotrypsin serine protease, family S1        1      0.0714285714285714     0.00226690125378559     0.01133450626
IPR001254   Serine protease, trypsin family    1           0.0714285714285714     0.00302693627282291     0.0121077450912916
IPR006209   EGF-like domain                    1           0.0714285714285714     0.0144705978840429      0.0434117936521288
IPR001687   ATP/GTP-binding site motif A (P-loop)          1      0.0714285714285714     0.186466724321013       0.37293344864
IPR001472   Bipartite nuclear localization signal          1      0.0714285714285714     0.32858823062321        0.32858823062
```

Status

```
Saved data tabl
Scoring entities ir
Scoring of biovec.
Scoring of biovec.
Scoring of biovec.
Scoring of biovec.
Scoring of biovec.
Scoring of biovec.
Scoring Finished
Saved data table f
```

Rank	En	Ex	Ip	Ke	GO	Te	Avg	Pval
1	TTR	G6PC	PAH	G6PC	IGF1	TTR		TTR
2	IGF1	TTR	IGF1	PAH	PAH	IGF1		PAH
3	CRP	ALB	TTR	RERE	G6PC	CRP		G6PC
4	HOXB6	HABP2	ALB	ERCC3	TTR	HOXB6		IGF1
5	ALB	PAH	HDC	ERCC3		ALB		ALB
6	NR4A2	IF	TLL2	ANKRD3		HMGA2		CRP
7	PAH		C1QR1	ARAF1	HDC	NR4A2		HABP2
8	HOXA11	IGF1	G6PC	PKD2	F13A1	PAH		IF
9	NFYA	CRP	HABP2	MTMR1	KCNN3	HOXA11	C13orf7	FST
10	C9	ARAF1	IF	HDC	CLIC1	NFYA	TTR	ARAF1
11	PKD2	GPR6	C9	ASPA	TM4SF13	C9	IGF1	HMGA2
12	BPAG1	GRIN2A	EPHA7		FST	FOXC2	PAH	C9
13	FOXA2	PCBP2	EPHA7	DUSP3		PKD2	G6PC	PCBP2
14	TGFB3	TGFBRA	HHIP	CDK6	IF	BPAG1	ALB	HOXB6
15	G6PC	FST	PIK4CB	TGFB3	CRH	FOXA2		RERE
16	GABPA	TLL2	ERCC3	CKMT1	PLUNC	TGFB3	HMGA2	HOXA11
17	PCBP2	DUSP3	ERCC3	RPL34	NR3C2	G6PC	CRP	CLIC1
18	F13A1	STX8	MAGED2	PLOD2	STX8	GABPA		ERCC3
19	MEIS1	FOXC2	ZNF207	CKMT1	PLOD2	PCBP2	FST	ERCC3

Status

```
Saved data table to file lps_test.bin
Scoring entities in test set...
Scoring of biovec.ExpressionModel_atlas succesful.
Scoring of biovec.EnsemblEstModel succesful.
Scoring of biovec.KeggModel succesful.
Scoring of biovec.IprModel succesful.
Scoring of biovec.GOModel succesful.
Scoring of biovec.TextModel succesful.
Scoring Finished succesfully.
Saved data table to file export
```

Figure 6.2: Screenshots of the ENDEAVOUR graphical user interface. (A) training set and a number of trained submodels are shown. The view of a test set is similar. (B) One view of the results is a sprintplot where each column contains a ranking according to one of the submodels. Each gene has a color, for example the TTR gene in red is ranked 1st, 2nd, 3rd, none, 4th, 1st for the respective submodels. Another view of the results that is not shown here is a spreadsheet-like table view with the genes as rows and the submodels as columns where the genes can be sorted according to either of the submodels.

ner as test genes. The following genes have a statistically significant ranking (p-value<0.05): *PLK, CDC2, NEK2, CKS1B, SHC1, HMG20B, PRG4, PRKAA2, DDX5, BPAG1, SEMA3C, TOP2B, and ORC6L.* Even more genes than reported in Section 5.5.3 (reported by manual literature searching) seem related with the training genes and it could therefore be worth investigating the role of the detected module for these genes.

6.7 Conclusions

In this chapter we discussed the design and implementation of a strategy to prioritize a set of test genes according to their similarity with a set of training genes, based on multiple genomic information sources. The cross-validation showed the capabilities of the system, and in the case studies the system could properly prioritize putative target genes of a *cis*-regulatory module.

The major limitation of the system regarding the validation of *cis*-regulatory module is that the presence of significantly highly ranked test genes cannot guarantee that the module is a real module. It can only help the investigator to choose the best candidate targets for further investigation (e.g., experimental validation). This problem might be caused by a too low stringency in the statistical procedure or by the ranking methodology in general. A possible solution to this problem is the use of pattern classification methods like Support Vector Machines (SVM).

6.7.1 Perspectives

We are currently investigating other data sources like phylogenetic data and promoter profiles of TFBSs (both vector data). Another data source that is being tested is the actual DNA sequence (the coding sequence) of the genes. This data type does not fall under the vector based system nor under the non-vector based system described in this chapter. Rather, each test gene is aligned with all training genes using BLAST and are ranked according to the best similarity score. Note that this gene-by-gene comparison instead of the comparison with the average training profile can also be interesting for the other data types. A test gene is then scored better if it is very similar to one training gene and not necessarily to all or the average of the training genes. A further perspective regarding the microarray data source is that the framework can potentially be linked directly with microarray repositories like ArrayExpress or GEO in the future (possibly via Ensembl). This would allow for the comparison of gene expression measured under hundreds of conditions.

6.7.2 Perspectives on the computational prioritization of candidate disease genes

It is clear that the ENDEAVOUR framework can be used to sort any list of genes according to a list of training genes. A particularly interesting application is the prioritization of candidate disease genes in chromosomal regions that are mapped to a disease by linkage and association studies. Such regions may contain ∼100-200 genes, which is too much for individual experimental validation[2] or large-scale SNP association studies in human populations. Therefore, a computational prioritization could lead to a reduction of the potential candidates and hopefully to a high success rate when testing only the top scoring genes (e.g., the top 10). The training set for this application can be constructed either by an expert in the genetics of the disease under study, or by using databases with gene-disease information like the Online Mendelian Inheritance in Man resource (OMIM).

Several studies have recently appeared that describe the computational prioritization of candidate disease genes [302, 138, 306, 210, 113, 231]. However, none of these has used a large coverage of data types as available in ENDEAVOUR.

---

[2] Experimental validation of candidate disease genes in model organisms is expensive and time consuming. For example, the construction of a mouse knockout takes up to one year and experimental tests in zebrafish using morpholino's [70, 141] can cost up to 1000 EUR per gene.

Chapter 7

Comprehensive analysis of the base composition around the transcription start site in Metazoa

THE transcription start site (TSS) of a metazoan gene remains poorly understood, mostly because there is no clear signal present in all genes. Now that several sequenced metazoan genomes have been annotated, we have been able to compare the base composition around the TSS for all genes across multiple genomes. The most prominent feature in the base compositions is a significant local variation in G+C content over a large region around the TSS. The change is present in all animal phyla but the extent of variation is different between distinct classes of vertebrates, and the shape of the variation is completely different between vertebrates and arthropods. Furthermore, the height of the variation correlates with CpG frequencies in vertebrates but not in invertebrates and it also correlates with gene expression, especially in mammals. We also detect GC and AT skews in all clades (%G≠%C or %A≠%T respectively) but these occur in a more confined region around the TSS and in the coding region. The dramatic changes in nucleotide composition in human are a consequence of CpG nucleotide frequencies and of gene expression, the changes in fugu could point to primordial CpG islands, and the changes in fly are of a totally different kind and unrelated to dinucleotide frequencies.

7.1 Background

Genomic DNA sequences display compositional heterogeneity on several scales—for example, long-range variations in G+C content (large blocks of DNA of different compositions are often referred to as "isochores" [27]), CpG suppression in vertebrate genomes [30], or skews caused by mutation biases intrinsic to mutation and repair mechanisms [127]. Both neutralist hypotheses and selectionist hypotheses have been made to explain the various compositional variations [102, 112]. Until recently it was difficult to investigate more local variations in base composition (for example, at one position relative to some genomic signal). Although there are currently many efforts to understand metazoan gene regulation and transcriptional control, we have only a limited knowledge of the exact start of transcription. In this study we re-evaluate the average base composition around the transcription start site of animal genes [5]. We could both confirm several aspects regarding nucleotide composition and we discovered new aspects, especially in invertebrates. It is most obvious from our results that the average nucleotide composition around the TSS across the genome is significantly different from the composition in the intergenic and coding regions and some aspects of these composition variations are furthermore different among the investigated species.

7.2 Data and methods

For each sequenced organism that is available in the Ensembl database Release 14 (*Homo sapiens, Mus musculus, Rattus norvegicus, Fugu rubripes, Danio rerio, Drosophila melanogaster* [109], *Anopheles gambiae, Caenorhabditis elegans* [135], and *Caenorhabditis briggsae* [135]), we have randomly selected 5000 stable gene identifiers [149]. These lists were used to retrieve 2000 base pairs (bp) of single stranded DNA from the synonymous strand upstream of the annotated starting point(s) of each gene and 1000 bp downstream. This was done using the EnsMart data mining tool (http://www.ensembl.org/EnsMart). The analysis and plotting of the average base pair composition of these sequences is done as follows. For each position in the 3000 bp long sequences the percentage of A, T, C, and G over the 5000 genes is calculated and this value is represented on the y axis of all profile figures. The x axis shows the position along the sequence and x=0 corresponds to position +1, the start of the annotated gene or the putative transcription start site. This way of representing the nucleotide composition at an aligned genomic position across many genes—as opposed to a classical average base composition calculated over a window along the DNA strands as in [264]—has been used before for purposes like the study of GC skews in Arabidopsis [291], the base composition of complete genes (introns, exons, etc.) [199, 193], and promoter prediction [46, 223]. Many genes have multiple alternative transcripts with a different TSS. Using DNA regions around each possible TSS of a gene or only around the furthest 5' reaching TSS did not influence the composition profiles (data not shown). For the sake of brevity we only discuss human, fly, and fugu profiles. Mouse and rat were very similar to human, profiles of mosquito were noisy and difficult to interpret, and *C. elegans* and *C. briggsae* are omitted because the interpretation would be too difficult because of trans-splicing at the 5' end of the genes [37, 312].

7.3 Results and discussion

7.3.1 Comparing Ensembl and DBTSS human gene start annotations

From the extraordinary shapes of the composition profiles calculated using the gene start annotations of Ensembl (Figure 7.1.B and Figure 7.2) it can already be postulated that a significant degree of correct start annotation must be present in Ensembl to get such high resolutions. To double check this statement (for human only) we have downloaded all human promoter sequences from DBTSS [287] for which the TSS has been determined experimentally. It can be seen from Figure 7.1 that the Ensembl data (using 5000 randomly selected genes with at least 100 bp 5'UTR) is noisier but that most of the composition characteristics (as discussed below) are also present in the profiles generated from the Ensembl data. The TATA box is less clear and the GC rise is lower for the Ensembl data than for the DBTSS data. The reason for the latter observation will be given below. We have also checked the quality of the *Drosophila* start points by comparing the nucleotide frequencies around Ensembl (i.e., annotation from FlyBase) gene starts with a data set of experimentally determined TSSs of [224], and they were highly similar (not shown).

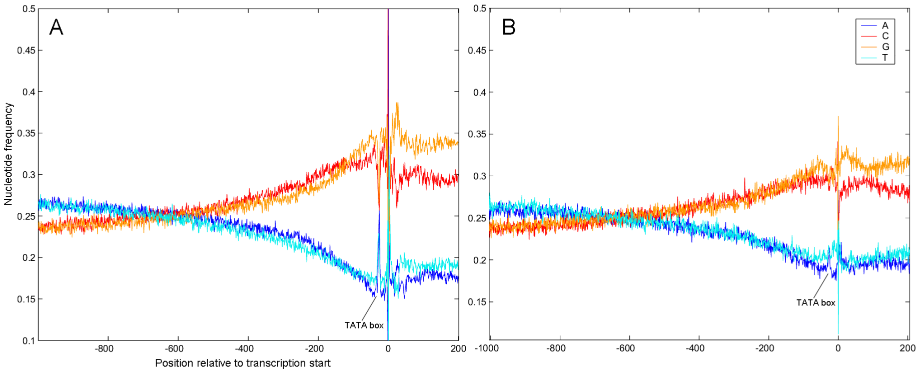


Figure 7.1: Comparing Ensembl gene starts with DBTSS. (A) Nucleotide frequencies around the experimentally determined transcription start site of all genes in DBTSS. (B) Frequencies around the annotated gene start in Ensembl for 5000 randomly selected genes.

7.3.2 Variations in base composition in different phyla

Figure 7.2 shows the nucleotide frequencies around TSS for human, fly, and fugu. A characteristic that is shared among all investigated species is that the A/T content (W, IUPAC alphabet) is greater than the G/C content (S, IUPAC alphabet) in the intergenic region; for example, at -2000 bp upstream of the TSS. This is the result of the fact that in general the G:C→A:T base pair

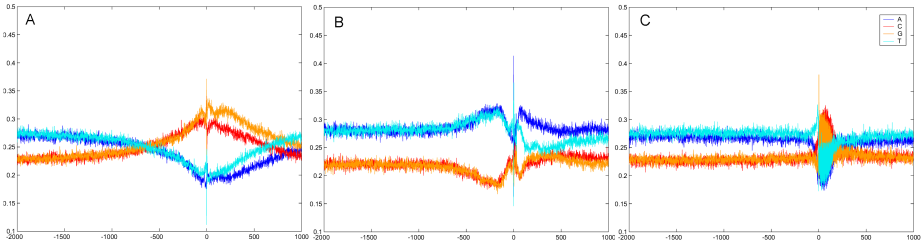


Figure 7.2: Nucleotide frequencies around the annotated gene start in Ensembl, calculated from 5000 randomly selected genes in human (A), *Drosophila* (B), and fugu (C).

transition frequency is significantly higher than that of the reverse T:A→C:G transition. Thus accumulation of neutral substitutions results in a generally GC-poor composition of mammalian genomes [198], and apparently also of other vertebrate and also invertebrate genomes. We will further denote this composition as the intergenic background composition (IBC), and we will denote a difference between the A+T content and the G+C content as ΔWS $=\#[(A+T)-(G+C)]/(A+T+G+C)$.

The most notable features of the composition profiles are the dramatic changes in ΔWS in the region [-1000,+1000] around the TSS. In human for example, ΔWS changes from $\sim$ 10% in the IBC to $\sim$ -20% at the TSS (also see Figure 7.8). A similar polarity switch of ΔWS can be seen in the other vertebrates: mouse, rat, fugu, and zebrafish (see Figure 7.2.C for fugu). The mouse patterns are similar to human (not shown). The fugu and zebrafish patterns also have the same shape with a polarity switch but the composition starts to change later than in mammals and is restored faster as well. The fast drop in G+C content might be caused by the fact that the 5'UTRs in fish are much shorter than in human so the coding region (where codon usage largely determines base composition) starts immediately after the TSS. A common explanation for the G+C rise that is seen here in the mammalian profile in the proximity of the TSS is the presence of CpG islands, which is related to DNA methylation, or more precisely to a lack of DNA methylation (see further). *Drosophila* (Figure 7.2.B) also shows a significant change in ΔWS, but without a polarity switch: it increases from $\sim$ 12% in the IBC to $\sim$ 26% at the TSS. The maximal difference between ΔWS_{IBC} and ΔWS_{TSS} is not reached at the TSS itself as in vertebrates, but about 150 base pairs before the TSS. We have no explanation for the *Drosophila* patterns that show almost an opposite behavior to that of vertebrates, but because of the absence of DNA methylation in *Drosophila*, a rise in G+C because of an over-representation of CpG dinucleotides would not be expected anyway (although DNA methylation in insects has been the subject of some debate [143]).

Interestingly, because *Drosophila* does have a change in ΔWS, namely an opposite change to that of vertebrates, there are perhaps factors other than DNA methylation that influence the base composition in this species. One factor could be the general presence of more AT-rich binding sites for transcription factors or histone modification factors [266]. An alternative hypothesis could be that another type of DNA modification than CpG methylation would be involved in a genome-wide marking of promoter regions in *Drosophila*.

7.3.3 Nucleotide composition and CpG islands

Above we have made the remark that the G+C rise in mammals and maybe generally in vertebrates is probably caused by the higher number of CpG dinucleotides in the promoter region. Normally CpGs are present at a frequency of only ~1.5% instead of their expected frequency of ~5% based on the individual frequencies of C and G (0.225×0.225). Indeed, most CpGs in the genome are methylated at the cytosine [29] and those methylated cytosines frequently mutate to thymines [71].

To investigate the relationship between CpG frequency and the observed composition profiles, we compared the base compositions between genes with and without a CpG island around the TSS. We did not use a CpG prediction algorithm however to separate CpG-related genes from non-CpG-related genes because CpG island prediction is done using an arbitrary threshold on the number of CpG doublets as compared to the genome frequency. Instead we have taken another approach by simply counting the CpG doublets in the [-400,+400] region around TSS. The same technique was used by Ioshikhes and colleagues [153]. A histogram of CpG numbers for 5000 randomly selected genes is bimodal for human, but not for fly nor fish (see Figure 7.3). For human, the first peak represents the genes with CpG numbers that correspond more or less to the genome frequency and the second peak represents genes with more than expected numbers of CpGs.

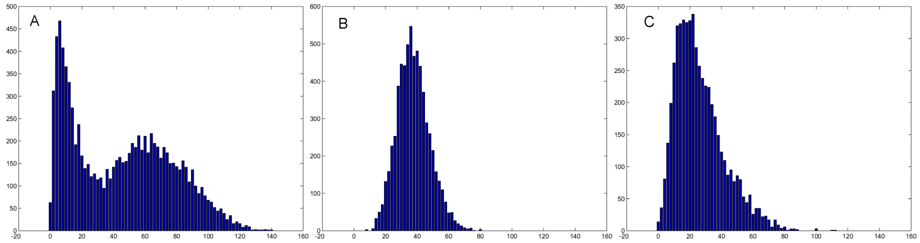


Figure 7.3: Dinucleotide frequencies. Frequency distributions of the CpG dinucleotide in the [-400,400] region around the TSS in human (A), fly (B), and fugu (C). The number of CG doublets in this window is on the x axis and the number of regions (genes) containing this number of CGs is on the y axis.

The histogram of CpG scores for fish shows almost no second peak, but the distribution is slightly broader than the first peak of the human distribution. This could mean that there is some DNA methylation and some CpG over-representation around TSS but not as much as in human. Auf der Maur et al. [16] have suggested that CpG islands of fish may represent a primordial stage of CpG island evolution. This could indeed be a plausible explanation for the fugu distribution.

The distribution of CpG frequencies in *Drosophila* is a normal distribution, which means that there is nothing special about CpG doublets in *Drosophila*. This is in agreement with the fact that there is no DNA methylation in *Drosophila*. To test whether another dinucleotide than CpG is over-represented around TSS in fly, we have performed the same analysis for the fifteen other possible dinucleotides and looked for a distribution like the CpGs for human or fish, but all dinucleotide frequencies were normally distributed and similar to the CpG distribution, although the WpW dinucleotides (AA, AT, TA, TT) had a slightly broader distribution and a higher mean.

To see the effect of the CpG concentration on the overall nucleotide composition, we have plotted the base composition profiles separately for the 15% lowest scoring and the 15% highest scoring genes (see Figure 7.4). In human (A-B), this shows that ΔWS can be completely attributed to CpG over-representation. The results for fugu (C-D) show that some genes could have CpG islands (D) since for those the nucleotide composition is similar to the mammalian profiles. This again can be in agreement with the hypothesis of primordial CpG island evolution in other vertebrates than mammals, although other tests are needed to check for a possible functional consequence of the differences between the extremes of the CpG frequency distribution. If we look at the two ends of the distribution of AT dinucleotides in *Drosophila*, we can see a similar breaking apart of the composition profiles into genes with a small ΔWS and genes with a large ΔWS (E-F). The question is whether these gene classes in fugu and fly also have a functional meaning like in human, or that these visualizations are artefacts due to plotting the extremes of the distributions. One clue that supports a functional meaning for such gene classes is the fact that the profiles of fly genes with many (respectively few) CpGs or few (respectively many) ApTs are exactly the same and that 50% of the genes classified in the "few CpG" category are also present in the "many ApT category". Below we will test the dependency of the composition profiles on gene expression.

7.3.4 Nucleotide composition and gene expression

It is generally known that the presence of a CpG island around the TSS is related to the expression pattern of the gene. Unmethylated DNA can have an open chromatine structure that facilitates the interaction of transcription factors with the promoter region. Housekeeping genes (HK genes), which are transcribed in all somatic cells and under all circumstances (and thus should be easily activated) frequently have a CpG island in their promoter region [175, 235]. Ponger et al. [235] showed that early embryo genes (both housekeeping and tissue specific genes) that are active at the totipotent cell stage or in the blastocyste are associated with CpG islands [235]. We have shown above that our composition profiles are caused by CpG islands, so we can expect to see differences in base composition between genes with different expression patterns. We identified sets of widely and narrowly expressed genes using microarray data using a similar

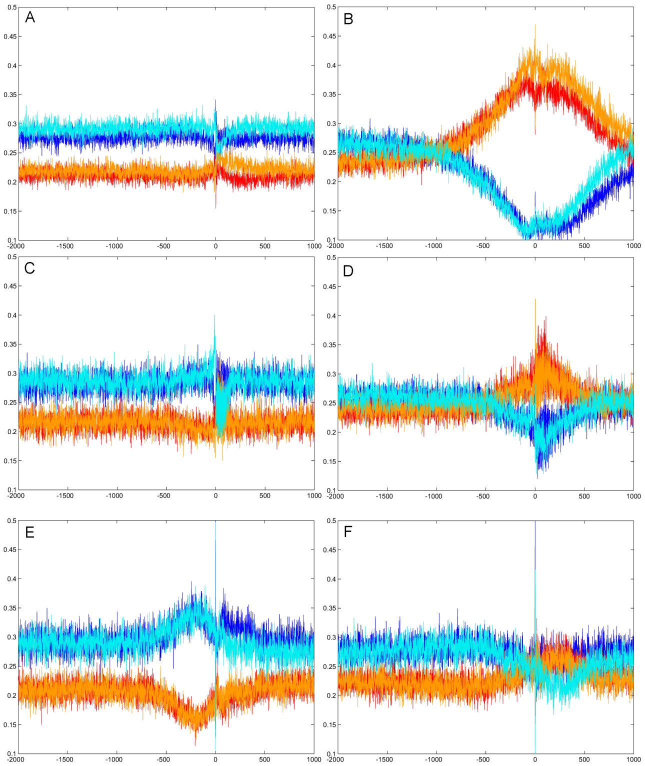


Figure 7.4: Nucleotide frequencies of several gene classes, separated according to the concentration of a dinucleotide in the [-400,400] region around the TSS. (A) Human genes with few CpG doublets. (B) Human genes with many CpG doublets. (C) Fugu genes with few CpGs. (D) Fugu genes with many CpGs. (E) Fly genes with many ApTs. (F) Fly genes with few ApT's.

analysis as Eisenberg and colleagues in [99]. We used microarray expression data from 101 different samples taken from 47 different human tissues and cell lines under normal physiological state [283]. The experiments measuring replicates of the same biological condition were averaged to reduce the measurement noise, resulting in 47 data points per probe. We have selected three probe sets with an average reading above 200 standard Affymetrix difference units [99] in the following conditions: (1) in all tissues, these are widely expressed genes; (2) in 20 to 29 out of 47 tissues (medium expression); and (3) in only 1 tissue (narrow expression). Then we mapped the Affymetrix probe identifiers to HUGO gene names using MatchMiner [55] and used these lists to retrieve the corresponding sequences using EnsMart. The size of the sets are respectively 647, 886, and 783 genes. Figure 7.5 shows the average base composition graphs for the three sets.

It can be seen that the more widely the genes are expressed, the more pro-

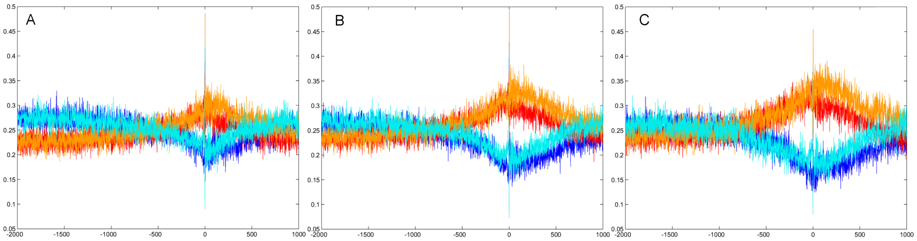


Figure 7.5: Nucleotide frequencies of three human gene groups: genes with a narrow expression pattern (A), a medium pattern (B), and a wide pattern(C).

nounced the variations in ΔWS are. These differences prove that the composition changes and the effect of methylation on gene expression are functionally conserved, and that the nucleotide compositions are not the result of some kind of mutational bias. The fact that widely expressed genes, regardless of the level of expression, would need a promoter that is easily accessible (e.g., by an open chromatine structure) would make sense in an evolutionary perspective. For these genes one could expect that their regulation depends less on specific *cis*-regulatory modules than on the accessibility of the proximal promoter. Also note that the fact that the profiles from DBTSS in Figure 7.1 show a steeper rise than the Ensembl-based profiles can be the result of a slight bias of DBTSS towards genes with high expression levels.

To test whether the nucleotide composition also depends on gene expression in fugu, we have worked under the assumption that fugu genes that are orthologous to human genes that are widely (or narrowly) expressed, are also widely (or narrowly) expressed. For each of the three human gene sets above, the fugu orthologous genes were retrieved from the Ensembl database and the nucleotide frequencies were calculated (see Figure 7.6). As opposed to human, almost no variation between the groups is observed. This can be due to the fact that the control of methylation (i.e., keeping promoters unmethylated; for human this is reflected in the second peak of the bimodal CpG distribution) is not or only slightly present in fugu. It cannot be ruled out however that the absence of a clear trend could be due to the fact that the expression patterns among orthologous genes are not well preserved or again that the gene start annotations are of too low quality.

Drosophila shows a completely different behavior in composition changes so we were interested to see whether these changes also vary with the level of gene expression. Unfortunately we could not find a similar microarray experiment in *Drosophila* that compares different tissues under normal circumstances, and the mapping of human genes to *Drosophila* orthologs results in too few genes. As an alternative we have selected gene sets with different EST expression patterns from the Unigene database, namely (1) Unigene clusters with only one expression site (and leaving out the clusters with *whole body* expression) (narrow expression) and (2) the 2000 clusters with the most expression sites (wide

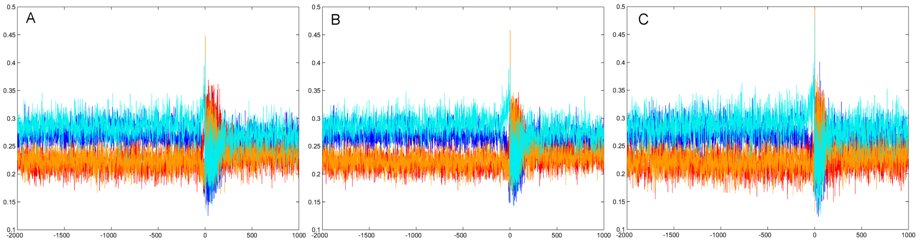


Figure 7.6: Nucleotide frequencies of three groups of Fugu genes that are orthologous to the respective gene groups in Figure 7.5.

expression). These two sets are displayed in Figure 7.7.

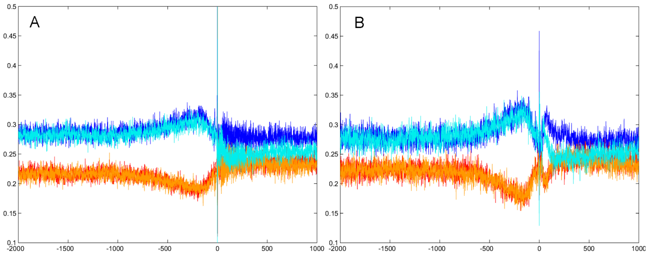


Figure 7.7: Nucleotide frequencies of two fly gene groups: genes with a narrow expression pattern (A), and a wide pattern(B) as determined by their EST expression pattern obtained from the Unigene database.

The difference between the profiles of wide and narrow expression is minimal. The A+T maximum and the G+C minimum are more or less the same, only the rise in A+T is a little bit steeper in the widely expressed genes. This finding however might be caused by the quality of the data set and since there is a small observable difference we would not rule out the possibility that differences could be seen in the future when more appropriate data sets are available.

Another way of visualizing the variation of ΔWS, is by directly plotting ΔWS, as done in Figure 7.8.

7.3.5 GC and AT skews around the TSS

Chargaff's second parity rule states that the number of As equals the number of Ts, and the number of Cs equals the number of Gs in a *single* strand over windows of sufficient size, often in the order of 1000 bp [62]. In our composition profiles, at least in the intergenic regions, the number of As also equals the number of Ts (and %G=%C), but this is measured at one position across 5000 genes. An "ergodic"[1] version of Chargaff's second parity rule seems to hold. This variant rule is broken in the [-60,+60] region around the TSS, and also further

---

[1] %G=%C along one sequence and also at one position along multiple sequences.

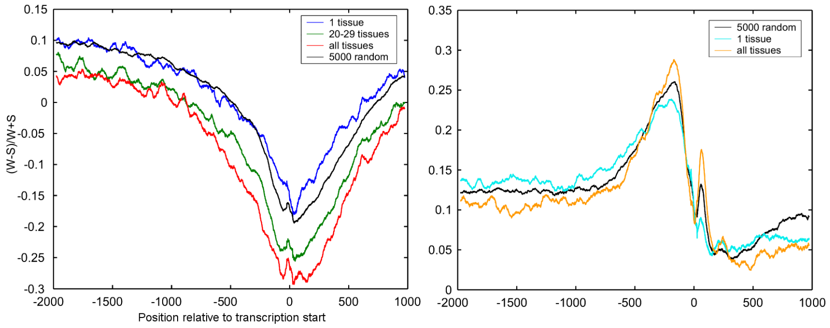


Figure 7.8: ΔWS plotted for different gene groups. (A) ΔWS for four sets of human genes: 5000 randomly selected genes, and the three groups with different expression properties as used in Figure 7.5. It can be seen that ΔWS is indeed varying more in in genes with wide expression. (B) Similar plots for three *Drosophila* gene groups: 5000 random genes and the two gene groups used in Figure 7.7. The differences in ΔWS variation between the groups is clearly less than for human, as discussed in the text.

downstream of the TSS in most species. In vertebrates %A>%T and %G>%C and in invertebrates %T>%A and %C>%G. Such differences are called AT and GC skews and they are measured as (A-T)/(A+T) and (G-C)/(G+C) respectively. The same observation was also made by Louie et al. [193]. The transcription process is asymmetric and might bias mutation patterns between the transcribed and nontranscribed strands by exposing the nontranscribed strand to DNA damage [111]. Both transcription-coupled repair and deamination have been shown experimentally to produce an excess of C→T mutations on the nontranscribed strand in *E. coli* [20, 225]. Green and colleagues have shown that A→G transitions can occur significantly more than T→C transitions in transcribed than in non-transcribed regions (in mammals), which can explain the GC skew (G>C) that is present in the whole region after the TSS in vertebrates [127] (we have used the nontranscribed or synonymous strand in all the analyses). In general, they show that transcripts have a significant G+T compositional excess, and we also see that T>A after TSS. Majewski performed a genome-wide study in human and reported the same mutational asymmetry and he further established a correlation between this symmetry and gene expression [198]. All of this however seems only to make sense for the vertebrate skews. Since A>T after the TSS in *Drosophila* (while the opposite is true in human), either the transcriptional machinery that causes the mutational bias differs between these organisms, or else the skews are functionally conserved with a different function in the two phyla. A last observation regarding skews is the sudden AT skew (where the A and T profiles separate in the plots) that occurs right before the TSS in vertebrates and right after the TSS in arthropods. A similar although less pronounced sudden GC skew can be seen right after the

TSS in vertebrates, but not in arthropods. For these observations we have no explanation.

7.4 Conclusions

In human there is a continuum in gene expression (low–medium–high or narrow-medium-broad) that goes hand in hand with a continuum of CpG doublet concentration around TSS and both are reflected in a continuum of nucleotide frequencies (small–medium–large ΔWS). In other words, genes can differ in their CpG content (and thus in their nucleotide composition) and this difference has a functional meaning (large ΔWS is needed for an "easy" expression, early in the embryo or in many tissues) and is therefore evolutionary conserved. For CpGs in fugu these relations are not so clear, perhaps because CpG islands in fish seem primordial. For *Drosophila* we could not find an analogy of CpG islands. A possible explanation for the A+T rise in the base composition in *Drosophila* could then be that fly genes differ in their AT-content because of differences in the concentration of AT-rich transcription factor binding sites around the TSS.

Chapter 8

General discussion

I have shown that, by integrating multiple data and multiple methods, meaningful results can be obtained for the regulatory sequence analysis in metazoan species, with their large intergenic regions. Several of the methods are new and in some cases combined with existing ones. Also, several of the data sources used have been created during this study and others were taken from public databases.

The general scope of this dissertation is broad and different aspects of the analysis of gene regulatory networks (GRN) are touched upon, some in more detail than others. However, although the principles of GRNs have steered this research, the actual inference, construction, or completion of GRNs is outside the scope of this work. Therefore, the achievements described can be seen as putting out feelers and smoothing the way for computational GRN analysis in Metazoa. To achieve the latter, improvements in bioinformatics techniques are required at the level of microarray data analysis, microarray data comparisons, the detection of transcription factor binding sites in regulatory regions and the detection of the regions themselves, *in silico* validation of computationally generated hypotheses, network structure inference, etc. In this dissertation we have looked at some of these building blocks of systems biology and suggested strategies to make them better or easier to use.

Microarray data analysis

The quantity of mRNA molecules of a gene is the immediate output of a GRN in action. If we want to reversely engineer a GRN, we can therefore use the respective mRNA levels. Of course, since for most GRNs the constituting genes are not yet known, we can only get the mRNA output of *all* GRNs. It is in this massive amount of data that we have to find correlations between genes that could imply a linkage between them. The most straightforward correlation is co-expression: genes with the same expression pattern are potentially co-regulated by the same set of transcription factors, that is they form a *gene battery*, peripheral in the GRN. The inference of linkages between internal genes (encoding TFs) is more difficult because they are generally expressed at lower levels and because their correlation requires the measurement of mRNA levels at higher resolution (e.g., at many conditions), which is currently still expensive.

We have studied gene batteries during neuronal differentiation. Before clustering, we had to remove systematic errors for which we applied state-of-the-art preprocessing techniques that appeared just then [327], and we had to filter the genes to remove bad measurements. For the latter we invented a genewise ANOVA filter to detect only those genes with significantly similar replicate measurements and a filter based on the correlation between the replicate measurements. The clustering was done with existing algorithms (K-means and AQBC [81]) and the clusters were interpreted manually by an expert to dissect the different gene batteries. We also suggested an alternative way to find particular gene batteries in the large heap of GRN outputs, namely by first selecting all genes that are known to be involved in the a particular process (using Gene Ontology) followed by a hierarchical clustering within this group. This resulted in the finding that for many processes there are gene batteries called into action early during neuronal differentiation, and other gene batteries are activated at later stages. The data set and the analysis system is available to the research community as a web-based software tool called **NEURODIFF**.

If the goal in systems biology is to reversely engineer all GRNs we will need massive amounts of mRNA expression levels during all possible conditions. We have written a review paper to give a perspective on the possibilities in this matter [215]. As was realized early by the MGED consortium, a clear description of the biological and experimental conditions of sampling is crucial for the data to be valuable for other investigators. An illustration of the fact that this is not always trivial can be seen in our own comparison of microarray data measured during neuronal differentiation *in vitro* and *in vivo*. It required *time warping* to find out that the differentiation process *in vitro* proceeded faster than *in vivo*. Thus caution should be taken when comparing or averaging conditions in the repositories.

Cis-regulatory sequence analysis

The human genome sequence was published in February 2001 [173, 311] and I started this research path in June 2001. The genomic sequences of worm [295] and fly [2] were available earlier. The sequences of mouse [318], rat, fugu, and zebrafish followed rapidly. Ensembl, a joint project between EMBL-EBI and the Sanger Institute, has taken a leading role in the management of the metazoan sequence assemblies and has created a software system for the automatic annotation of the sequences. Our work has gone together with the developments of Ensembl.

Relevant information in Ensembl for this work include (1) a catalogue of genes for each species with their chromosomal location and an unambiguous identifier, (2) external references for each gene (HUGO name, LocusLink ID, GO associations, protein domains, etc.), (3) the gene structure with exons, introns, 5'UTR, and 3'UTR. The latter was extremely useful to retrieve the 5' upstream sequence of a gene. We have analyzed the properties of such upstream sequences in several metazoans in Chapter 6 and discovered some intriguing phenomena regarding nucleotide skews and regarding differences in nucleotide composition around the transcription start site between clades. Moreover, we have used these upstream sequences to select putative proximal promoter sequences and to find conserved non-coding sequences (CNS) between human and mouse.

These genomic sequences were treated as putative regulatory sequences and were subjected to further analysis, namely the detection of transcription factor binding sites (TFBS). To overcome the high level of false positive predictions we proposed a new method to score a sequence with a position specific frequency matrix (PSFM) that discriminates between a motif and the background and that estimates the number of instances (the **MotifScanner**). Although this resulted in a more robust scoring scheme—showing less variation with the varying parameter—, the short TFBS sequences still occur everywhere by chance without necessarily being functional. We therefore used an approach that is around for about a decade now, namely using only those putative TFBS that are present in a all or many genes of a gene battery. Until now, this approach was used to detect *new* motifs (i.e., DNA words of a certain length) in a sequence set (see Section 2.5.2). By taking advantage of the growing databases of PSFMs like TRANSFAC, we have taken an alternative route by considering only instances of known PSFMs. We used the binomial statistic to measure a significant over-representation. The combination of phylogenetic footprinting (selecting human-mouse conserved regions), the MotifScanner, the known PSFMs, and the over-representation is thus the resulting strategy that, at least for the benchmark data sets, gives meaningful results regarding the transcriptional control of human genes. The system was made available to the research community in the form of a Java software tool called **TOUCAN**.

Then we moved one step further by not selecting the *single* over-represented TFBSs in a gene battery, but instead detecting a significant *combination* of TFBSs, following the biological model of combinatorial transcriptional control in the form of *cis*-regulatory modules. The resulting **ModuleSearcher** algorithm was proven to work on artificial data sets, and the results on a cell cycle gene battery were promising. We realized that since the ModuleSearcher lacks a statistical measure of significance for a module, supporting evidence was needed to check the functional meaning of a certain combination of TFBSs. To this end, we developed the **ModuleScanner** to scan all human-mouse CNSs of the human genome for clusters of these binding sites. The functional coherence of the top scoring genes validated this approach. Further, we developed a strategy to prioritize the top scoring genes according to their functional similarity with the original gene battery where the module was found. The combination of microarray-based expression data, EST-based expression data, information in scientific abstracts, KEGG pathway membership information, predicted protein domains, and Gene Ontology associations was proven to give better prioritizations than the individual sources. The computational prioritization is also being made available as a Java tool called **ENDEAVOUR**.

There are several severe limitations for both versions of the strategy (the single over-representation and the modules). Firstly, the availability of high quality PSFMs is still limited. TRANSFAC contains 493 vertebrate matrices but the recent high quality non-redundant database JASPAR (based on SELEX-determined matrices) contains only 111 matrices, while the estimated number of transcription factors is ∼1850 [311]. Secondly, for noise reduction we have restricted the detection of TFBS detection to CNSs while not all experimentally determined TFBSs or modules are located in such CNSs. Thirdly, the TFBS prediction by the MotifScanner depends on the sequence length, which makes the calculation of expected frequencies for the binomial formula difficult.

Future research directions

For some of these limitations, the improvement of the computational efficiency in gene regulation bioinformatics depends on biological advances in field, like more high-quality PSFMs, more high quality and annotated microarray data sets, improved gene start annotations, more experimentally verified enhancers, etc. Regardless of that, we can think of some bioinformatics improvements *per se* as well.

A follow-up of the ModuleSearcher algorithm will have to use more than only the CNSs. One can imagine an algorithm that uses CNSs as seeds or anchors to find an initial module, then to extend the search space to the complete upstream, downstream, and intronic sequences, and improve the module iteratively. A further improvement could be a more thorough phylogenetic footprinting (PF) using this module again as a seed, and comparing the module structures (e.g., the arrangement of the binding sites therein) between multiple species, and taking the evolutionary distances into account. Such a PF-aware ModuleSearcher could also be used to find modules in a single gene represented by multiple orthologous sequences of different species, instead of being restricted to gene batteries. In this respect however, the question rises whether modular structures are conserved well enough across species to be detected. An alternative to the detailed PF, the selection of CNSs as anchors could also be extended to regions that are conserved across multiple species, instead of using only human-mouse conservation. In that case, multiple alignment algorithms like MAVID [44] (instead of AVID that we used for two species) can be used. Another possible improvement in the ModuleSearcher can be a more complex score function, for example based on hidden Markov models like in [115].

If the performance of the current, or an improved version of the ModuleSearcher proves to be high enough, a database of hypothetical modules could be constructed where biologists can find the predictions for their "pet gene" or their own hypotheses generated in the lab by comparing them with the computationally generated hypotheses in the database.

The ENDEAVOUR system for computational prioritization currently implements a simple ranking scheme where the test genes that are ranked best are considered as the best candidates to be related to the training genes. Although the ranking is accompanied with a p-value, this only indicates the probability of observing the combined ranking by chance. If one can construct a set of negative training genes (or maybe randomly selected genes from the genome), then a pattern classification approach could improve the prioritization. Support Vector Machines could be suitable to this end, by combining the kernel matrices that could be generated for each data source.

A last, more general perspective can be given regarding the analysis of gene regulatory networks. It should be feasible to construct an analysis pipeline for a certain process under study, using the methods discussed in this work and other available methods. The pipeline could start with the analysis of all available microarray data in public repositories measured in conditions relevant for the process, to select high-quality gene batteries or even to infer linkages between transcription factors. This can be followed by TFBS and module detection and *in silico* validation of the predicted GRN linkages. This way, complete subnetworks can be predicted that could be experimentally validated. Such an analysis however requires a biological research environment or at least a collaboration between bioinformatics researchers and molecular biologists.

Appendix A. Glossary

This glossary has been constructed mainly from the 2can glossary of the European Bioinformatics Institute, available at http://www.ebi.ac.uk/2can/, and from [59].

activator A protein that positively regulates transcription of a gene.

annotation A combination of comments, notations, references, and citations, either in free format or utilising a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools.

base pair A pair of nitrogenous bases (a purine and a pyrimidine), held together by hydrogen bonds, that form the core of DNA and RNA i.e the A:T, G:C and A:U interactions.

Bilateria The bilaterally symmetrical animals, including all protostomes and deuterostomes, but not sponges, cnidarians, or ctenophores.

binding site A place on cellular DNA to which a protein (such as a transcription factor) can bind. Typically, binding sites might be found in the vicinity of genes, and would be involved in activating transcription of that gene (promoter elements), in enhancing the transcription of that gene (enhancer elements), or in reducing the transcription of that gene (silencers). Note that whether the protein in fact performs these functions may depend on some condition, such as the presence of a hormone, or the tissue in which the gene is being examined. Binding sites could also be involved in the regulation of chromosome structure or of DNA replication.

chromatin The material into which DNA in the cell nucleus is packaged with proteins.

***cis*-regulatory element** A discrete region of DNA that affects transcription of a gene.

***cis*-regulatory module** See cis-regulatory element. The word module is used because of the modular organization of cis-regulatory elements.

CpG island CpG refers to a C nucleotide immediately followed by a G. The 'p' in 'CpG' refers to the phosphate group linking the two bases. Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands. It has been estimated that about half of all mammalian genes have a CpG-rich region around their 5' end. It is said that all mammalian housekeeping genes have a CpG island. Non-mammalian vertebrates have some CpG islands that are associated with genes, but the association gets equivocal in the farther taxonomic groups. Finding a CpG island upstream of predicted exons or genes is good contributory evidence.

enhancer An enhancer is a nucleotide sequence to which transcription factor(s) bind, and which increases the transcription of a gene. It is *not* part of a promoter; the basic difference being that an enhancer can be moved around anywhere in the general vicinity of the gene (within several thousand nucleotides on either side or even within an intron), and it will still function. It can even be clipped out and spliced back in backwards, and will still operate. A promoter, on the other hand, is position- and orientation-dependent. Some enhancers are ”conditional”—in other words, they enhance transcription only under certain conditions, for example in the presence of a hormone.

gene A unit of DNA that performs one function. Usually, this is equated with the production of one RNA or one protein. A gene contains coding regions, introns, untranslated regions, and control regions.

gene battery A set of target genes of a regulatory network that are co-regulated by the same set of regulators

housekeeping genes Genes that encode proteins required for basic functions required in all cells.

Metazoa Multicellular animals, including diploblasts and triploblasts.

microarray Microarrays allow snapshots to be made of expression levels for thousands of genes in a single experiment.

orthologs Homologous genes in different species that arose from a single gene in the last common ancestor of these species.

paralogs Homologous genes that are related by duplication of an ancestral gene.

parsimony Refers to a rule used to choose among possible trees, which states that the tree implying the least number of changes in character states is the best.

phylogeny The evolutionary relationships among organisms; the patterns of lineage branching produced by the true evolutionary history of the organisms being considered.

promoter The genomic sequence immediately upstream of the transcriptional start site defined by the 5’ end of an mRNA. It is this region that is presumed to bind the *trans*-acting factors required to transcribe the gene.

pseudogene The remnant of a gene that has been rendered nonfunctional through the accumulation of mutations.

regulatory evolution Evolutionary changes in gene regulation.

regulon A set of co-regulated genes (they have the same *cis*-regulatory elements).

repressor A transcription factor that negatively regulates the expression of a gene, often by binding directly to DNA sequences in a *cis*-regulatory element.

TATA-box A sequence found in the promoter (part of the 5’ flanking region) of many genes. Deletion of this site (the binding site of transcription factor TFIID) causes a marked reduction in transcription, and gives rise to heterogeneous transcription initiation sites.

transcription The process of copying DNA to produce an RNA transcript. This is the first step in the expression of any gene. The resulting RNA, if it codes for a protein, will be spliced, polyadenylated, transported to the cytoplasm, and by the process of translation will produce the desired protein molecule.

transcription factor A protein that is involved in the transcription of genes. These usually bind to DNA as part of their function (but not necessarily). A transcription factor may be general (i.e. acting on many or all genes in all tissues), or tissue-specific (i.e., present only in a particular cell type, and activating the genes restricted to that cell type). Its activity may be constitutive, or may depend on the presence of some stimulus; for example, the glucocorticoid receptor is a transcription factor that is active only when glucocorticoids are present.

Appendix B. Software and databases

Table 8.1: Software developed in this work

Tool	Description	URL	Ref.
NEURODIFF[1]	Interactive web application for the functional analysis of microarray data from *in vitro* differentiation of mouse hippocampal neurons	http://www.esat.kuleuven.ac.be/neurdiff	[77]
MotifScanner[2]	Probabilistic algorithm for the prediction of transcription factor binding sites	http://www.esat.kuleuven.ac.be/\~thijs/Work/MotifScanner.html	[296, 4]
MotifLocator[2]	Algorithm for the prediction of transcription factor binding sites with score thresholds		[296]
ModuleSearcher[1]	Algorithms (branch-and-bound and Genetic Algorithm) for the detection of new over-represented *cis*-regulatory modules	http://www.esat.kuleuven.ac.be/\~saerts/software/modulesearcher.html	[7, 6]
ModuleScanner[1]	Algorithm to detect *cis*-regulatory modules as known combinations of position weight matrices		[7]
TOUCAN[1]	Java standalone software application for *cis*-regulatory sequence analysis in metazoan genomes	http://www.esat.kuleuven.ac.be/\~saerts/software/toucan.html	[4]
GO4G[1]	Web application to find statistically over-represented Gene Ontology terms in gene sets	http://www.esat.kuleuven.ac.be/\~saerts/software/go4g.html	[67]
ENDEAVOUR[3]	Java standalone application to rank test genes according to their similarity with a set of test genes using multiple information sources	http://www.esat.kuleuven.ac.be/\~saerts/software/endeavour.php	[3]

[1] Developed fully in this work. [2] Joint work with Gert Thijs. [3] Joint work with Bert Coessens. ENDEAVOUR is still under development.

Table 8.2: Used third party and open source algorithms, tools, and libraries

Tool	Description	URL	Ref.
AVID	AVID is a global alignment algorithm optimized for large intergenic regions.	http://baboon.math.berkeley.edu/avid/	[42]
VISTA	Visualization of a global alignment as a percent identity plot.	http://www-gsd.lbl.gov/vista/index.shtml	[209]
FootPrinter	Phylogenetic footprinting algorithm without alignment. FootPrinter is accessible as a SOAP web service in TOUCAN.	http://abstract.cs.washington.edu/\~blanchem/FootPrinterWeb/FootPrinterInput2.pl	[35, 34, 36]
BioJava	The BioJava library is an open-source Java package for processing biological data.	http://www.biojava.org/	
BioPerl	The Bioperl library is a collection of open source Perl tools for bioinformatics, genomics and life science research.	http://www.bioperl.org	[271]
SOAP	SOAP (Simple Object Access Protocol) is a lightweight XML based protocol for exchange of information in a decentralized, distributed environment.	http://ws.apache.org/soap/, http://ws.apache.org/axis/	
RMI	Java RMI (Remote Method Invocation) enables the invocation of methods of remote objects from other Java virtual machines, using object serialization.	http://java.sun.com/products/jdk/rmi/	
ensj-core	Ensj is a Java API (Application Programming Interface) to interact with Ensembl databases.	http://www.ensembl.org/java/	
MySQL	MySQL is an open source database management system.	http://www.mysql.com/	

Table 8.3: Used third party and open source databases

Database	Contents	URL	Ref.
TRANSFAC	Over 600 position weight matrices (motif models) of transcription factors.	http://www.biobase.de	[323]
Ensembl	Genomic DNA sequences and annotation of several animals.	http://www.ensembl.org	[149]
SMD	SMD (Stanford Microarray database) stores raw and normalized data from microarray experiments.	http://genome-www5.stanford.edu/	[263]
SOURCE	SOURCE is a unification tool which dynamically collects and compiles data from many scientific databases about genes.	http://source.stanford.edu/	[86]
ArrayExpress	ArrayExpress is a public repository for microarray data, which is aimed at storing well annotated data in accordance with MGED recommendations.	http://www.ebi.ac.uk/arrayexpress/	[47]
GEO	GEO (Gene Expression Omnibus) is a high-throughput gene expression data repository.	http://www.ncbi.nlm.nih.gov/geo/	[96]
OMIM	OMIM (Online Mendelian Inheritance in Man) is a catalog of human genes and genetic disorders.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	[320]
LocusLink	LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci.	http://www.ncbi.nlm.nih.gov/LocusLink/	[320]
Jaspar	Jaspar is a high-quality transcription factor binding profile database.	http://jaspar.cgb.ki.se/	[252]
InterPro	InterPro is a database of protein families, domains and functional sites.	http://www.ebi.ac.uk/interpro/	[216]
GO	GO (Gene Ontology) is a controlled vocabulary produced by the GO Consortium that can be used to annotate genes and proteins of any organism.	http://www.geneontology.org	[15]

Table 8.3: Used third party and open source databases (continued)

Database	Contents	URL	Ref.
MGI	MGI (Mouse Genome Informatics) provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse.	http://www.informatics.jax.org/	[33]
Unigene	UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene	[320]
MEDLINE/-PubMed	PubMed includes over 14 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed	
DBTSS	DBTSS contains exact information of the genomic positions of the transcriptional start sites for a number of human and mouse genes.	http://elmo.ims.u-tokyo.ac.jp/dbtss/	[286, 287]
EPD	EPD (Eukaryotic Promoter Database) is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.	http://www.epd.isb-sib.ch/	[256]
Flybase	*Drosophila* genome database.	http://flybase.bio.indiana.edu/	[109]
Wormbase	*C. elegans* genome database.	http://www.wormbase.org/	[135]

Nederlandse samenvatting

Computationele detectie van cis-regulatorische modules in dierlijk DNA

Inleiding

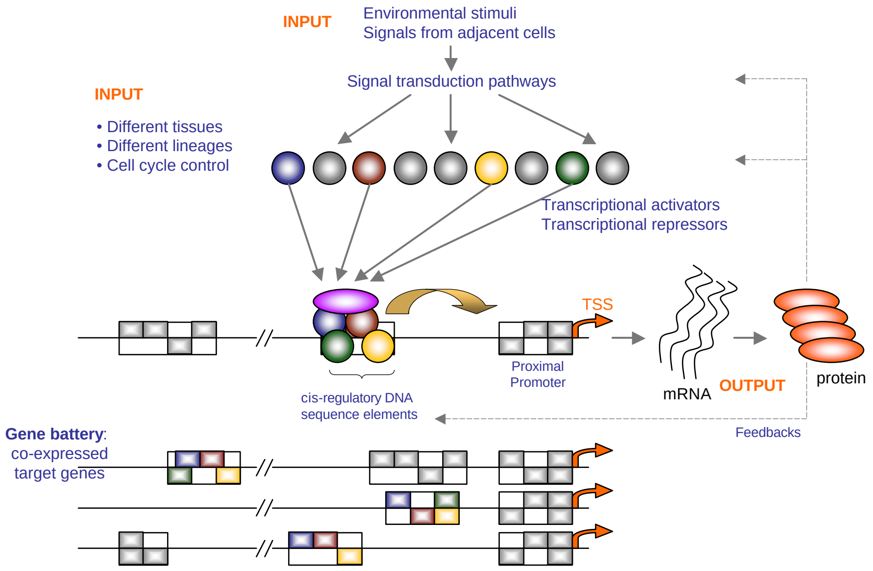
Bij de recentelijke voltooiing van verschillende genoomprojecten [173, 311, 318, 2] bleek het aantal opgetekende genen minder dan verwacht. Zo is het aantal genen in het menselijk genoom (∼25000) slechts het dubbel van het aantal genen in de fruitvlieg, en zowat 10% van de menselijke genen zijn bovendien duidelijk verwant aan bepaalde genen in vlieg en worm. Daarom is de motivatie om te begrijpen hoe genen gereguleerd worden sterker als nooit tevoren en er wordt aangenomen dat evolutie en ontwikkeling beide uitingen zijn van de erfelijke regulatorische programma's die bepalen hoe de morfologische kenmerken van elke soort eruit zien [146, 78, 59].

De rol van de bio-informatica bij de studie van genregulatie is groter geworden tijdens het afgelopen decennium, zowel door de enorme hoeveelheden sequentie- en annotatiegegevens die beschikbaar worden, als door het gebruik van "high-throughput" metingen van genexpressie met behulp van DNA microarrays waarvoor computationele analysemethoden vereist zijn.

Genregulatorische netwerken en *cis*-regulatorische modules

Bij dieren, en ook meer algemeen bij Eukaryoten, zijn diverse mechanismen in werking om genexpressie te reguleren, waaronder chromatinecondensatie, DNA-methylatie, transcriptie-initiatie, alternatieve "splicing" van RNA, mRNA-stabiliteit, translationele controles, verschillende vormen post-translationele modificatie, intracellulaire traffiek, en eiwitdegradatie [183]. Van al deze categorieën ligt de voornaamste controle bij de snelheid van transcriptie-initiatie [178].

Slechts enkele genen in een eukaryotische cel komen tot expressie op een bepaald moment. De verhouding en de samenstelling van de afgeschreven genen verandert aanzienlijk tijdens de levenscyclus, tussen celtypes, en als respons op veranderende fysiologische en omgevingsfactoren. Gegeven dat eukaryotische genomen qua grote-orde 0.5 tot 5 $\times 10^4$ genen bevatten, vereist de differentiële regulatie ervan een buitengewoon complex verzameling van specifieke fysische interacties tussen macromolecules. De vorm van de machinerie die transcriptie controleert is dat van een genregulatorisch netwerk (GRN, zie Figuur 8.1). Een GRN bepaalt de transiënte regulatorische toestanden in een cel en de batterijen van stroomafwaartse genen die ze tot expressie brengen [146, 325].


INPUT
Environmental stimuli
Signals from adjacent cells
Signal transduction pathways
INPUT
• Different tissues
• Different lineages
• Cell cycle control
Transcriptional activators
Transcriptional repressors
TSS
Proximal Promoter
cis-regulatory DNA sequence elements
mRNA
OUTPUT
protein
Feedbacks
Gene battery: co-expressed target genes

Figuur 8.1: Een imaginair generegulatorisch netwerk met de *cis*-regulatory modules als centrale elementen.

Zoals weergegeven in Figuur 2.1 zijn de centrale elementen in een GRN de *cis*-regulatorische modules (CRM of kortweg module) waarop transcriptiefactoren (TF) en co-factoren kunnen assembleren. CRM's behandelen op die manier alle informatie van de stroomopwaartse biochemische signaaltransductiebanen en door te communiceren met het basaal transcriptie-apparaat dirigeren zij de snelheid van transcriptie-initiatie. Een CRM wordt operationeel gedefinieerd als een cluster van transcriptiefactorbindingsplaatsen (TFBP) die een discreet aspect van het totale transcriptieprofiel van een gen produceert. De meest gehanteerde termen in de literatuur voor een CRM zijn enhancer en silencer. Een module bevat typisch ongeveer 6 tot 15 bindingsplaatsen en er binden 4 tot 8 TF'en [78].

Het doel van dit werk is om nieuwe methoden te ontwikkelen om groepen van genen te vinden die samen tot expressie komen, via microarray data-analyse, en om de TFBP'en en de modules te ontdekken die verantwoordelijk zijn voor deze co-expressie. Wat betreft de TFBP'en wordt er gebruik gemaakt van bestaande modelleringsmethoden. De twee voornaamste modellen voor een TFBP zijn de consensus sequentie en de positiespecifieke frequentiematrix (PSFM). Beide worden opgebouwd of getraind vertrekkende van een alignering van experimenteel bepaalde bindingsplaatsen[1]. De consensus sequentie komt goed, maar niet noodzakelijk exact, overeen met elk van de individuele plaatsen en er is een altijd een afweging tussen het aantal toegelaten afwijkingen, the ambiguteit in de consensus, en de specificiteit en sensitiviteit van de voorstelling. De voorstelling door een PSFM, die wij voornamelijk zullen hanteren, is een benadering van de bindingsenergie van een TF op een bindingsplaats [24] en wordt als volgt voorgesteld:

$$\mathbf{\Theta} = \begin{pmatrix} w_{\texttt{A},1} & w_{\texttt{A},2} & \dots & w_{\texttt{A},L} \\ w_{\texttt{C},1} & w_{\texttt{C},2} & \dots & w_{\texttt{C},L} \\ w_{\texttt{G},1} & w_{\texttt{G},2} & \dots & w_{\texttt{G},L} \\ w_{\texttt{T},1} & w_{\texttt{T},2} & \dots & w_{\texttt{T},L} \end{pmatrix}, \tag{8.1}$$

waar $w_{b,j}$ de probabiliteit is om nucleotide b te vinden op positie j in de bindingsplaats van lengte L. Om nu instanties van een dergelijk motiefmodel terug te vinden in een sequentie, wordt voor elk sequentiesegment $\mathbf{x}$ een score berekend:

$$W(\mathbf{x}) = \sum_{j=1}^{L} w_{b,j}. \tag{8.2}$$

Een algemeen toegepaste techniek is om deze segmenten $\mathbf{x}$ te weerhouden als mogelijke bindingsplaats of positieve “hit”, wanneer de genormaliseerde score groter is dan een bepaalde drempelwaarde.

Gezien de beperkte lengte van een TFBP en gezien de ontaarding van een TFBP (er is veel variatie toegelaten op de bindingssequentie), komen TFBP’en overal voor in het genoom, maar slechts een beperkt aantal voorkomens zijn functionele TFBP’en. Wij zullen via drie benaderingswijzen het aantal vals positieve TFBP-voorspellingen trachten te reduceren: ten eerste door een meer robuuste voorspelling van TFBP’en met PSFM’en; ten tweede door het verkleinen van de zoekruimte, hetgeen kan op twee mogelijke manieren, namelijk door “phylogenetic footprinting” of het in rekening brengen van evolutionaire conservering en door het gebruik van “co-expressie”, gemeten met DNA microarrays; en ten derde door het zoeken naar combinaties van TFBP’en als modules i.p.v. naar enkelvoudige TFBP’en.

Microarray data-analyse: een casus in de neurobiologie

Hier beschrijven we enkele nieuwe en reeds bestaande technieken om microarray expressiegegevens te analyseren, gebruik makende van een casus betreffende de genprofilering in hippocampale neuronen in muis tijdens differentiatie [77]. Deze

---

[1] Een voorbeeld van een *in vitro* methode om TFBP’en te bepalen is SELEX (systematische evolutie van liganden door exponentiële verrijking).

analyse werd uitgevoerd in samenwerking met het Laboratorium voor Neuronale Celbiologie van de KULeuven o.l.v. Professor B. De Strooper. Van 17-dagen oude muisembryo's werden neuronen uit de hippocampus in cultuur gebracht en onder bepaalde condities tot differentiatie gebracht. Na 7h, 18h, 33h, 72h, 8 dagen en 12 dagen werd mRNA uit de cellen geëxtraheerd en telkens tweemaal op vijf microarray "slides" gehybridiseerd, waarbij mRNA van de volledige hersenen als controle werd gebruikt.

De logaritmen van de ruwe expressie-ratio's (test/referentie) werden eerst voorbehandeld om systematische fouten weg te filteren. In het bijzonder werd een LOWESS fit gebruikt om de data te corrigeren voor sterk optredende kleurstof-specifieke fouten. Vervolgens werd een filter ontworpen op basis van een ANOVA procedure om de genen te selecteren die consistent hetzelfde tijdsprofiel vertoonden voor de vier herhaalde metingen[2]. Dit resulteerde in minstens 2314 genen die een verandering in expressie vertonen met een statistische betrouwbaarheid (p-waarde<0.01).

De gefilterde en genormaliseerde profielen werden vervolgens onderworpen aan een clusteranalyse (zie Figuur 3.5). De duidelijkste trends in genexpressie tijdens neuronale differentiatie zijn gestadige "up"- en "down"-regulatie. Dit patroon is zichtbaar in de globale analyse, maar ook binnen de meeste functionele groepen (bv. de synaptische vesikelcyclus), hetgeen resulteert in het vervangen van vroege genen door late genen met (schijnbaar) gelijkaardige functies. De resultaten van een "K-means" clustering met K=20 werden door een bioloog geïnterpreteerd en als volgt samengevat. Tijdens een eerste fase van de cultuur is er een hoog expressieniveau van genen voor DNA- en proteïnesynthese, die dan geleidelijk minderen in expressie. De latere differentiatiestadia worden gekarakteriseerd door een sterke vooruitgang van de systemen voor proteïnetransport en energie-ontwikkeling, en het aanzetten van specifieke neuronale functies, zoals de synaptische vesikelcyclus. Voor deze laatste functie werden alle betrokken gekende genen die in hun expressie veranderen bekeken en vergeleken met bestaande literatuur. De creatie van dergelijke functionele groepen van genen is arbeidsintensief en vereist een uitgebreide biologische kennis. Om gengroepen van andere functies te onderzoeken werd een webapplicatie ontwikkeld (NEURODIFF) die toelaat om eerst een groep van genen te selecteren o.b.v. hun functie, hetgeen wordt bewerkstelligd door "Gene Ontology" associaties, en om vervolgens deze gengroep verder op te splitsen d.m.v. een hiërarchische clustering van hun expressieprofielen. Hieruit bleek inderdaad dat de gestadige stijging en daling in expressie ook binnen de functionele groepen van kracht was. De NEURODIFF toepassing is vrij beschikbaar voor vorsers van academische instellingen. Dergelijke groepen van genen die een gelijkaardig expressieprofiel vertonen én bij hetzelfde biologisch proces betrokken zijn, worden mogelijk gecontroleerd door dezelfde groep transcriptiefactoren en zouden bijgevolg een aantal transcriptiefactorbindingsplaatsen gemeenschappelijk kunnen hebben in proximale of distale modules. Het onderzoek naar modules in deze groepen is

---

[2] Op elke slide staat tweemaal dezelfde kloon en het hele experiment werd tweemaal herhaald.

momenteel aan de gang [76].

Het systeem van *in vitro* differentiatie werd reeds veelvuldig gebruikt om verschillende aspecten van neuronale differentiatie te bestuderen. Om na te gaan of de genetische programma's die *in vitro* en *in vivo* worden aangewend dezelfde zijn, werden de hier bekomen genexpressieprofielen vergeleken met deze bekomen tijdens *in vitro* differentiatie, zoals gemeten door Mody et al. [214]. Voor de laatste werden echter verschillende tijdspunten als voor de eerste gebruikt. Daarom werd de optimale overeenkomst tussen de experimenten bepaald via "time warping" (letterlijk "het doen krommen van de tijd"), waarbij het aligneren van de tijdsprofielen kan vergeleken worden met sequentiealignering. De correlatie tussen de expressieprofielen van de genen die in beide experimenten werden gemeten was gemiddeld 0.646 en de mediaan was 0.787. Deze hoge correlatie leidde tot de volgende conclusies: (1) hetzelfde genetisch programma is actief tijdens *in vitro* en *in vivo* differentiatie van hippocampale neuronen in de muis en dus is het *in vitro* systeem een goed model om genexpressie tijdens differentiatie te bestuderen; (2) aangezien het *in vivo* experiment gebeurde met Affymetrix oligo-DNA-chips en het *in vitro* experiment met cDNA microarrays blijkt dat de resultaten van beide platformen vergelijkbaar kunnen zijn, mits een goede voorbehandeling en filtering van de data. In een review [215] zijn wij verder ingegaan op het vergelijken van microarray data en op de mogelijkheden van microarray "repositories".

Detectie van transcriptiefactorbindingsplaatsen in genen bij dieren

De casus rond genexpressie-analyse illustreerde dat het mogelijk is om met behulp van de microarray technologie clusters van genen te vinden die een gelijkaardig expressiepatroon vertonen tijdens een proces onder studie. In deze sectie introduceren we enkele methoden om TFBP'en te detecteren in de regulatorische sequenties van de genen in zo'n cluster die mogelijk verantwoordelijk zijn voor de co-expressie. Aangezien het testen van nieuwe methoden best gebeurt met gegevens waarvoor de uitkomst op voorhand geweten is, of waarvoor de uitkomst makkelijk geïnterpreteerd kan worden, worden onze methoden gevalideerd m.b.v. "benchmark" datasets.

Vooreerst werd er aandacht besteed aan het verkleinen van de zoekruimte waarin gezocht wordt naar TFBP'en. De DNA sequenties waarin regulatorische elementen kunnen liggen, spreiden zich bij hogere eukaryoten namelijk uit tot tientallen kilobasen stroomopwaarts van een gen, stroomafwaarts van een gen, en in de grote intronische gebieden van een gen. Er werden twee manieren gebruikt om de zoekruimte te verkleinen: (1) het gebruik van groepen van genen die samen tot expressie komen, zodanig dat een statistische over-representatie van een TFBP in een set berekend kan worden; en (2) het gebruik van "phylogenetische footprinting" door enkel deze DNA sequenties te onderzoeken die geconserveerd zijn tussen twee gerelateerde organismen. In deze studie werden daarvoor orthologe genenparen van mens en muis gebruikt. Het tijdstip van evolutionaire divergentie van mens en muis (de speciatie) is zodanig dat blokken van geconserveerde sequentie makkelijk gedetecteerd kunnen worden. Bovendien hebben deze een relatief grote kans hebben om een regulatorische functie te herbergen [329]. De gespecialiseerde algoritmen die vereist zijn voor de alignering van grote genomische sequenties werden beschikbaar tijdens dit werk (hier werd AVID [43] gebruikt). Aangezien echter algemeen aanvaard wordt dat er in de proximale promotergebieden altijd functionele TFBP'en aanwezig kunnen zijn, onafgezien van een mogelijke evolutionaire conservering, werd tevens een manier bestudeerd om dergelijke proximale gebieden van voldoende kwaliteit te bekomen. Na een analyse van de nucleotidesamenstelling rond de startplaats van menselijke genen zoals die in de Ensembl databank geannoteerd is (zie ook verder), werd geconcludeerd dat deze startplaatsen in voldoende mate overeenstemmen met de werkelijke transcriptiestartplaatsen. Bijgevolg kan het onmiddellijk gebied stroomopwaarts ervan (bv. 400 bp) als proximale promoter gebruikt worden. Deze aanpak heeft als voordeel dat er geen promoter-predictie algoritmen gebruikt dienen te worden. Van dergelijke algoritmen is de performantie vaak immers onvoldoende (bv. tot 70%), en het gebruik ervan afhangt af van het organisme (promoterpredictie bij menselijke genen verschilt bijvoorbeeld van promoterpredictie bij fruitvlieggenen).

Nadat de gewenste sequenties geselecteerd zijn voor een groep van genen, wordt elke individuele sequentie "gescoord" met alle PSFM's uit de TRANSFAC databank [323]. De reeds gekende scoringsalgoritmen werken alle door elk subsegment in de sequentie als een bindingplaats (of "hit") te beschouwen waarvoor de genormaliseerde som van de logaritmen van alle probabiliteiten overeenkomstig de PSFM boven een ingestelde drempelwaarde ligt (bv. 0.75). Wij hebben een nieuwe scoringsmethode ontwikkeld en getest die twee nieuwe elementen introduceert. Het eerste is het gebruik van een hoger-orde achtergrondmodel waarmee het subsegment ook gescoord wordt en waarmee de score volgens de PSFM vergeleken wordt. Dit zorgt voor een bevoordeling van motiefinstanties die 'uit de achtergrond springen':

$$W(\mathbf{x}) = \log\left(\frac{P(\mathbf{x}|\boldsymbol{\Theta})}{P(\mathbf{x}|S,\mathcal{B}_m)}\right) = \sum_{j=1}^{W}[\log(\boldsymbol{\theta}_j^{b_j}) - \log(P(b_j|S,\mathcal{B}_m))]. \tag{8.3}$$

De eerste term in deze formule beschrijft de kans dat een segment een instantie is van een motiefmodel $\boldsymbol{\theta}_.$ De tweede term beschrijft de kans dat het segment overeenkomt met de background. Ten tweede worden niet alle instanties boven de drempelwaarde als positief aanschouwd, maar wordt het verwachte aantal instanties berekend in een probabilistisch model dat vergelijkbaar is met het schatten van het aantal instanties tijdens het zoeken naar nieuwe motieven met "Gibbs sampling" [298]. De drempelwaarde-parameter is dan vervangen door een parameter die de kans weergeeft dat er een instantie in de sequentie aanwezig zal zijn (de 'prior', bijvoorbeeld 0.2). Deze prior wordt in de formule van Bayes gebruikt waarmee het aantal instanties wordt afgeleid. Het resulterende algoritme is de MotifScanner.

Om de statistische over-representatie te berekenen voor alle PSFM's die bij het scoren gebruikt werden, wordt een binomiale analyse uitgevoerd. De verwachte frequentie aan instanties van een PSFM wordt hierin vergeleken met de geobserveerde frequentie en resulteert in een p-waarde voor elke PSFM. Na een correctie die dient te gebeuren omdat multipele tests worden uitgevoerd, kunnen die PSFM's als over-gerepresenteerd beschouwd worden, die een voldoende hoge significantie vertonen. Om de verwachte frequenties van elke PSFM te berekenen werden alle mens-muis geconserveerde gebieden in 10 kilobasen stroomopwaarts van de genstart gescoord.

Deze procedure van sequentieophaling uit Ensembl, mens-muis sequentiealignering, PSFM scoring, en binomiale analyse werd geïmplementeerd in een software tool TOUCAN, waarbij de PSFM scoring niet lokaal gebeurt, maar op de ESAT servers. De communicatie tussen de TOUCAN-klant en de ESAT servers gebeurt d.m.v. een XML-gebaseerd protocol SOAP genaamd ("Simple Object Access Protocol"). TOUCAN kan worden gedownload van onze website en is vrij te gebruiken door vorsers van academische instellingen.

TOUCAN en de voorgestelde strategie werden gevalideerd door de significant over-gerepresenteerde bindingsplaatsen voor een set van spierspecifieke en een set van leverspecifieke genen te bepalen. De resultaten kwamen goed overeen met de verwachte resultaten die werden afgeleid uit twee voorheen verschenen studies [168, 316] waarin de experimenteel bepaalde TFBP'en werden samengevat en gemodelleerd. Deze studie toont aan dat functionele TFBP'en kunnen gevonden worden in co-gereguleerde menselijke genen, hetgeen door de grote intergenische gebieden voorheen niet in op een "high-throughput" computationele manier kon worden uitgevoerd. Wij postuleren dat deze resultaten erop wijzen dat ook clusters van genen die enkel *vermoedelijk* co-gereguleerd zijn (bijvoorbeeld clusters van genen bekomen met microarray data) betrouwbare resultaten kunnen opleveren. Dit geldt des te meer indien er daarenboven bijkomende aanwijzingen zijn voor co-regulatie (bv. indien de genen naast een gelijkaardig expressieprofiel ook in hetzelfde of een gelijkaardige proces actief zijn). Er dient verder opgemerkt dat de keuze van de sequentieset om de verwachte frequenties te berekenen een significante invloed heeft op de resultaten hetgeen een negatief effect heeft op de robuustheid van de aanpak (een pragmatische oplossing hiervoor wordt gegeven in de volgende sectie). Tevens is de methode afhankelijk van beschikbare PSFM's in databanken als TRANSFAC of JASPAR [252], waarvan het aantal en de kwaliteit tot op heden nog niet optimaal zijn. Een laatste beperking van de aanpak is de restrictie van de zoekruimte tot geconserveerde gebieden. Door deze beperking neemt het aantal vals positieve voorspellingen inderdaad gunstig af, maar het aantal vals negatieven neemt toe. De beperkte biologische kennis van distale regulatorische gebieden en hun conservering maken dat een schatting van deze percentages momenteel nog moeilijk is.

Ten slotte werd TOUCAN gebruikt voor regulatorische sequentie-analyses tijdens twee gezamenlijke projecten met moleculaire biologen van het Centrum Menselijke Erfelijkheid betreffende de regulatorische analyse van mogelijke

TCF-3-β-catenine doelgenen [83] en van het gen *Atonal* in *Drosophila melanogaster* m.b.v. phylogenetische footprinting.

Detectie van *cis*-regulatorische modules

Verscheidene experimentele studies hebben uitgewezen dat de transcriptionele regulatie van een gen door een *combinatie* van transcriptiefactoren eerder regel is dan uitzondering. Dit gegeven kan worden aangewend bij de voorspelling van TFBP'en door enkel deze TFBP'en te beschouwen die in gewenste combinatie met andere TFBP'en voorkomen binnen een beperkt DNA gebied (bv. 200-500 bp). In plaats van de statistische over-representatie van elke individuele PSFM afzonderlijk te bekijken, werd een algoritme ontwikkeld om de beste combinatie van PSFM's te vinden. Een dergelijke module omvat PSFM's waarvoor de instanties vaak samen voorkomen binnen een DNA-venster, en dit in zoveel mogelijk sequenties van een set. Een score-functie wordt gehanteerd die voor een PSFM-combinatie het logaritme van de MotifScanner scores van elke PSFM-instantie binnen een DNA-venster sommeert en die vervolgens de scores van de beste scorende DNA-vensters van elke sequentie sommeert. Het best scorende wordt als volgt berekend:

$$\mathcal{S}_{\mathcal{M}}(\mathbf{set}) = \sum_{i=1}^{n} \mathcal{S}_{\mathcal{M}}(seq) = \sum_{i=1}^{n} \max_{T} p(\mathbf{m}) \times \sum_{\mathbf{x}\in\mathbf{m}} W(\mathbf{x}), \tag{8.4}$$

met $W(\mathbf{x})$ de PSFM score zoals beschreven in 8.3, $\mathbf{m}$ een module (een verzameling van $\mathbf{x}$'en), $\mathbf{x}$ een instantie van een $\boldsymbol{\Theta}$ en T een verzameling van alle *geldige* $\mathbf{m}$'s, of met andere woorden alle mogelijke instanties van $\mathcal{M}$ in een sequentie *seq*.

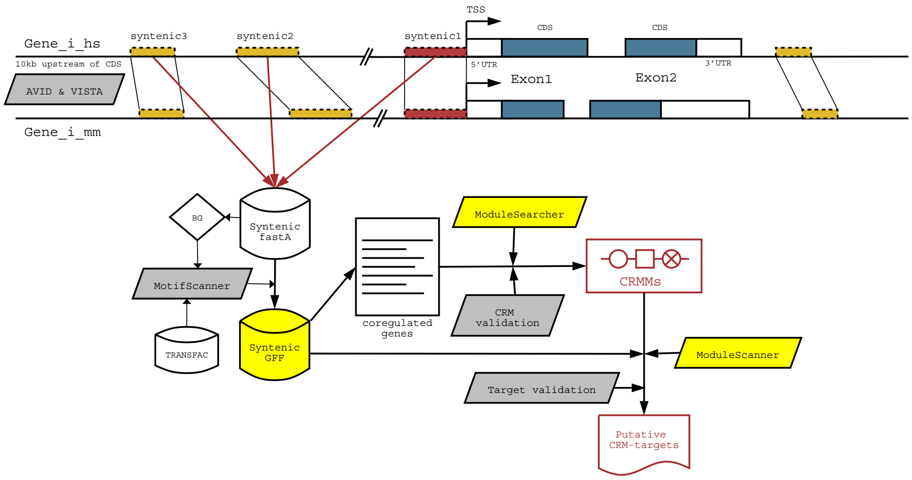
Aangezien nu het aantal PSFM's van vertebraten in TRANSFAC 493 is (voor versie 7.3), is het computationeel niet mogelijk om voor alle combinaties van bijvoorbeeld 5 PSFM's de score te berekenen om zo de meest optimale te selecteren. Er werden twee efficiënte zoekalgoritmen geïmplementeerd om de zoekruimte te doorzoeken. Het eerste is gebaseerd op A*, waarbij een boom doorzocht wordt waarin elke knoop een deel is van een mogelijke oplossing. De "root" van de boom is de lege module, en op elk niveau wordt één PSFM toegevoegd. Met behulp van een heuristiek wordt de boom op een intelligente manier doorzocht zodat niet alle knopen moeten worden afgegaan. A* vindt gegarandeerd het optimale pad in de boom, en dus de optimale module [136]. Met de "A* ModuleSearcher" werd de voorgestelde aanpak gevalideerd.

De ModuleSearcher werd getest op semi-artificiële datasets: random sequenties werden gegenereerd door "sampling" van een 3de orde achtergrondmodel (gemaakt van alle mens-muis geconserveerde gebieden) en daarin werden op willekeurige posities niet-overlappende instanties geïmplanteerd door "sampling" van een PSFM. Het scoren van deze sequenties met alle PSFM's uit TRANSFAC resulteert in een enorm aantal vals positieve TFBP-predicties (er zijn namelijk veel "hits" van PSFM op plaatsen waar geen instantie werd geïmplanteerd).

Het algoritme was echter in staat om 80 % (4/5) tot 100 % (5/5) van de geïmplanteerde elementen correct terug te vinden als module.

Als biologische test werd een groep van genen samengesteld die samen tot expressie komen tijdens de celcyclus. Hiervoor werd gebruik gemaakt van de SOURCE [86] webapplicatie die toelaat naar genen te zoeken in een microarray experiment die een gelijkaardig expressieprofiel vertonen, als een "query". Als query gebruikten wij *CCNB2* (cycline B2) en als microarray experiment de metingen tijdens de celcyclus van HeLa cellen van Whitfield et al. [322]. Voor die genen waarvoor een muisortholoog gekend is in Ensembl, werd de 10 kb stroomopwaartse sequentie afgehaald en de mens-muis orthologe gebieden werder gealigneerd met AVID. In de geconserveerde niet-coderende sequenties (CNS) werd naar de optimale combinatie van 4 TF'en gezocht binnen DNA vensters van 200 bp. De beste module was [CEBPA-STAF-NFY-TCF4].

Om dit resultaat te valideren, werd een strategie ontwikkeld die gebaseerd is op de hypothese dat een "query" op het hele menselijke genoom (of minstens op alle CNS'en) met een functionele module, een aantal van de doelgenen van deze module moet kunnen terugvinden (i.e., een goede sensitiviteit) zonder te veel vals positieve "hits" (i.e., een goede specificiteit). De ModuleScanner methode werd ontwikkeld en is gebaseerd op de scorefunctie van de ModuleSearcher. Indien de CNS'en met de hoogste score voor de geteste module in de buurt liggen van genen die alle tot hetzelfde proces behoren, dan is de module plausibel. Deze aanpak werd getest door de IFN-β-enhancer [IRF-HMGIY-NFKB] als query te gebruiken. Als alle CNS'en gerangschikt worden volgens score, staat het IFN-β gen op nummer 4 (van meer dan 10.000 genen). Enkele andere genen uit de top 10 zijn duidelijk verwant met het de anti-virale functie van IFN-β. Een meer systematische aanpak werd vervolgens uitgewerkt om manuele opzoekingen in de literatuur en subjectieve beslissingen te vermijden. Zoals er voor de analyse van genexpressiegegevens gebruik gemaakt werd van Gene Ontology associaties om genen met een bepaalde functie te groeperen, zullen wij hier dezelfde bron van informatie gebruiken om de functionele coherentie van een gengroep als volgt te bepalen. Voor elke geannoteerde term van een gen worden alle termen bij de genannotatie gevoegd die op een van de mogelijke paden liggen naar de 'root' in de "directed acyclic graph" (DAG). Vervolgens wordt voor elke term in deze uitgebreide annotatie de frequentie berekend in de hele genset. Met de binomiale formule wordt dan een p-waarde berekend die aangeeft hoe waarschijnlijk deze geobserveerde frequentie is t.o.v. de verwachte frequentie. Deze laatste kan geschat worden door de frequentie van elke term te berekenen in een set van alle genen van het genoom. Voor de IFN-β-enhancer zijn er enkele termen significant over-gerepresenteerd in de top 10 scorende genen die met de ModuleScanner bekomen werden, zoals onder meer "apoptosis", "cytokine activity" en "innate immune response". Voor de nieuwe [CEBPA-STAF-NFY-TCF4] module die in de *CCNB2*-set werd gevonden waren GO-termen als "mitosis", "nuclear division" en "cell proliferation" over-gerepresenteerd. De nieuwe module is dus in staat om celcyclus-specifieke genen uit het genoom te selecteren en is daarom potentieel ook functioneel. Het systeem om modules te zoeken en te valideren is schematisch voorgesteld in Figuur 8.2.



Figuur 8.2: Overzicht van het systeem om *cis*-regulatorische modules te detecteren. Alle geconserveerde niet-coderende sequenties (CNS) vanaf 100 bp, gevonden door globale alignering van alle mens-muis orthologe genparen, zijn opgeslagen. Alle CNS'en werden gescoord met alle beschikbare PSFM'en voor vertebraten uit TRANSFAC, waarvan het resultaat als GFF wordt bewaard. Voor een te bestuderen genset worden de relevante GFF rijen uit de GFF-databank opgehaald, en kunnen gebruikt worden als input voor de ModuleSearcher. Deze vindt het beste module-model in de set, samen met de instanties van het model op alle sequenties. Het model kan tevens gebruikt worden om mogelijke doelgenen te vinden in de GFF-databank m.b.v. de ModuleScanner.

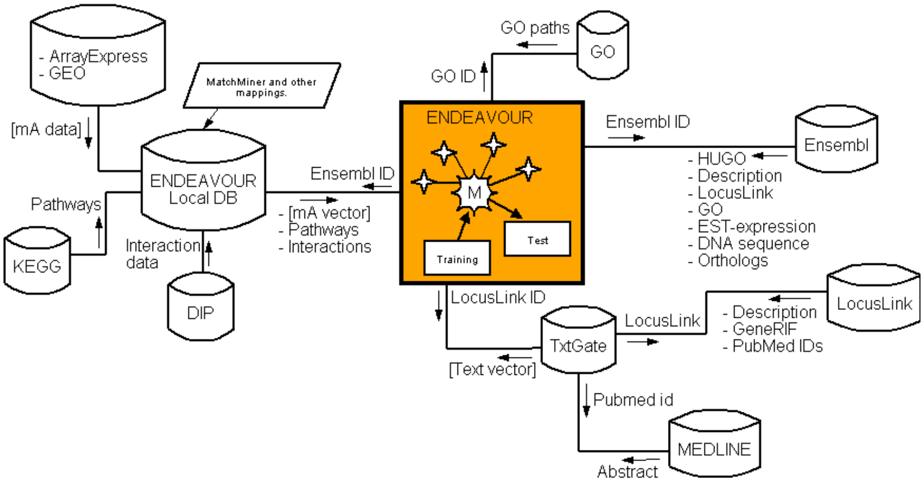
Aangezien de A* implementatie traag kan zijn voor grote sequentiesets of grote modules (bv. meer dan 5 elementen), werd een alternatieve zoekmethode ontworpen die gebaseerd is op Genetische Algoritmen (GA). Hierbij wordt vertrokken van een populatie van random gegenereerde modules die volgens dezelfde scorefunctie als in A* gerangschikt worden. De beste "survivors" worden onderling gekruist (totdat dezelfde grote van populatie opnieuw bereikt is) en sommige PSFM's kunnen muteren tot een andere PSFM. Dit proces wordt een vooraf opgegeven aantal iteraties herhaald (i.e., de generaties). Deze GA-ModuleSearcher werd gevalideerd door de resultaten op de *CCNB2*-genset voor verschillende parameterinstellingen te vergelijken met de uitkomst van de A*-ModuleSearcher. Indien GA twee tot driemaal herhaald wordt met 100 generaties, is de uitkomst dezelfde als voor A* en de winst aan snelheid is significant (enkele minuten voor GA i.p.v. uren of zelfs dagen voor A*).

Als conclusie kunnen we stellen dat de ModuleSearcher samen met de ModuleScanner en de statistiek voor GO-overrepresentatie—alle in dit werk ontwikkeld—, een valabele methode is om nieuwe *cis*-regulatorische modules te ontdekken in dierlijke genomen, hetgeen tot voorheen niet mogelijk was. Er zijn diverse methoden gepubliceerd die zoals de ModuleScanner naar genomische instanties zoeken van gekende combinaties van TF'en (gebruik makende van de respectievelijke PSFM's). Het CREME algoritme [262] heeft hetzelfde doel als de ModuleSearcher maar gebruikt andere technieken (o.m. een hashing algoritme om combinaties van PSFM's te genereren) en werd onafhankelijk van en gelijktijdig met de ModuleSearcher [7] gepubliceerd.

Data integratie voor de validatie van modules en doelgenen

Het testen van de functionele coherentie van potentiële doelgenen die door de ModuleScanner gevonden worden kan worden aangewend om deze doelgenen te karakteriseren en om de query-module te valideren. Naast de hierboven aangewende Gene Ontology associaties van de doelgenen kan echter ook andere data gebruikt worden. Wij hebben een aantal genomische informatiebronnen geïntegreerd in een software systeem ENDEAVOUR die voor elk data type, net als voor GO, de karakteristieken van een genset (bv. de top N scorende genen uit ModuleScanner) weergeeft. In een tweede stap stellen wij een methode voor om de doelgenen te rangschikken volgens hun overeenkomst met de co-gereguleerde genset waarin de query-module werd ontdekt met de ModuleSearcher. Dit is nuttig in een moleculair biologische omgeving waarin hypothetische doelgenen experimenteel gevalideerd worden. De scores van de ModuleScanner verschillen namelijk weinig en er wordt aangenomen dat de genen waarvoor volgens externe informatie geweten is dat zij betrokken zijn bij het bestudeerde proces het meeste kans hebben om ook *in vivo* werkelijke doelgenen van de module te zijn.

Er worden twee data types gebruikt die genexpressie uitdrukken, namelijk microarray genexpressiedata en EST-gebaseerde expressie data. Voor de EST-data wordt dezelfde techniek gehanteerd met de binomiale statistiek zoals hierboven beschreven voor GO. Zo kan bijvoorbeeld de anatomische plaats "Nervous → brain → diencephalon → hypothalamus" over-gerepresenteerd zijn in een genset. Microarray data (meerdere datasets zijn geïntegreerd) worden momenteel niet aangewend om de expressionele coherentie te meten, maar enkel voor de sortering van de doelgenen (zie verder). Andere data types die ook met de binomiale statistiek worden behandeld zijn "KEGG pathways" die aangeven welke pathway over-gerepresenteerd is in een genset, en InterPro data die aangeven welk proteïne-domein over-gerepresenteerd is (bv. "DNA-binding" kan duiden op een set van transcriptiefactoren). Tenslotte wordt tekstuele data gebruikt die met "text mining" technieken [121] geëxtraheerd wordt uit teksten die de functie van een gen beschrijven. Deze teksten zijn o.a. een omschrijving van een gen en korte "GeneRIF" zinnen uit de LocusLink databank, en enkele "abstracts" van artikels die het gen in kwestie behandelen (hiervoor worden de PubMed ID's gebruikt die in een LocusLink record van gen terug te vinden zijn). De tekstuele bron resulteert in een aantal (bv. 20) "keywords" die over-gerepresenteerd zijn in de beschrijvende teksten van een genset. Een IT-overzicht van de data-integratie wordt voorgesteld in Figuur 8.3.



Figuur 8.3: Informatie-technologisch overzicht van ENDEAVOUR. De M binnen het gekleurde vierkant is het model dat getraind wordt, en is verbonden met alle informatiesubmodellen (ISM), voorgesteld als sterren. Voor elke link met een externe of lokale databank zijn de communicatiedetails weergegeven. Bij het trainen van een ISM wordt voor elk gen in de trainingset de beschikbare data uit de relevante databanken opgevraagd en verwerkt. Hetzelfde gebeurt voor elk gen van de testset dat gescoord wordt met elk getraind ISM. De databronnen die rechts van het centrale vierkant zijn weergegeven worden direct vanuit ENDEAVOUR aangesproken via MySQL queries of via SOAP services. De databronnen die links van het centrale vierkant zijn weergegeven werden off-line voorbehandeld en beschikbaar gemaakt in een MySQL databank.

Om de ModuleScanner doelgenen te rangschikken worden de bovenvermelde samenvattingen van alle gehanteerde data types als een *submodel* beschouwd en alle submodellen samen vormen een *model* van een *trainingset* (een trainingset kan bijvoorbeeld de originele co-gereguleerde genset zijn waarin een module werd gevonden). Het *scoren* van een *testset* (bijvoorbeeld de top 200 doelgenen uit de ModuleScanner) gebeurt voor elke submodel afzonderlijk. Voor vector-gebaseerde data types (microarray data en tekstuele data) wordt de Pearson correlatie (i.e., de cosinus van de hoek tussen twee vectoren) berekend tussen elk testgen en het gemiddelde van de trainingsgenen. De testgenen worden gesorteerd volgens die correlatie. Voor niet-vector-gebaseerde data types (de overige, namelijk GO, EST, KEGG, InterPro) worden de p-waarden van deze attributen in de trainingset (zoals zij door de binomiale analyse bepaald werden) die voor het testgen relevant zijn gecombineerd tot een nieuwe p-waarde m.b.v. Fisher's Chi-square methode.

De testgenen kunnen op n mogelijke manieren worden gerangschikt, waarbij n het aantal gebruikte submodellen voorstelt (n kan groter zijn dan het aantal data types; voor microarray data bijvoorbeeld wordt voor elke data set een apart submodel getraind). De significantie van alle rangschikkingen kan berekend worden met "order statistics":

$$P(r_1, r_2 ..., r_n) = n! \int_0^{r_1} \int_0^{r_2} \dots \int_0^{r_n} ds_1 ds_2 ... ds_n \tag{8.5}$$

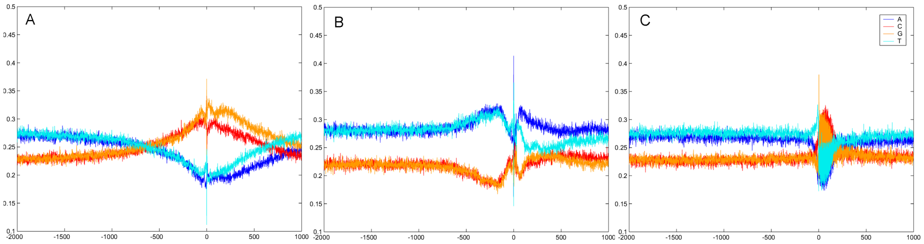
Genen met een p-waarde kleiner dan bijvoorbeeld (0.05 × aantal geteste genen) kunnen als gelijkaardig met de trainingsgenen beschouwd worden. Het systeem en de orde-statistiek werden getest d.m.v. een “cross-validatie” en de sortering van potentiële doelgenen van een *cis*-regulatorische module werd getest voor een gekende leverspecifieke en een nieuwe celcyclusspecifieke module. Deze tests tonen aan dat het voorgestelde systeem biologisch zinvolle resultaten oplevert betreffende de computationele prioritisering van de mogelijke doelgenen van een combinatie van transcriptiefactoren.

Uitgebreide analyse van de base-samenstelling rond de transcriptiestartplaats in Metazoa

De Ensembl databank is essentieel geweest voor het onderzoek dat in dit werk beschreven is. In TOUCAN worden genomische sequenties stroomopwaarts van een gen geselecteerd en opgehaald uit Ensembl, gebaseerd op de annotatie van de startplaats van het gen (i.e., de start van Exon 1). Deze annotatie is afgeleid van een mapping van beschikbare cDNA’s op de genomische sequentie. Indien deze cDNA’s volledig zijn qua lengte, dan komt de startplaats overeen met de werkelijke transcriptiestartplaats (TSP) van het gen. De algemene kennis van de TSP in hogere eukaryoten (met mogelijks lange 5’UTR’s) is nog steeds beperkt, onder meer doordat er niet één bepaald sequentiesignaal aanwezig is rond de TSP van alle genen. Enerzijds om de kwaliteit van de startplaatsannotaties te verifiëren (om ze te gebruiken voor de identificatie van proximale promoters in TOUCAN), en anderzijds om een vergelijking te treffen tussen verschillende dierlijke klassen, werd de nucleotidesamenstelling van de sequentie rond de TSP geanalyseerd.

Voor elk organisme waarvoor de genomische sequentie beschikbaar is in Ensembl werd voor 5000 random geselecteerde genen 3000 bp sequentie gedownload, 2000 bp stroomopwaarts en 1000 bp stroomafwaarts van de TSP. De karakteristieken van zoogdieren (gerepresenteerd door mens), andere vertebraten (gerepresenteerd door de Japanse Fugu of Kogelvis *Fugu rubripes*) en invertebraten (gerepresenteerd door de fruitvlieg *Drosophila melanogaster*) worden besproken en in het licht van bestaande literatuur geplaatst. Om de base-samenstelling visueel voor te stellen wordt op elke positie het gemiddeld voorkomen berekend van A, C, G, en T en uitgezet t.o.v. de positie van de TSP (zie Figuur 8.4).

Uit Figuur 8.4 blijkt dat voor alle drie de base-samenstelling drastisch verandert rond de TSP. Laten we deze verandering noteren als ΔWS = [(A+T)-(G+C)]/(A+T+G+C). Bij zoogdieren verandert ΔWS van $\sim$10% in de achtergrond (bv. op -2000) tot $\sim$-20% op de TSP. Voor andere vertebraten zoals fugu is de vorm van ΔWS gelijkaardig, maar is veel minder uitgespreid. Dit kan een gevolg zijn van het feit dat de 5’UTR in fugu onbestaande of zeer kort zijn,



Figuur 8.4: Nucleotide frequencies around the annotated gene start in Ensembl, calculated from 5000 randomly selected genes in human (A), Drosophila (B), and fugu (C).

waardoor de overheersende periodische samenstelling van de coderende sequentie snel de overhand haalt. Invertebraten, en in het bijzonder de fruitvlieg[3], vertonen ook een significante verandering van ΔWS, gaande van ∼12% in de achtergrond tot ∼26% aan de TSP. Dit is een tegengestelde verandering als bij vertebraten, en het maximale verschil wordt niet op de TSP zelf bereikt, maar ongeveer 150 bp ervoor.

Om een beeld te krijgen van de fenomenen die dergelijke fluctuaties in basesamenstelling over grote afstanden veroorzaken, werden enkele analyses uitgevoerd. Daaruit is gebleken dat de ΔWS veranderingen bij zoogdieren praktisch volledig veroorzaakt worden door het gehalte aan CpG dinucleotiden. Door de DNA methylatie in deze species (methylatie van cytosine, die dan vaak deamineert tot thymine) zijn CpG doubletten onder-gerepresenteerd in het genoom. Echter, rond de TSP is dit voor vele genen (vnl. genen die makkelijk of veel tot expressie moeten komen) niet het geval, hetgeen de uitgesproken profielen veroorzaakt. Dit fenomeen staat bekend als CpG eilanden. Uit gelijkaardige tests in fugu bleek dat er mogelijk “primordiale” CpG eilanden aanwezig zijn, dus veel minder uitgesproken. In fruitvlieg is er geen DNA methylatie en zijn er geen CpG eilanden. De profielen in fruitvlieg zouden mogelijks te wijten kunnen zijn aan het voorkomen van A/T-rijke transcriptiefactorbindingsplaatsen in de nabijheid van de TSP.

Conclusies

In dit werk is aangetoond dat het mogelijk is om betekenisvolle resultaten te bekomen in de regulatorische sequentie-analyse in dierlijke genomen, voornamelijk door de integratie van meerdere computationele methoden en meerdere data types. Het beschreven werk heeft een brede omvang en er worden verschillende aspecten behandeld voor de analyse van genregulatorische netwerken

---

[3] Op de profielen van mug waren werd veel ruis waargenomen, waarschijnlijk door gebrekkige annotatie, en die van de wormen *Caenorhabditis elegans* en *C. briggsae* worden niet besproken omdat het “trans-splicing” fenomeen in deze species de interpretatie bemoeilijkt [37, 312].

(GRN), sommige in meer detail dan andere. Ondanks het feit dat de principes van GRN's dit onderzoek hebben gestuurd, valt de eigenlijke computationele inferentie, constructie, of vervollediging van GRN's buiten het draagvlak van dit werk. De beschreven prestaties kunnen daarom gezien worden als een aftasten en een nivelleren van de weg die leidt naar de computationele analyse van GRN's in Metazoa. Om dit laatste te bereiken zijn verbeteringen van de bioinformatica-technieken nodig op niveau van microarray data-analyse, vergelijkingen van microarray data, detectie van transcriptiefactorbindingsplaatsen in regulatorische sequenties en de detectie zelf van deze sequenties, *in silico* validatie van computationeel gegenereerde hypothesen, netwerkstructuur-inferentie, enzovoort. In dit proefschrift werden enkele van deze bouwblokken van "Systems Biology" behandeld en werden strategieën voorgesteld —vaak door "Proof of Concept"— om ze te verbeteren of om het gebruik ervan te vergemakkelijken.

Bibliography

[1] J. Aach and G. M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.

[2] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–95, 2000.

[3] S. Aerts, B. Coessens, P. Glenisson, D. Lambrechts, and De Moor B. In silico validation of cis-regulatory modules and putative target genes using multiple genomic information sources. Technical Report 04-06, K.U.Leuven, ESAT-SCD, Leuven, Belgium, 2004.

[4] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–1764, 2003.

[5] S. Aerts, G. Thijs, M. Dabrowski, Y. Moreau, and De Moor B. Comprehensive analysis of the base composition around the transcription start site in metazoa. Technical Report 04-07, K.U.Leuven, ESAT-SCD, Leuven, Belgium, 2004.

[6] S. Aerts, P. Van Loo., Y. Moreau, and B. De Moor. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, In press, 2004.

[7] S. Aerts, P. Van Loo., G. Thijs, Y. Moreau, and B. De Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2:II5–II14, 2003.

[8] G. Ahnert-Hilger, U. Kutay, I. Chahoud, T. Rapoport, and B. Wiedenmann. Synaptobrevin is essential for secretion but not for the development of synaptic processes. *Eur J Cell Biol*, 70(1):1–11, 1996.

[9] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–10106, 2000.

[10] P. Antal, P. Glenisson, G. Fannes, J. Mathys, B. De Moor, and Y. Moreau. On the potential of domain literature for clustering and bayesian network learning. In *Proceedings of the 8th ACM-SIGKDD Int. Conf. on Knowledge Discovery and Datamining*, pages 405–414, 2002.

[11] F. Antequera. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci*, 60(8):1647–58, 2003.

[12] F. Antequera and A. Bird. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A*, 90(24):11995–11999, 1993.

[13] M.N. Arbeitman, E.E.M. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M.A. Krasnow, M.P. Scott, R.W. Davis, and K.P. White. Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, 297(5590):2270–2275, 2002.

[14] M.I. Arnone and E.H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–64, 1997.

[15] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[16] A. Auf der Maur, T. Belser, G. Elgar, O. Georgiev, and W. Schaffner. Characterization of the transcription factor MTF-1 from the Japanese pufferfish (Fugu rubripes) reveals evolutionary conservation of heavy. *Biol Chem*, 380(2):175–85, 1999.

[17] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995.

[18] T.L. Bailey and W.S. Noble. Searching for statistically significant regulatory modules. *Bioinformatics*, 19 Suppl 2:II16–II25, 2003.

[19] P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

[20] A. Beletskii and A. S. Bhagwat. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 93(24):13919–13924, 1996.

[21] P.V. Benos, M.L. Bulyk, and G.D. Stormo. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–51, 2002.

[22] E. Berezikov, V. Guryev, R.H. Plasterk, and E. Cuppen. CONREAL: Conserved Regulatory Elements Anchored Alignment Algorithm for Identification of Transcription Factor Binding Sites by Phylogenetic. *Genome Res*, 14(1):170–8, 2004.

[23] O. Berezovska, P. McLean, R. Knowles, M. Frosh, F. M. Lu, S. E. Lux, and B. T. Hyman. Notch1 inhibits neurite outgrowth in postmitotic primary neurons. *Neuroscience*, 93(2):433–9, 1999.

[24] O.G. Berg and P.H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–50, 1987.

[25] M. Bergmann, D. Grabs, and G. Rager. Developmental expression of dynamin in the chick retinotectal system. *J Histochem Cytochem*, 47(10):1297–306, 1999.

[26] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci U S A*, 99(2):757–762, 2002.

[27] G. Bernardi. The human genome: organization and evolutionary history. *Annu Rev Genet*, 29:445–476, 1995.

[28] A. Bird. DNA methylation de novo. *Science*, 286(5448):2287–2288, 1999.

[29] A. P. Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*, 8(7):1499–1504, 1980.

[30] A. P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213, 1986.

[31] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, et al. Ensembl 2004. *Nucleic Acids Res*, 32(1):D468–70, 2004.

[32] E. Birney and R. Durbin. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc Int Conf Intell Syst Mol Biol*, 5:56–64, 1997.

[33] J. A. Blake, J. E. Richardson, C. J. Bult, J. A. Kadin, and J. T. Eppig. The mouse genome database (mgd): the model organism database for the laboratory mouse. *Nucleic Acids Res*, 30(1):113–5, 2002.

[34] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *J Comput Biol*, 9(2):211–223, 2002.

[35] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, 12(5):739–748, 2002.

[36] M. Blanchette and M. Tompa. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res*, 31(13):3840–3842, 2003.

[37] T. Blumenthal, D. Evans, C.D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W.L. Chiu, K. Duke, M. Kiraly, and S.K. Kim. A global analysis of Caenorhabditis elegans operons. *Nature*, 417(6891):851–4, 2002.

[38] D. Boffelli, J. McAuliffe, D. Ovcharenko, K.D. Lewis, I. Ovcharenko, L. Pachter, and E.M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, 2003.

[39] C. Boyer, T. Schikorski, and C. F. Stevens. Comparison of hippocampal dendritic spines in culture and in brain. *J Neurosci*, 18(14):5294–300, 1998.

[40] F. Bradke and C. G. Dotti. Changes in membrane trafficking and actin dynamics during axon formation in cultured hippocampal neurons. *Microsc Res Tech*, 48(1):3–11, 2000.

[41] F. Bradke and C. G. Dotti. Establishment of neuronal polarity: lessons from cultured hippocampal neurons. *Curr Opin Neurobiol*, 10(5):574–81, 2000.

[42] N. Bray, I. Dubchak, and L. Pachter. AVID: A Global Alignment Program. *Genome Res*, 13(1):97–102, 2003.

[43] N. Bray, A. Fabrikant, J. Lord, J. Schwartz, I. Dubchak, and L. Pachter. AVID: A global alignment program for large genomic sequences. *In preparation*, 2001.

[44] N. Bray and L. Pachter. MAVID multiple alignment server. *Nucleic Acids Res*, 31(13):3525–6, 2003.

[45] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4):365–371, 2001.

[46] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*, 8(11):1202–1215, 1998.

[47] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G.G. Lara, A. Oezcimen, P. Rocca-Serra, and S.A.. Sansone. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 31(1):68–71, 2003.

[48] M. Brudno, C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, and S. Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721–31, 2003.

[49] M.L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol*, 5(1):201, 2003.

[50] M.L. Bulyk, X. Huang, Y. Choo, and G.M. Church. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A*, 98(13):7158–63, 2001.

[51] M.L. Bulyk, P.L. Johnson, and G.M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–61, 2002.

[52] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, 1997.

[53] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18):10096–10100, 2000.

[54] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–171, 2001.

[55] K.J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay, and J.N. Weinstein. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol*, 4(4):R27, 2003.

[56] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A*, 97(22):12182–6, 2000.

[57] D. Capiati, S. Benassati, and R.L. Boland. 1,25(OH)2-vitamin D3 induces translocation of the vitamin D receptor (VDR) to the plasma membrane in skeletal muscle cells. *J Cell Biochem*, 86(1):128–135, 2002.

[58] M. Carey. The enhanceosome and transcriptional synergy. *Cell*, 92(1):5–8, 1998.

[59] S.B. Carroll, J.K. Grenier, and S.D. Weatherbee. *From DNA to diversity*. Blackwell Science Inc, Massachusetts USA, 2001.

[60] J. E. Casanova, X. Wang, R. Kumar, S. G. Bhartur, J. Navarre, J. E. Woodrum, Y. Altschuler, G. S. Ray, and J. R. Goldenring. Association of rab25 and rab11a with the apical recycling system of polarized madin-darby canine kidney cells. *Mol Biol Cell*, 10(1):47–61, 1999.

[61] M. Caselle, F. Di Cunto, and P. Provero. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics*, 3(1):7, 2002.

[62] E. Chargaff. Structure and function of nucleic acids as cell constituents. *Fed Proc*, 10:654–659, 1951.

[63] Q. K. Chen, G. Z. Hertz, and G. D. Stormo. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci*, 11(5):563–566, 1995.

[64] Y. A. Chen and R. H. Scheller. Snare-mediated membrane fusion. *Nat Rev Mol Cell Biol*, 2(2):98–106, 2001.

[65] T. J. Chilcote, T. Galli, O. Mundigl, L. Edelmann, P. S. McPherson, K. Takei, and P. De Camilli. Cellubrevin and synaptobrevins: similar subcellular localization and biochemical properties in pc12 cells. *J Cell Biol*, 129(1):219–31, 1995.

[66] J.M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet*, 6(10):1735–44, 1997.

[67] B. Coessens, G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys, and B. De Moor. INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res*, 31(13):3468–3470, 2003.

[68] P.R. Cook. Predicting three-dimensional genome structure from transcriptional activity. *Nat Genet*, 32(3):347–52, 2002.

[69] T. Cook, B. Gebelein, and R. Urrutia. Sp1 and its likes: Biochemical and functional predictions for a growing family of zinc finger transcription factors. *Ann NY Acad Sci*, 880:94–102, 1999.

[70] D.R. Corey and J.M. Abrams. Morpholino antisense oligonucleotides: tools for investigating vertebrate development. *Genome Biol*, 2(5):REVIEWS1015, 2001.

[71] C. Coulondre, J. H. Miller, P. J. Farabaugh, and W. Gilbert. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*, 274(5673):775–780, 1978.

[72] A. J. Courey. Cooperativity in transcriptional control. *Curr Biol*, 11(7):250–252, 2001.

[73] M.L. Crowe, C. Serizet, V. Thareau, S. Aubourg, P. Rouze, P. Hilson, J. Beynon, P. Weisbeek, P. van Hummelen, P. Reymond, J. Paz-Ares, W. Nietfeld, and M. Trick. CATMA: a complete Arabidopsis GST database. *Nucleic Acids Res*, 31(1):156–8, 2003.

[74] E. M. Crowley, K. Roeder, and M. Bina. A statistical model for locating regulatory regions in genomic DNA. *J Mol Biol*, 268(1):8–14, 1997.

[75] X. Cui and G.A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210, 2003.

[76] M. Dabrowski, S. Aerts, B. De Strooper, and Y. Moreau. Singular value decomposition of gene expression data suggests shared regulatory inputs. Belgian Bioinformatics Conference 2003, Leuven, Belgium, 2003.

[77] M. Dabrowski, S. Aerts, P. Van Hummelen, K. Craessaerts, B. De Moor, W. Annaert, Y. Moreau, and B. De Strooper. Gene profiling of hippocampal neuronal culture. *J Neurochem*, 85(5):1279–1288, 2003.

[78] E. H. Davidson. *Genomic Regulatory Systems.* Academic Press, San Diego, USA, 2001.

[79] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, M. J. Schilstra, P. J. C. Clarke, A. G. Rust, Z. Pan, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev Biol*, 246(1):162–190, 2002.

[80] R. V. Davuluri, I. Grosse, and M. Q. Zhang. Computational identification of promoters and first exons in the human genome. *Nat Genet*, 29(4):412–417, 2001.

[81] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor, and Y. Moreau. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–46, 2002.

[82] H. Denys. *β-Catenin target genes in desmoid tumors.* PhD thesis, K.U.Leuven, 2003.

[83] H. Denys, A. Jadidizadeh, S. Amini Nik, K. Van Dam, S. Aerts, B.A. Alman, J.J. Cassiman, and S. Tejpar. Identification of IGFBP-6 as a significantly downregulated gene by beta-catenin in desmoid tumors. *Oncogene*, 23(3):654–664, 2004.

[84] P. P. Di Fiore and P. De Camilli. Endocytosis and signaling. an inseparable partnership. *Cell*, 106(1):1–4, 2001.

[85] E. Diaz, Y. Ge, Y. H. Yang, K. C. Loh, T. A. Serafini, Y. Okazaki, Y. Hayashizaki, T. P. Speed, J. Ngai, and P. Scheiffele. Molecular analysis of gene expression in the developing pontocerebellar projection system. *Neuron*, 36(3):417–34, 2002.

[86] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J.C. Matese, T. Hernandez-Boussard, C.A. Rees, J.M. Cherry, D. Botstein, P.O. Brown, and A.A. Alizadeh. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res*, 31(1):219–223, 2003.

[87] C. Dieterich, H. Wang, K. Rateitschak, H. Luz, and M. Vingron. CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res*, 31(1):55–57, 2003.

[88] H.H. Do, K. Toralf, and E. Rahm. Comparative evaluation of microarray-based gene expression databases. In G. Weikum, H. Schning, and Rahm E., editors, *GI-Edition Lecture Notes in Informatics*, volume P-26, pages 482–502, 2003.

[89] S. Dornan, A. P. Jackson, and N. J. Gay. Alpha-adaptin, a marker for endocytosis, is expressed in complex patterns during *Drosophila* development. *Mol Biol Cell*, 8(8):1391–403, 1997.

[90] C. G. Dotti, C. A. Sullivan, and G. A. Banker. The establishment of polarity by hippocampal neurons in culture. *J Neurosci*, 8(4):1454–68, 1988.

[91] T. A. Down and T. J. P. Hubbard. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res*, 12(3):458–461, 2002.

[92] I. Dubchak, M. Brudno, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res*, 10(9):1304–1306, 2000.

[93] A.M. Dudley, J. Aach, M.A. Steffen, and G.M. Church. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci U S A*, 99(11):7554–9, 2002.

[94] S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods foridentifying differentially expressed genes in replicated cdna microarray experiments. Technical Report 578, Stanford University, 2000.

[95] S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919–929, 2001.

[96] R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, 2002.

[97] Editorial. Coming to terms with microarrays. *Nat Genet*, 32(Suppl):333–334, 2002.

[98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, 1998.

[99] E. Eisenberg and E.Y. Levanon. Human housekeeping genes are compact. *Trends Genet*, 19(7):362–365, 2003.

[100] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. Genome-wide *in silico* identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res*, 13(5):773–80, 2003.

[101] E.G. Emberly, N. Rajewsky, and E.D. Siggia. Conservation of Regulatory Elements between two species of *Drosophila*. *BMC Bioinformatics*, 4(57), 2003.

[102] A. Eyre-Walker. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, 152(2):675–683, 1999.

[103] K. Faire, F. Trent, J. M. Tepper, and E. M. Bonder. Analysis of dynamin isoforms in mammalian brain: dynamin-1 expression is spatially and temporally regulated during postnatal development. *Proc Natl Acad Sci U S A*, 89(17):8376–80, 1992.

[104] H.F. Farhadi, P. Lepage, R. Forghani, H.C.H. Friedman, W. Orfali, L. Jasmin, W. Miller, T.J. Hudson, and A.C. Peterson. A combinatorial network of evolutionarily conserved myelin basic protein regulatory sequences confers distinct glial-specific phenotypes. *J Neurosci*, 23(32):10214–10223, 2003.

[105] J.W. Fickett and A.G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Res*, 7(9):861–78, 1997.

[106] T. L. Fletcher and G. A. Banker. The establishment of polarity by hippocampal neurons: the relationship between the stage of a cell's development in situ and its subsequent development in culture. *Dev Biol*, 136(2):446–54, 1989.

[107] T. L. Fletcher, P. Cameron, P. De Camilli, and G. Banker. The distribution of synapsin i and synaptophysin in hippocampal neurons developing in culture. *J Neurosci*, 11(6):1617–26, 1991.

[108] T. L. Fletcher, P. De Camilli, and G. Banker. Synaptogenesis in hippocampal cultures: evidence indicating that axons and dendrites become competent to form synapses at different stages of neuronal development. *J Neurosci*, 14:6695–706, 1994.

[109] FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*, 31(1):172–5, 2003.

[110] W.B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, 1992.

[111] M. P. Francino and H. Ochman. Strand asymmetries in DNA evolution. *Trends Genet*, 13(6):240–245, 1997.

[112] A. C. Frank and J. R. Lobry. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, 238(1):65–77, 1999.

[113] J. Freudenberg and P. Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18 Suppl 2:S110–5, 2002.

[114] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–20, 2000.

[115] M. C. Frith, U. Hansen, and Z. Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889, 2001.

[116] M.C. Frith, J.L. Spouge, U. Hansen, and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*, 30(14):3214–3224, 2002.

[117] M. Gardiner-Garden and T. G. Littlejohn. A comparison of microarray databases. *Brief Bioinform*, 2(2):143–158, 2001.

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

[119] R.A. Gibbs, G.M. Weinstock, M.L. Metzker, D.M. Muzny, E.J. Sodergren, S. Scherer, G. Scott, D. Steffen, K.C. Worley, P.E. Burch, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, 2004.

[120] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput*, pages 391–402, 2003.

[121] P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. Text-based gene profiling with domain-specific views. In *Proc. of the First International Workshop on Semantic Web and Databases (SWDB 2003)*, pages 15–31, 2003.

[122] P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. Txtgate : Profiling gene groups with text-based information. *Genome Biol*, In press, 2004.

[123] A. Goldbeter. Computational approaches to cellular rhythms. *Nature*, 420(6912):238–245, 2002.

[124] J. Gollub, Catherine. A. Ball, Gail. Binkley, Janos. Demeter, David. B. Finkelstein, Joan. M. Hebert, Tina. Hernandez-Boussard, Heng. Jin, Miroslava. Kaloper, John. C. Matese, Mark. Schroeder, Patrick. O. Brown, David. Botstein, and Gavin. Sherlock. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*, 31(1):94–96, 2003.

[125] K. Goslin and G. Banker. Rat hippocampal neurons in low-density culture. In K. Goslin, editor, *Culturing Nerve Cells*, Cellular and Molecular Neuroscience Series, pages 251–282. The MIT Press, Cambrige, Massachusetts, 1991.

[126] B. Gottgens, L. M. Barton, J. G. Gilbert, A. J. Bench, M. J. Sanchez, S. Bahn, S. Mistry, D. Grafham, A. McMurray, M. Vaudin, E. Amaya, D. R. Bentley, A. R. Green, and A. M. Sinclair. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol*, 18(2):181–186, 2000.

[127] P. Green, B. Ewing, W. Miller, P.J. Thomas, and E.D. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*, 33(4):514–517, 2003.

[128] D. GuhaThakurta and G. D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, 2001.

[129] C. A. Haas and L. J. DeGennaro. Multiple synapsin i messenger rnas are differentially regulated during neuronal development. *J Cell Biol*, 106(1):195–203, 1988.

[130] T. K. Hale and A. W. Braithwaite. The adenovirus oncoprotein E1a stimulates binding of transcription factor ETF to transcriptionally activate the p53 gene. *J Biol Chem*, 274(34):23777–23786, 1999.

[131] A.S. Halees, D. Leyfer, and Z. Weng. PromoSer: A large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Res*, 31(13):3554–3559, 2003.

[132] M. S. Halfon, Y. Grad, G. M. Church, and A. M. Michelson. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res*, 12(7):1019–1028, 2002.

[133] S. Hannenhalli and S. Levy. Promoter prediction in the human genome. *Bioinformatics*, 17 Suppl 1:S90–6, 2001.

[134] R.C. Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet*, 16(9):369–72, 2000.

[135] T.W. Harris, N. Chen, F. Cunningham, M. Tello-Ruiz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, J. Chan, et al. WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res*, 32(1):D411–7, 2004.

[136] P.E. Hart, N.J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern*, SSC-4:100–107, 1968.

[137] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, and R.A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput*, pages 437–49, 2002.

[138] M.A. Hauser, Y.J. Li, S. Takeuchi, R. Walters, M. Noureddine, M. Maready, T. Darden, C. Hulette, E. Martin, E. Hauser, H. Xu, D. Schmechel, J.E. Stenger, F. Dietrich, and J. Vance. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum Mol Genet*, 12(6):671–7, 2003.

[139] Peter. M. Haverty, Zhiping. Weng, Nathan. L. Best, Kenneth. R. Auerbach, Li.-Li Hsiao, Roderick. V. Jensen, and Steven. R. Gullans. HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res*, 30(1):214–217, 2002.

[140] Z. He, K. C. Wang, V. Koprivica, G. Ming, and H. J. Song. Knowing how to navigate: mechanisms of semaphorin signaling in the nervous system. *Sci STKE*, 2002(119):RE1, 2002.

[141] J. Heasman. Morpholino oligos: making sense of antisense? *Dev Biol*, 243(2):209–14, 2002.

[142] A.W. Helms, A.L. Abney, N. Ben-Arie, H.Y. Zoghbi, and J.E. Johnson. Autoregulation and multiple enhancers control Math1 expression in the developing nervous system. *Development*, 127(6):1185–96, 2000.

[143] B. Hendrich and S. Tweedie. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet*, 19(5):269–277, 2003.

[144] M.J. Herrgard, M.W. Covert, and B.O. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res*, 13(11):2423–34, 2003.

[145] G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999.

[146] V.F. Hinman, A.T. Nguyen, R.A. Cameron, and E.H. Davidson. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc Natl Acad Sci U S A*, 100(23):13356–13361, 2003.

[147] L. Hood and D. Galas. The digital code of DNA. *Nature*, 421(6921):444–8, 2003.

[148] F. W. Hopf, J. Waters, S. Mehta, and S. J. Smith. Stability and plasticity of developing synapses in hippocampal neuronal cultures. *J Neurosci*, 22(3):775–81, 2002.

[149] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. The Ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41, 2002.

[150] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*, 296(5):1205–1214, 2000.

[151] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[152] F. J. Iborra, D. A. Jackson, and P. R. Cook. Coupled transcription and translation within nuclei of mammalian cells. *Science*, 293(5532):1139–1142, 2001.

[153] I. P. Ioshikhes and M. Q. Zhang. Large-scale human promoter mapping using CpG islands. *Nat Genet*, 26(1):61–63, 2000.

[154] V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, and P.O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819):533–8, 2001.

[155] R. Jahn and T. C. Sudhof. Membrane fusion and exocytosis. *Annu Rev Biochem*, 68:863–911, 1999.

[156] M. Jareb and G. Banker. Inhibition of axonal growth by brefeldin a in hippocampal neurons in culture. *J Neurosci*, 17(23):8955–63, 1997.

[157] N. Jareborg, E. Birney, and R. Durbin. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res*, 9(9):815–824, 1999.

[158] N. Jarousse and R. B. Kelly. Endocytotic mechanisms in synapses. *Curr Opin Cell Biol*, 13(4):461–9, 2001.

[159] R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol Cell Biol*, 14(1):116–127, 1994.

[160] A.G. Jegga, S.P. Sherwood, J.W. Carman, A.T. Pinski, J.L. Phillips, J.P. Pestian, and B.J. Aronow. Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res*, 12(9):1408–1417, 2002.

[161] O. Johansson, W. Alkema, W.W. Wasserman, and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19(Suppl 1):I169–I176, 2003.

[162] K. S. Katula, K. L. Wright, H. Paul, D. R. Surman, F. J. Nuckolls, J. W. Smith, J. P. Ting, J. Yates, and J. P. Cogswell. Cyclin-dependent kinase activation and S-phase induction of the cyclin B1 gene are linked through the CCAAT elements. *Cell Growth Differ*, 8(7):811–820, 1997.

[163] A.E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–9, 2003.

[164] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6):819–837, 2000.

[165] O.D. King and F.P. Roth. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*, 31(19):e116, 2003.

[166] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[167] S. Knudsen. Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics*, 15(5):356–61, 1999.

[168] W. Krivan and W. W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*, 11(9):1559–1566, 2001.

[169] M.M. Kulkarni and D.N. Arnosti. Information display by transcriptional enhancers. *Development*, 130(26):6569–6575, 2003.

[170] S. Kullback. *Information Theory and Statistics.* John Wiley & Sons, New York, USA, 1959.

[171] W. P. Kuo, T. K. Jenssen, A. J. Butte, L. Ohno-Machado, and I. S. Kohane. Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–12, 2002.

[172] A.K. Kutach and J.T. Kadonaga. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol Cell Biol*, 20(13):4754–64, 2000.

[173] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[174] L. A. Lapierre, R. Kumar, C. M. Hales, J. Navarre, S. G. Bhartur, J. O. Burnette, Jr. Provance, D. W., J. A. Mercer, M. Bahler, and J. R. Goldenring. Myosin vb is associated with plasma membrane recycling systems. *Mol Biol Cell*, 12(6):1843–57, 2001.

[175] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz. CpG islands as gene markers in the human genome. *Genomics*, 13(4):1095–1107, 1992.

[176] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[177] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D. B. Gordon, B. Ren, J.J. Wyrick, J. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, and R.A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[178] B. Lemon and R. Tjian. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev*, 14(20):2551–2569, 2000.

[179] B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W.W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(13), 2003.

[180] M. Lermen and K. Reinert. The practical use of the A* algorithm for exact multiple sequence alignment. *J Comput Biol*, 7(5):655–671, 2000.

[181] Y.F. Leung and D. Cavalieri. Fundamentals of cDNA microarray data analysis. *Trends Genet*, 19(11):649–659, 2003.

[182] M.P. Levesque and P.N. Benfey. Systems biology. *Curr Biol*, 14(5):179–180, 2004.

[183] B. Lewin. *Genes VII.* Oxford University Press, Oxford, 2000.

[184] H. Y. Li, M. Kotaka, S. Kostin, S. M. Lee, L. D. Kok, K. K. Chan, S. K. Tsui, J. Schaper, R. Zimmermann, C. Y. Lee, K. P. Fung, and M. M. Waye. Translocation of a human focal adhesion LIM-only protein, FHL2, during myofibrillogenesis and identification of LIM2 as the principal determinants of FHL2 focal adhesion localization. *Cell Motil Cytoskeleton*, 48(1):11–23, 2001.

[185] R. C. Lin and R. H. Scheller. Mechanisms of synaptic vesicle exocytosis. *Annu Rev Cell Dev Biol*, 16:19–49, 2000.

[186] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–24, 1999.

[187] R. Liu and D.J. States. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res*, 12(3):462–9, 2002.

[188] T. E. Lloyd, P. Verstreken, E. J. Ostrin, A. Phillippi, O. Lichtarge, and H. J. Bellen. A genome-wide search for synaptic vesicle cycle proteins in *Drosophila*. *Neuron*, 26(1):45–50, 2000.

[189] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80, 1996.

[190] A. T. Look. Oncogenic transcription factors in the human acute leukemias. *Science*, 278(5340):1059–1064, 1997.

[191] G. G. Loots, R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–140, 2000.

[192] G. G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak, and E. M. Rubin. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*, 12(5):832–839, 2002.

[193] E. Louie, J. Ott, and J. Majewski. Nucleotide frequency variation across human genes. *Genome Res*, 13(12):2594–601, 2003.

[194] G.F. Luger. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison Wesley, Harlow, England, 4 edition, 2001.

[195] A.V. Lukashin and M. Borodowsky. Genemark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26:1107–1115, 1998.

[196] N.M. Luscombe, S.E. Austin, H.M. Berman, and J.M. Thornton. An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):REVIEWS001, 2000.

[197] N.M. Luscombe, R.A. Laskowski, and J.M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*, 29(13):2860–74, 2001.

[198] J. Majewski. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet*, 73(3):688–692, 2003.

[199] J. Majewski and J. Ott. Distribution and characterization of regulatory elements in the human genome. *Genome Res*, 12(12):1827–36, 2002.

[200] L.J. Maltais, J.A. Blake, J.T. Eppig, and M.T. Davisson. Rules and guidelines for mouse gene nomenclature: a condensed version. International Committee on Standardized Genetic Nomenclature for Mice. *Genomics*, 45(2):471–6, 1997.

[201] P. Marc, F. Devaux, and C. Jacq. yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res*, 29(13):63–63, 2001.

[202] K. Marchal, K. Engelen, J. De Brabanter, S. Aerts, B. De Moor, T Ayoubi, and P. Van Hummelen. Comparison of different methodologies to identify differentially expressed genes in two-sample cdna arrays. *Journal of Biological Systems*, 10(4):409–430, 2002.

[203] K. Marchal, K. Engelen, J. De Brabanter, and B. De Moor. A guideline for the analysis of two sample microarray data. Technical Report 02-87, K.U.Leuven, ESAT-SCD, Leuven, Belgium, 2002.

[204] Kathleen. Marchal, Gert. Thijs, Sigrid. De. Keersmaecker, Pieter. Monsieurs, Bart. De. Moor, and Jos. Vanderleyden. Genome-specific higher-order background models to improve motif detection. *Trends Microbiol*, 11(2):61–66, 2003.

[205] M. Markstein, P. Markstein, V. Markstein, and M. S. Levine. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc Natl Acad Sci U S A*, 99(2):763–768, 2002.

[206] L. Marsan and M. F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol*, 7(3-4):345–362, 2000.

[207] S. Martinez-Arca, S. Coco, G. Mainguy, U. Schenk, P. Alberts, P. Bouille, M. Mezzina, A. Prochiantz, M. Matteoli, D. Louvard, and T. Galli. A common exocytotic mechanism mediates axonal and dendritic outgrowth. *J Neurosci*, 21(11):3830–8, 2001.

[208] C. Mathe, M.F. Sagot, T. Schiex, and P. Rouze. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*, 30(19):4103–17, 2002.

[209] C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. S. Pachter, and I. Dubchak. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047, 2000.

[210] M.I. McCarthy, D. Smedley, and W. Hide. New methods for finding disease-susceptibility genes: impact and potential. *Genome Biol*, 4(10):119, 2003.

[211] M. Mietus-Snyder, F. M. Sladek, G. S. Ginsburg, C. F. Kuo, J. A. Ladias, J. E. Jr. Darnell, and S. K. Karathanasis. Antagonism between apolipoprotein AI regulatory protein 1, Ear3/COUP-TF, and hepatocyte nuclear factor 4 modulates apolipoprotein CIII gene expression in liver and intestinal cells. *Mol Cell Biol*, 12(4):1708–1718, 1992.

[212] J.A. Mitchell, A.R. Aronson, J.G. Mork, L.C. Folk, S.M. Humphrey, and J.M. Ward. Gene Indexing: Characterization and Analysis of NLM's GeneRIFs. *Proc AMIA Symp*, pages 460–4, 2003.

[213] S. Mochida. Protein-protein interactions in neurotransmitter release. *Neurosci Res*, 36(3):175–82, 2000.

[214] M. Mody, Y. Cao, Z. Cui, K. Y. Tay, A. Shyong, E. Shimizu, K. Pham, P. Schultz, D. Welsh, and J. Z. Tsien. Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc Natl Acad Sci U S A*, 98(15):8862–7, 2001.

[215] Y. Moreau, S. Aerts, B. De Moor, B. De Strooper, and M. Dabrowski. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*, 19(10):570–577, 2003.

[216] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, 31(1):315–8, 2003.

[217] N. Munshi, Y. Yie, M. Merika, K. Senger, S. Lomvardas, T. Agalioti, and D. Thanos. The IFN-beta enhancer: a paradigm for understanding activation and repression of inducible gene expression. *Cold Spring Harb Symp Quant Biol*, 64:149–159, 1999.

[218] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970.

[219] Y. Nemoto and P. De Camilli. Recruitment of an alternatively spliced form of synaptojanin 2 to mitochondria by the interaction with the pdz domain of a mitochondrial outer membrane protein. *Embo J*, 18(11):2991–3006, 1999.

[220] W. Jr. Nikovits, J. H. Mar, and C. P. Ordahl. Muscle-specific activity of the skeletal troponin I promoter requires interaction between upstream regulatory sequences and elements contained within the first transcribed exon. *Mol Cell Biol*, 10(7):3468–3482, 1990.

[221] M.A. Nobrega, I. Ovcharenko, V. Afzal, and E.M. Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, 2003.

[222] J.C. Oeltjen, T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, and J.W. Belmont. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res*, 7(4):315–29, 1997.

[223] U. Ohler, H. Niemann, G.c. Liao, and G.M. Rubin. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17 Suppl 1:S199–206, 2001.

[224] Uwe. Ohler, Guo.-chun Liao, Heinrich. Niemann, and Gerald. M. Rubin. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol*, 3(12):RESEARCH0087, 2002.

[225] A. R. Oller, I. J. Fijalkowska, R. L. Dunn, and R. M. Schaaper. Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 89(22):11036–11040, 1992.

[226] F. Parlati, O. Varlamov, K. Paz, J. A. McNew, D. Hurtado, T. H. Sollner, and J. E. Rothman. Distinct snare complexes mediating membrane fusion in golgi transport based on combinatorial specificity. *Proc Natl Acad Sci U S A*, 99(8):5424–9, 2002.

[227] P. Pavlidis, T. S. Furey, M. Liberto, D. Haussler, and W. N. Grundy. Promoter region-based classification of genes. *Pac Symp Biocomput*, pages 151–163, 2001.

[228] A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak. DNA structure in human RNA polymerase II promoters. *J Mol Biol*, 281(4):663–673, 1998.

[229] A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak. The biology of eukaryotic promoter prediction–a review. *Comput Chem*, 23(3-4):191–207, 1999.

[230] L.A. Pennacchio and E.M. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet*, 2(2):100–9, 2001.

[231] C. Perez-Iratxeta, P. Bork, and M.A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31(3):316–9, 2002.

[232] Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, 29(2):153–9, 2001.

[233] J. Pokorny and T. Yamamoto. Postnatal ontogenesis of hippocampal ca1 area in rats. i. development of dendritic arborisation in pyramidal neurons. *Brain Res Bull*, 7(2):113–20, 1981.

[234] R. Pollock and R. Treisman. A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Res*, 18(21):6197–6204, 1990.

[235] Loic. Ponger and Dominique. Mouchiroud. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 18(4):631–633, 2002.

[236] D.S. Prestridge. Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol*, 249(5):923–32, 1995.

[237] D.S. Prestridge. SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput Appl Biosci*, 12(2):157–60, 1996.

[238] M. Ptashne and A. Gann. *Genes & signals*. Cold Spring Harbor Laboratory Press, New York, USA, 2002.

[239] L. G. Puskas, A. Zvara, Jr. Hackler, L., and P. Van Hummelen. Rna amplification results in reproducible microarray data with slight ratio bias. *Biotechniques*, 32(6):1330–4, 1336, 1338, 1340, 2002.

[240] J. Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501, 2002.

[241] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, 23(23):4878–4884, 1995.

[242] N. Rajewsky, M. Vergassola, U. Gaul, and E.D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, 3(1):30, 2002.

[243] Sridhar. Ramaswamy, Ken. N. Ross, Eric. S. Lander, and Todd. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nat Genet*, 33(1):49–54, 2003.

[244] M. Rebeiz, N. L. Reeves, and James. W. P. SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc Natl Acad Sci U S A*, 99(15):9888–9893, 2002.

[245] L. Redmond, S. R. Oh, C. Hicks, G. Weinmaster, and A. Ghosh. Nuclear notch1 signaling and the regulation of dendritic development. *Nat Neurosci*, 3(1):30–40, 2000.

[246] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–75, 2003.

[247] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–9, 2000.

[248] J.C. Reyes, C. Muchardt, and M. Yaniv. Components of the human SWI/SNF complex are enriched in active chromatin and are associated with the nuclear matrix. *J Cell Biol*, 137(2):263–74, 1997.

[249] R. Rivera-Pomar and H. Jackle. From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. *Trends Genet*, 12(11):478–83, 1996.

[250] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–35, 2000.

[251] G. Russo, C. Zegar, and A. Giordano. Advantages and limitations of microarray technology in human cancer. *Oncogene*, 22(42):6497–507, 2003.

[252] A. Sandelin, W. Alkema, P. Engstrom, W.W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(1):91–94, 2004.

[253] M. Schaub, E. Myslinski, C. Schuster, A. Krol, and P. Carbon. Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III. *EMBO J*, 16(1):173–181, 1997.

[254] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[255] M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol*, 297(3):599–606, 2000.

[256] C.D. Schmid, V. Praz, M. Delorenzi, R. Perier, and P. Bucher. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res*, 32 Database issue:D82–5, 2004.

[257] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100, 1990.

[258] T.D. Schneider, G.D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J Mol Biol*, 188(3):415–31, 1986.

[259] S. Schwartz, L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, E.D. Green, R.C. Hardison, and W. Miller. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res*, 31(13):3518–24, 2003.

[260] S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res*, 10(4):577–586, 2000.

[261] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, 2003.

[262] R. Sharan, I. Ovcharenko, A. Ben-Hur, and R.M. Karp. CREME: a framework for identifying *cis*-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19(Suppl 1):I283–I291, 2003.

[263] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, and J. M. Cherry. The Stanford Microarray Database. *Nucleic Acids Res*, 29(1):152–155, 2001.

[264] T.S. Shimizu, K. Takahashi, and M. Tomita. CpG distribution patterns in methylated and non-methylated species. *Gene*, 205(1-2):103–7, 1997.

[265] S. Sinha, E. Van Nimwegen, and E.D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(Suppl 1):I292–I301, 2003.

[266] S. Sitzler, I. Oldenburg, G. Petersen, and E. K. Bautz. Analysis of the promoter region of the housekeeping gene DmRP140 by sequence comparison of *Drosophila melanogaster* and *Drosophila virilis*. *Gene*, 100:155–162, 1991.

[267] T. Skutella and R. Nitsch. New molecules for hippocampal development. *Trends Neurosci*, 24(2):107–13, 2001.

[268] P. Smolen, D. A. Baxter, and J. H. Byrne. Mathematical modeling of gene networks. *Neuron*, 26(3):567–580, 2000.

[269] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, C.J.Jr. Stoeckert, and A. Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 3(9):RESEARCH0046, 2002.

[270] P.T. Spellman and G.M. Rubin. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol*, 1(5), 2002.

[271] J.E. Stajich, D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C.J. Mungall, B.I. Osborne, M.R. Pocock, P. Schattner, M. Senger, L.D. Stein, E. Stupka, M.D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, 2002.

[272] L. Stein. Creating a bioinformatics nation. *Nature*, 417(6885):119–120, 2002.

[273] E. Sterrenburg, R. Turk, J.M. Boer, G.B. van Ommen, and J.T. den Dunnen. A common reference for cDNA microarray hybridizations. *Nucleic Acids Res*, 30(21):e116, 2002.

[274] O. Stettler, B. Tavitian, and K. L. Moya. Differential synaptic vesicle protein expression in the barrel field of developing cortex. *J Comp Neurol*, 375(2):321–32, 1996.

[275] Christian. J. Jr. Stoeckert, Helen. C. Causton, and Catherine. A. Ball. Microarray databases: standards and ontologies. *Nat Genet*, 32 Suppl:469–473, 2002.

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

[277] G.D. Stormo. Consensus patterns in DNA. *Methods Enzymol*, 183:211–21, 1990.

[278] G.D. Stormo. Gene-finding approaches for eukaryotes. *Genome Res*, 10(4):394–7, 2000.

[279] G.D. Stormo and Hartzell GW. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A*, 86(4):1183–7, 1989.

[280] A. Stride and A.T. Hattersley. Different genes, different diabetes: lessons from maturity-onset diabetes of the young. *Ann Med*, 34(3):207–216, 2002.

[281] K. Struhl. Gene regulation. A paradigm for precision. *Science*, 293(5532):1054–5, 2001.

[282] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003.

[283] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, and J.B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A*, 99(7):4465–4470, 2002.

[284] T. Sugiyama, T. Shinoe, Y. Ito, H. Misawa, T. Tojima, E. Ito, and T. Yoshioka. A novel function of synapsin ii in neurotransmitter release. *Brain Res Mol Brain Res*, 85(1-2):133–43, 2000.

[285] Y. Sun, L.Y. Jan, and Y.N. Jan. Transcriptional regulation of atonal during development of the *Drosophila* peripheral nervous system. *Development*, 125(18):3731–40, 1998.

[286] Y. Suzuki, R. Yamashita, K. Nakai, and S. Sugano. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res*, 30(1):328–331, 2002.

[287] Y. Suzuki, R. Yamashita, S. Sugano, and K. Nakai. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res*, 32(1):D78–81, 2004.

[288] D.A. Tagle, B.F. Koop, M. Goodman, J.L. Slightom, D.L. Hess, and R.T. Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental. *J Mol Biol*, 203(2):439–55, 1988.

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

[290] B. L. Tang. Protein trafficking mechanisms associated with neurite outgrowth and polarized sorting in neurons. *J Neurochem*, 79(5):923–30, 2001.

[291] T. Tatarinova, V. Brover, M. Troukhan, and N. Alexandrov. Skew in CG content near the transcription start site in Arabidopsis thaliana. *Bioinformatics*, 19 Suppl 1:I313–I314, 2003.

[292] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(3):281–285, 1999.

[293] S. Tejpar, C. Li, C. Yu, R. Poon, H. Denys, R. Sciot, E. Van Cutsem, J. J. Cassiman, and B. Alman. Tcf-3 expression and -catenin mediated transcriptional activation in aggressive fibromatosis (desmoid tumor). *Br J Cancer*, 85:98–101, 2001.

[294] O. Tetsu and F. McCormick. Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature*, 398(6726):422–426, 1999.

[295] The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 282(5396):2012–8, 1998.

[296] G. Thijs. *Probabilistic methods to search for regulatory elements in sets of coregulated genes*. PhD thesis, K.U.Leuven, 2003.

[297] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor., P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, 2001.

[298] G. Thijs, K. Marchal, M. Lescot, S. Rombouts, B. De Moor, P. Rouze, and Y. Moreau. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–464, 2002.

[299] L. A. Tolar and L. Pallanck. Nsf function in neurotransmitter release involves rearrangement of the snare complex downstream of synaptic vesicle docking. *J Neurosci*, 18(24):10250–6, 1998.

[300] J.T. Tou and R.C. Gonzalez. Pattern classification by distance functions. In R.C. Gonzalez, editor, *Pattern recognition principles.*, pages 75–109. Addison-Wesley, Reading (Mass.), 1979.

[301] S. C. Tsai, R. Adamik, M. Tsuchiya, P. P. Chang, J. Moss, and M. Vaughan. Differential expression during development of adp-ribosylation factors, 20-kda guanine nucleotide-binding protein activators of cholera toxin. *J Biol Chem*, 266(13):8213–9, 1991.

[302] F.S. Turner, D.R. Clutterbuck, and C.A. Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4(11):R75, 2003.

[303] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21, 2001.

[304] A. Ureta-Vidal, L. Ettwiller, and E. Birney. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4(4):251–62, 2003.

[305] J. van de Peppel, P. Kemmeren, H. van Bakel, M. Radonjic, D. van Leenen, and F.C. Holstege. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep*, 4(4):387–93, 2003.

[306] M.A. van Driel, K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, and H.G. Brunner. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet*, 11(1):57–63, 2003.

[307] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281(5):827–842, 1998.

[308] J. van Helden, B. Andre, and J. Collado-Vides. A web site for the computational analysis of yeast regulatory sequences. *Yeast*, 16(2):177–187, 2000.

[309] J. van Helden., A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res*, 28(8):1808–1818, 2000.

[310] B. van Steensel., J. Delrow, and S. Henikoff. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat Genet*, 27(3):304–308, 2001.

[311] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[312] C. von Mering and P. Bork. Teamed up for transcription. *Nature*, 417(6891):797–8, 2002.

[313] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15(10):776–784, 1999.

[314] T. Wang and G.D. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–80, 2003.

[315] V.Y. Wang, B.A. Hassan, H.J. Bellen, and H.Y. Zoghbi. *Drosophila* atonal fully rescues the phenotype of Math1 null mice: new functions evolve in new cellular contexts. *Curr Biol*, 12(18):1611–1616, 2002.

[316] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–181, 1998.

[317] W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet*, 26(2):225–228, 2000.

[318] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 2002.

[319] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[320] D.L. Wheeler, D.M. Church, R. Edgar, S. Federhen, W. Helmberg, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, E. Sequeira, T.O. Suzek, T.A. Tatusova, and L. Wagner. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*, 32 Database issue:D35–40, 2004.

[321] J.A. White, P.J. McAlpine, S. Antonarakis, H. Cann, J.T. Eppig, K. Frazer, J. Frezal, D. Lancet, J. Nahmias, P. Pearson, J. Peters, A. Scott, H. Scott, N. Spurr, J.r. Talbot C, and S. Povey. Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. *Genomics*, 45(2):468–71, 1997.

[322] M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*, 13(6):1977–2000, 2002.

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

[324] P. P. Wong, N. Daneman, A. Volchuk, N. Lassam, M. C. Wilson, A. Klip, and W. S. Trimble. Tissue distribution of snap-23 and its subcellular localization in 3t3-l1 cells. *Biochem Biophys Res Commun*, 230(1):64–8, 1997.

[325] G.A. Wray, M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–1419, 2003.

[326] J. J. Wyrick and R. A. Young. Deciphering gene expression regulatory networks. *Curr Opin Genet Dev*, 12(2):130–136, 2002.

[327] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.

[328] M.K. Yeung, J. Tegner, and J.J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9):6163–8, 2002.

[329] C. Yuh, C.T. Brown, C.B. Livi, L. Rowen, P.J.C. Clarke, and E.H. Davidson. Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin. *Dev Biol*, 246(1):148–161, 2002.

[330] C. H. Yuh, H. Bolouri, and E. H. Davidson. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279(5358):1896–1902, 1998.

[331] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003.

[332] M.Q. Zhang. Identification of human gene core promoters *in silico*. *Genome Res*, 8(3):319–26, 1998.

[333] Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol*, 2(11), 2003.

[334] J. Zheng, J. Wu, and Z. Sun. An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res*, 31(7):1995–2005, 2003.

[335] J. Zhu, J. S. Liu, and C. E. Lawrence. Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, 14(1):25–39, 1998.

Curriculum vitae

Stein Aerts was born in Heusden-Zolder on February 11, 1976. He obtained a Master of Science in Engineering in Cellular and Gene Technology ("Bio-ingenieur") at the Katholieke Universiteit Leuven in 1999. His master thesis on the genetics of *Rhizobium*-plant interactions was performed at the Center for Nitrogen Fixation in Cuernavaca, Mexico. After graduation, he worked as an assistant IT project leader for Janssen Pharmaceutica (Johnson & Johnson) and as a bioinformatician for Data4s, a spin-off company of the department of Electrical Engineering ESAT-SCD of the K.U.Leuven. In 2001 he joined ESAT-SCD as a research assistant of the K.U.Leuven. In 2002 he obtained a degree in Advanced Studies in Applied Computer Sciences at the Vrije Universiteit Brussel.