COMPUTATIONAL ANALYSES ON TRANSCRIPTIONAL REGULATION IN MAMMALS

SEBASTIAN SCHMEIER

Thesis presented in fulfilment of the requirements for the Degree of *Doctor Philosophiae* in Bioinformatics at the South African National Bioinformatics Institute, University of the Western Cape.

Advisor: Prof. Vladimir Bajic

May 2009

Abstract

The genomes of various organisms have been sequenced and their transcriptome elucidated. With the information about genes and gene products readily available it has become of the utmost importance to decipher the underlying biological mechanisms that are involved in the transcriptional control of these genes. Transcription initiation is a fundamental process in living cells. It involves the interaction of transcription factors with DNA to regulate the transcription of a gene. Despite significant research during the last few decades into transcription factors and their role in gene regulation we are still far from understanding the complete transcriptional machinery that acts within biological systems.

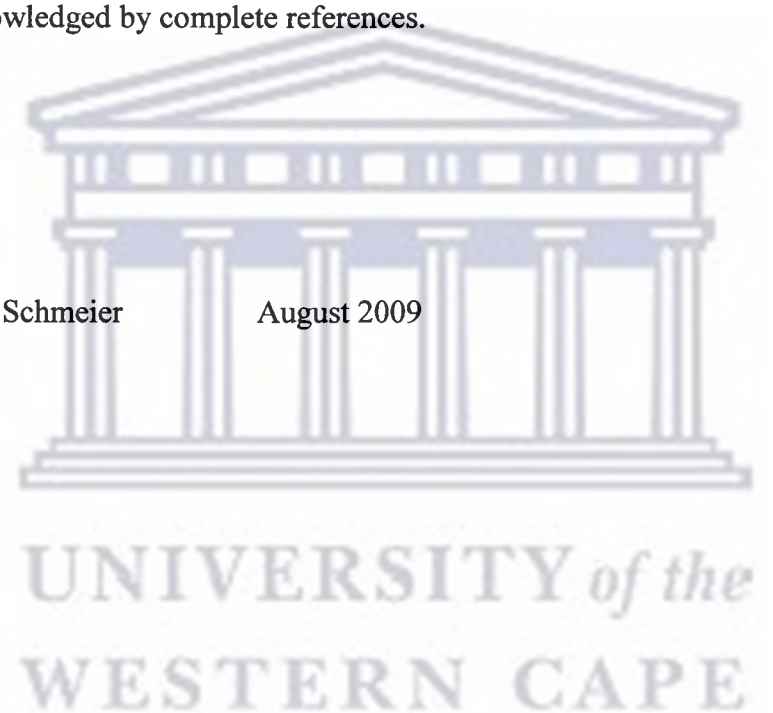
In this dissertation two computational approaches are presented to contribute to a better understanding of the transcriptional control of genes in mammals. The first addresses the transcriptional regulation of microRNA genes and its influence on the microRNA gene expression during monocytic differentiation. This is the first large-scale approach to decipher how microRNA genes are regulated by transcription factors during monocytic differentiation. The second approach relates to combinatorial gene regulation and the physical interaction of transcription factors. Here, a computational approach is used together with a novel form of numerical representation of transcription factors to predict their interactions. In this setup, the information necessary to predict the transcription

factor interactions is kept at the lowest level to minimise the data acquisition overhead that often occurs in computational prediction tasks. Both approaches enhance our insights into transcriptional control and have an impact on the further study of gene regulation.

Declaration

I declare that "*Computational Analyses on Transcriptional Regulation in Mammals*" is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have utilised or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Sebastian Schmeier August 2009



Acknowledgements

First and foremost I would like to thank my family whose unconditional support made this all possible. Their unquestioning faith in my abilities and enduring encouragement along the path of my studies provided me with the opportunity to pursue my interests. Thank you.

Furthermore, I would like to thank my advisor, Professor Vladimir Bajic, whose experience and sound advice have guided the course of this research to its successful conclusion. In him I have found a caring, fair, honest, and hard working person, characteristics I have learned that do not come naturally in our line of work. He also provided me with several opportunities to collaborate internationally that broadened my horizons. It has been a privilege and pleasure to work under his supervision.

I am grateful to my colleagues at SANBI who have assisted me with insightful comments, advice, and helpful discussions during the course of this work.

Finally, I owe thanks to old and new friends who were a source of revitalisation and supported me on innumerable occasions throughout my time in South Africa.

Publications arising from this thesis

Schmeier S, MacPherson CR, Essack M, Kaur M, Schaefer U, Suzuki S, Hayashizaki Y, and Bajic VB. **Deciphering the transcriptional circuitry of microRNA genes expressed during human monocytic differentiation.** *BMC Genomics, accepted.*

Essack M, Radovanovic A, Schaefer U, Schmeier S, Seshadri SV, Christoffels A, Kaur M, and Bajic VB. **DDEC: Dragon Database of Genes Implicated in Esophageal Cancer.** *BMC Cancer, 2009 Jul 6;9:219.*

Bajic VB, Schmeier S, and MacPherson CR. **Computational Methods to Identify Transcription Factor Binding Sites Using CAGE Information.** *In: CAP-ANALYSIS GENE EXPRESSION (CAGE) The Science of Decoding Genes Transcription, edited by Piero Carninci, Pan Stanford Publishing Pte Ltd, 2009, to appear.*

Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJL, Katayama S, Schroder K, Carninci P, Tomaru Y, Kanamori-Katayama M, Kubosaki A, Akalin A, Ando Y, Arakawa T, Arner E, Asada M, Asahara H, Bailey T, Bajic VB, Bauer D, Beckhouse AG, Bertin N, Björkegren J, Brombacher F, Bulger E, Chalk AM, Chiba J, Cloonan N, Dawe A, Dostie J, Engström PG, Essack M, Faulkner G, Fink JL, Fredman D, Fujimori K, Fukuda S, Furuno M, Gojobori T, Gough J, Grimmond SM, Gustafsson M, Hashimoto M, Hashimoto T, Hatakeyama M, Heinzel S, Hide W, Hofmann O, Hörnquist M, Huminiecki L, Ikeo K, Imamoto N, Imamura K, Inoue S, Inoue Y, Ishihara R, Iwayanagi T, Jacobsen A, Kaur M, Kawai J, Kawaji H, Kerr MC, Kimura R, Kimura S, Kimura Y, Kitano H, Koga H, Kojima T, Kondo S, Konno T, Krogh A, Kruger A, Kumar A, Lenhard B, Lennartsson A, Lindow M, Lizio M, MacPherson CR, Maeda N, Maher CA, Maqungo M, Mar JC, Matigian NA, Matsuda H, Mattick JS, Meier S, Miyamoto S, Miyamoto-Sato E, Nakabayashi K, Nakachi Y, Nygaard S, Okayama T, Okazaki Y, Okuda-Yabukami H, Orlando V, Otomo J, Pachkov M, Petrovsky N, Plessy C, Quackenbush J, Radovanovic A, Rehli M, Saito R, Sandelin A, Sano T, Schmeier S, Schönbach C, Schwartz AS, Semple CA, Sera M, Severin J, Shirahige K, Simons C, St. Laurent G, Suzuki M, Suzuki T, Sweet MJ, Taft RJ, Takeda S, Takenaka Y, Tan K, Taylor MS, Teasdale RD, Tegnér J, Teichmann S, Valen E, Wahlestedt C, Waki K, Waterhouse A, Wells CA, Winther O, Wu L, Yamaguchi K, Yanagawa H, Yasuda J, Zavolan M, and Hume DA. **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nature Genetics, 2009 May;41(5):553-62.*

Kaur M, Radovanovic A, Essack M, Schaefer U, Maqungo M, Kibler T, Schmeier S, Christoffels A, Narasimhan K, Choolani M, and Bajic VB. **Database for exploration of functional context of genes implicated in ovarian cancer.** *Nucleic Acids Res. 2009 Jan;37(Database issue):D820-3.*

Kaur M, Schmeier S, MacPherson CR, Hofmann O, Hide WA, Taylor S, Willcox N, Bajic VB. **Prioritizing genes of potential relevance to diseases affected by sex hormones: an example of Myasthenia Gravis.** *BMC Genomics. 2008 Oct 13;9:481.*

Schmeier S, MacPherson CR, Bajic VB. **Predicting interactions between transcription factors using promoter and factor properties.** *Poster at the 15th Annual International Conference on Intelligent Systems in Molecular Biology (ISMB), Vienna, Austria, 2007.*

Publications pending

Kaur M, MacPherson CR, Schmeier S, Sagar S, Schönbach C, Ravasi T, Schwegmann A, Brombacher F, Tegnér J , Suzuki H, Hayashizaki Y, and Bajic VB. **Time-dependent role of pathways and transcriptional control during differentiation of monocytes to macrophages in THP-1 cells.** *BMC Genomics, under review.*

Ravasi T, Katayama S, Bajic VB, Tan K, Schmeier S, Kanamori-Katayama M, Bertin N, Carninci P, Daub CO, Forrest ARR, Gough J, Grimmond S, Han J, Hashimoto T, Hide W, Hofmann O, Kaur M, Kawaji H, Lassmann T, van Nimwegen E, MacPherson CR, Ogawa C, Radovanovic A, Schwartz A, Teasdale RD, Tegnér J, Teichmann SA, Hume DA, Ideker T. RIKEN Omics Science Center: Arakawa T, Ninomiya N, Murakami K, Tagami M, Fukuda S, Imamura K, Kai C, Ishihara R, Kitazume Y, Kawai J, Suzuki H, and Hayashizaki Y. **An atlas of combinatorial transcriptional regulation in mouse and man.** *Cell, under review.*

Schmeier S, Dawe A, and Bajic VB. **Predicting Human Transcription Factor Interactions from Primary Structure.** *In preparation for submission.*

Table of Contents

List of Figures

List of Tables

Abbreviations

AA	amino acid
abs	absolute value
bp	base pairs
ChIP	chromatin immunoprecipitation
CPU	central processing unit
fc	fold-change
GB	gigabyte
GHz	gigahertz
hr	hour
log	logarithm
max	maximum
min	minimum
miRNA	microRNA
nt	nucleotides
PCC	Pearson's correlation coefficient
PPI	protein-protein interaction
SVM	Support vector machine
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site

Chapter 1

Introduction

Numerous genomes from a variety of species have been sequenced in the last decade [1]. The sheer volume of data produced is overwhelming making computational tools essential for processing and analysis of the information. Sequencing of complete genomes is mainly driven by the need to find and characterise the complete gene set available within an organism [2]. Elucidation of the transcriptome, the part of the genome that is transcribed into mRNA or other functional RNA, has opened avenues of research into transcriptional gene regulation. Even though the genes of an entire genome are known, it is of great interest to establish the underlying mechanisms that control the transcription of these genes.

Based on the central dogma of molecular biology, sequence information flows from DNA to proteins via RNA [3].The process of transcription is the synthesis of RNA from DNA [4]. Prior to the actual transcription step of a gene, transcription factors (TFs) bind to DNA in regulatory regions of a gene. Such regions, proximal to transcription start sites (TSSs), are known as promoters, while the control regions remote from the TSS are known as enhancers or silencers. The TFs bound to DNA mediate the binding of RNA polymerase II to the DNA. Promoter regions are essential for initiating a gene's transcription [5]. The definition and characterisation of the promoter regions is difficult,

experimentally as well as computationally, e.g. in eukaryotes mediators of transcription bind not only the immediate surroundings of the TSS but have also been found to bind several thousands of nucleotides away from the TSS [4]. The complex interplay among TFs and TFs with the promoter regions of genes with the aim to enhance or repress their transcription is generally denoted as transcriptional regulation. Interaction of TFs results in specific transcriptional responses. The combinatorial interplay of TFs is referred to as combinatorial or cooperative gene regulation and enables complex regulatory mechanisms within organisms [6,7]. TFs bind DNA via specific transcription factor binding sites (TFBSs), short DNA motifs recognised by the TFs [8]. TFs possess specific protein domains, DNA binding domains, with which they recognise specific TFBSs on the DNA. Examples of DNA binding domains are helix-turn-helix (HTH) domain, zinc finger domain, and basic leucine zipper (bZIP) domain. A primary step towards the elucidation of transcriptional gene regulation is the discovery of the specific TFBSs that TFs can bind.

Numerous high-throughput methods to experimentally determine TFBSs exist, such as Systemic Evolution of Ligands by EXponential enrichment (SELEX, [9,10]), Phage Display (PD, [11]) or Chromatin ImmunoPrecipitation (ChIP, [12,13]), as reviewed recently [14]. Whilst experimental methods to decipher TFBSs are essential, common to all methods are artefacts that come with experimental technologies that impact the accuracy of TFBS determination. In general, experimental methods are time consuming, elaborate to conduct, and

expensive. Currently, it is not feasible to determine experimentally on a global-scale all TFBSs for all known TFs.

Constraints of experimental methods emphasise the importance of complementary computational approaches for TFBS identification. One type of computation-based methods for predicting TFBSs uses position frequency matrices (PFMs) or position weight matrices (PWMs). A PFM/PWM is usually derived from multiple sequence alignments of experimentally verified and aligned DNA binding sites for single TFs or a class of TFs and represents a simple statistical model that reflects the relative distribution and conservation of all nucleotides within the set of binding sites. A PFM can be converted to a PWM by weighting normalised nucleotide frequencies by the background probabilities of the nucleotides in genome-wide DNA (e.g. human DNA) [15]. Databases with curated PFMs/PWMs are available, e.g. TRANSFAC [16,17] (http://www.gene-regulation.com/), and JASPAR [18] (http://jaspar.cgb.ki.se/). Figure 1 shows examples of PFMs and their visual representation in the form of sequence logos produced using enoLOGOS [19]. Sequence logos, displaying the information content, are a good means to gain a quick overview of the nucleotide conservation on specific binding site positions [20]. Differences in the matrices from different databases for the same TFs (see Figure 1), occur due to the utilisation of different binding sites for producing the PFMs/PWMs. However, PFMs/PWMs, disregarding from which source, are used instead of the real binding sites to analyse DNA sequences and

computationally map putative TFBSs with a certain confidence to the DNA. Examples of tools for mapping PFMs/PWMs to DNA sequences include MATCH^{TM} [21], MatInspector [22,23], and ConSite [24].

Computational prediction of TFBSs in DNA sequences has been the focus of various studies [25-36]. Even though the technological possibilities for experimental verification of binding sites for TFs are steadily increasing, computational methods are still indispensable for performing large-scale studies on transcriptional regulation. The prediction accuracy of computational tools for predicting TFBSs is still not sufficient in rivalling experimental methods [15,37,38]. However, while recognizing that computational methods cannot replace laboratory experiments, in the current stage of research they are essential and useful to gain a broad overview of the transcriptional regulatory mechanisms involved in cells when used in synergy.

To enhance general insights into transcriptional regulation within mammals, two distinct biological questions with relevance to the field of research have been investigated in this thesis. One of the gene products discovered within the last two decades, is microRNAs (miRNAs) [39]. MiRNAs are ~22 nucleotides (nt) long non-coding RNAs that influence post-transcriptional regulation, by degrading and repressing the translation of protein-coding mRNA [40,41]. Even though it is already clear that miRNAs have a great influence within cells, not much is known how the transcription of miRNA genes is regulated. A

recent study showed that miRNAs are transcribed by RNA Polymerase II [42], which led to the assumption that they might be controlled in a similar manner as protein-coding genes. The effects of TFs on the transcriptional regulation of miRNA genes are the focus of Chapter 2. Here, computational TFBS analysis is combined with gene expression data of miRNAs and TFs to discover these regulatory mechanisms during a distinct biological process, monocytic differentiation.

Chapter 3 focuses on an integral part of combinatorial gene regulation, the physical interaction of TFs. A method is devised to computationally predict if two TFs interact. The task of predicting interacting TFs can be seen as a subtask of the more general protein-protein interaction (PPI) prediction task. Often, the bottleneck in studies of computational PPI prediction is the acquisition of appropriate data for the numerical representation of the entities involved. The core of the approach in Chapter 3 is a novel form of numerical representation of TFs that incorporates amino acid properties of the primary protein sequence of TFs. An artificial intelligence system is employed to build a model that is able to predict interactions among TFs based on these representations.

Both computational analyses enhance our insights into transcriptional regulation. The aim of this research is to build the groundwork for further in

depth investigations into transcriptional regulation of miRNAs and cooperative transcriptional control of genes.

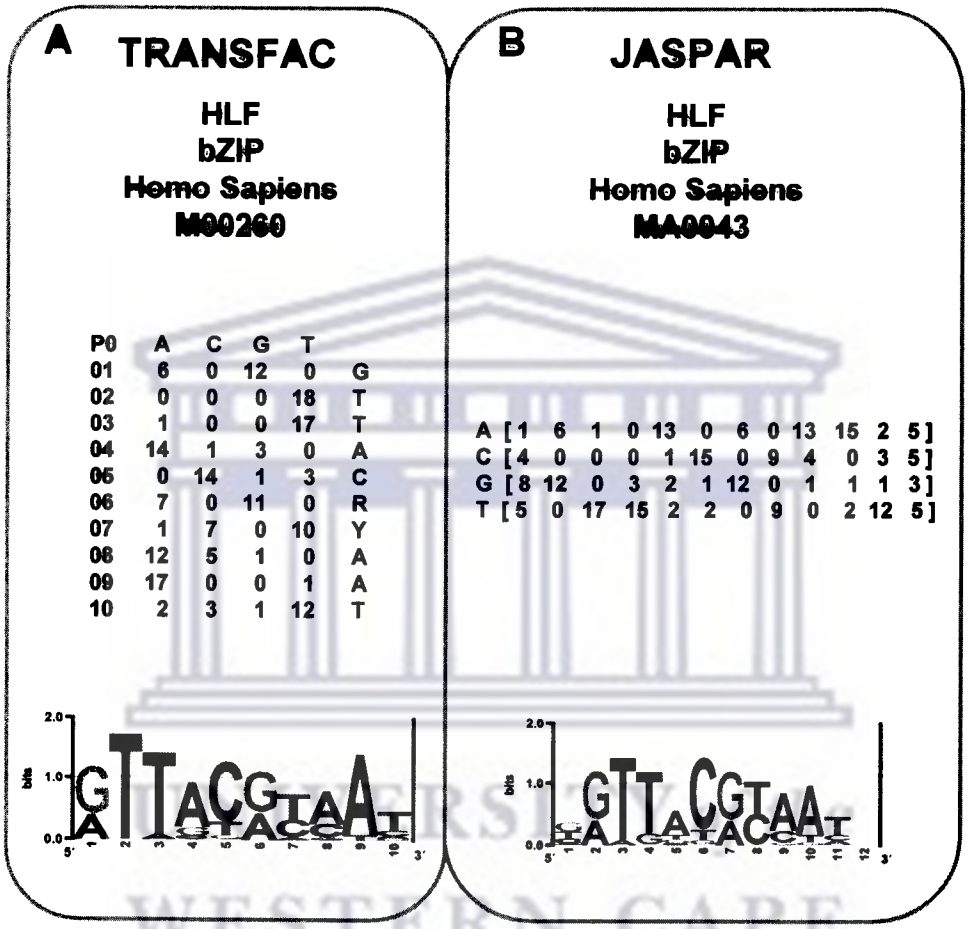


Figure 1. Example of PFMs from TRANSFAC and JASPAR
A/ Example of a PFM from the TRANSFAC Professional database (version 11.4). Shown is a matrix for the HLF transcription factor. The middle panel shows the actual matrix. The matrix consists of 10 positions arranged in a manner where each position is represented by a row in the matrix. The first column contains the actual number of the position in the matrix. Columns two to five contain counts of the nucleotides A, C, G, and T. The last column represents the consensus letter for the corresponding position. The lower panel shows a sequence logo for the matrix, another representation that visualises the conserved information content of each position. **B/** Example of a PFM from the JASPAR database (version 3.0). Shown is a matrix for the HLF transcription factor. The middle panel shows the actual matrix. The matrix consists of 12 positions arranged in a manner where each position is represented by a column in the matrix. Each row represents a nucleotide and gives for each position the number counts of the nucleotide occurrences. The lower panel shows a sequence logo for the matrix.

Chapter 2

Deciphering the Transcriptional Circuitry of MicroRNA Genes Expressed during Human Monocytic Differentiation.

2.1 Abstract

Macrophages are immune cells involved in various biological processes including host defence, homeostasis, differentiation, and organogenesis. Disruption of macrophage function is linked to increased pathogen infection, inflammation and malignant diseases. Differential gene expression observed in monocytic differentiation is primarily regulated by interacting TFs. Current research suggests that miRNAs degrade and repress translation of mRNA, and may also target genes involved in differentiation. The aim of this research is to investigate the transcriptional circuitry regulating miRNA genes expressed during monocytic differentiation.

Analysis of the transcriptional circuitry of miRNA genes during monocytic differentiation was performed computationally using *in vitro* time-course expression data for TFs and miRNAs. A set of TF→miRNA associations was derived from predicted TFBSs within promoter regions of miRNA genes. Time-lagged expression correlation analysis was utilised to evaluate the

TF→miRNA associations. 12 TFs were identified that potentially play a central role in regulating miRNAs throughout monocytic differentiation. Six of these 12 TFs (ATF2, E2F3, HOXA4, NFE2L1, SP3, and YY1) have not previously been described to be important for monocytic differentiation. The remaining six TFs are CEBPB, CREB1, ELK1, NFE2L2, RUNX1, and USF2. The impact on monocytic differentiation through the inferred transcriptional regulation of several miRNAs (miR-21, miR-155, miR-424, and miR-17-92), is presented. This study demonstrates that miRNAs and their transcriptional regulatory control are integral molecular mechanisms during differentiation. Furthermore, it is the first study to decipher on a large-scale, how miRNAs are controlled by TFs during human monocytic differentiation. Subsequently, 12 candidate key controllers (TFs) of miRNAs during human monocytic differentiation were discovered.

2.2 Introduction

The mononuclear phagocyte system is defined as a family of cells comprising of bone marrow progenitors and is derived from hematopoietic stem cells. Hematopoietic stem cells sequentially differentiate into monoblasts, promonocytes, monocytes and terminal macrophage cells [43]. The human monocytic leukemic cell line, THP-1 [44], is an accepted model system utilised to explore molecular events surrounding monocytic differentiation. The chemical Phorbol 12-myristate 13-acetate (PMA) induces the differentiation of monocytic THP-1 cells into macrophages/mature THP-1 cells [45]. Before

inducing differentiation, PMA first inhibits cell growth and blocks THP-1 cells in G1-phase of the cell cycle by up-regulating the expression of $p21^{WAF1/CIP1}$, enhancing binding of the SP1 factor to the $p21^{WAF1/CIP1}$ promoter. PMA inhibition of cell growth is mediated by several signalling pathways such as MAPK and ROS-dependent Raf/MEK/ERK pathway [46]. Human monocytic maturation incorporates lipid and protein metabolic processes together with several G-protein coupled receptors (GPCRs) [47].

Differential gene expression that results in human monocytic differentiation is regulated by numerous interacting TFs [46-48]. Current research suggests that miRNAs target several genes that are differentially expressed in the differentiation process [49]. MiRNAs are ~22 nucleotides (nt) long non-coding RNAs, which play a key role in the repression of translation and degradation of coding mRNA [40,41,50-52]. Several computational tools are available for miRNA target prediction [51,53-56].

Canonical miRNA biogenesis (see Figure 2) begins with the transcription of pri-miRNAs by RNA polymerase II [42,57,58]. The generation of pri-miRNAs by RNA polymerase II suggests that miRNA genes are controlled through the same regulatory machinery as protein-coding genes. These pri-miRNAs are cleaved into 60~70nt pre-miRNAs by the microprocessor complex Drosha (RNase II endonuclease) and DGCR8, a double-stranded RNA binding protein [59,60]. Pre-miRNAs are exported to the cytoplasm with the help of Exportin-5

and its co-factor RanGTP [61]. Dicer, a RNase III endonuclease, cleaves 22-nucleotides at the Drosha cleavage site yielding after strand separation mature miRNA [40,62].

Even though most miRNAs have their own transcriptional units [40], several miRNAs are transcribed together as a single pri-miRNA [63-65]. These clustered miRNAs are thus co-regulated. On the other hand, miRNAs can also be transcribed together with a protein-coding host gene [40]. In addition, a mature miRNA can be produced from several locations in the genome [40,66]. Furthermore, it is not clear how to define the regulatory regions for miRNA genes. Thus, a straightforward analysis of the transcriptional regulation of miRNA genes is difficult. Current research suggests that at transcription start sites (TSSs) of genes, histones are generally trimethylated at lysine 4 residues [67,68]. This has led to a potential definition of promoter regions for miRNAs in human embryonic stem cells using such determined TSSs as reference points [69].

As the transcriptional regulation of miRNAs is not well understood, the focus of this research is the analysis of transcriptional miRNA gene regulation during human monocytic differentiation. Gene expression of miRNAs and TFs was measured prior to PMA stimulation and over a 96 hour time-course, post-PMA stimulation. A general method was utilised to identify miRNAs whose expression levels differed due to PMA stimulation in THP-1 cells. Promoter

regions for these miRNAs were extracted and TFBSs computationally mapped to the promoter sequences. Time-lagged expression correlation analysis [70,71] was employed to evaluate the predicted TF→miRNA associations by combining the *in silico* TFBS analysis with the measured *in vitro* expression data. Similar types of time-lagged expression correlation analyses have been used to either predict or score TF→gene or gene→gene associations [72-74]. From these TF→miRNA associations, 12 TFs were identified to be plausibly playing a central role in regulating miRNAs throughout the considered differentiation process. Six of these 12 TFs (ATF2, E2F3, HOXA4, NFE2L1, SP3, and YY1) have not been previously described as important for monocytic differentiation. The remaining six TFs, CEBPB, CREB1, ELK1, NFE2L2, RUNX1, and USF2, although known to be involved in monocytic differentiation, were not known to play a role in transcriptional regulation of miRNAs in this process. The analysis was concluded by highlighting several inferred regulatory networks that suggest interplay of TFs, miRNAs, and miRNA targets and that are likely to have an impact on the differentiation process.

This research is the first large-scale study that attempts to decipher the transcriptional circuitry that regulates the expression of miRNAs during human monocytic differentiation and identifies potential new avenues for further research.

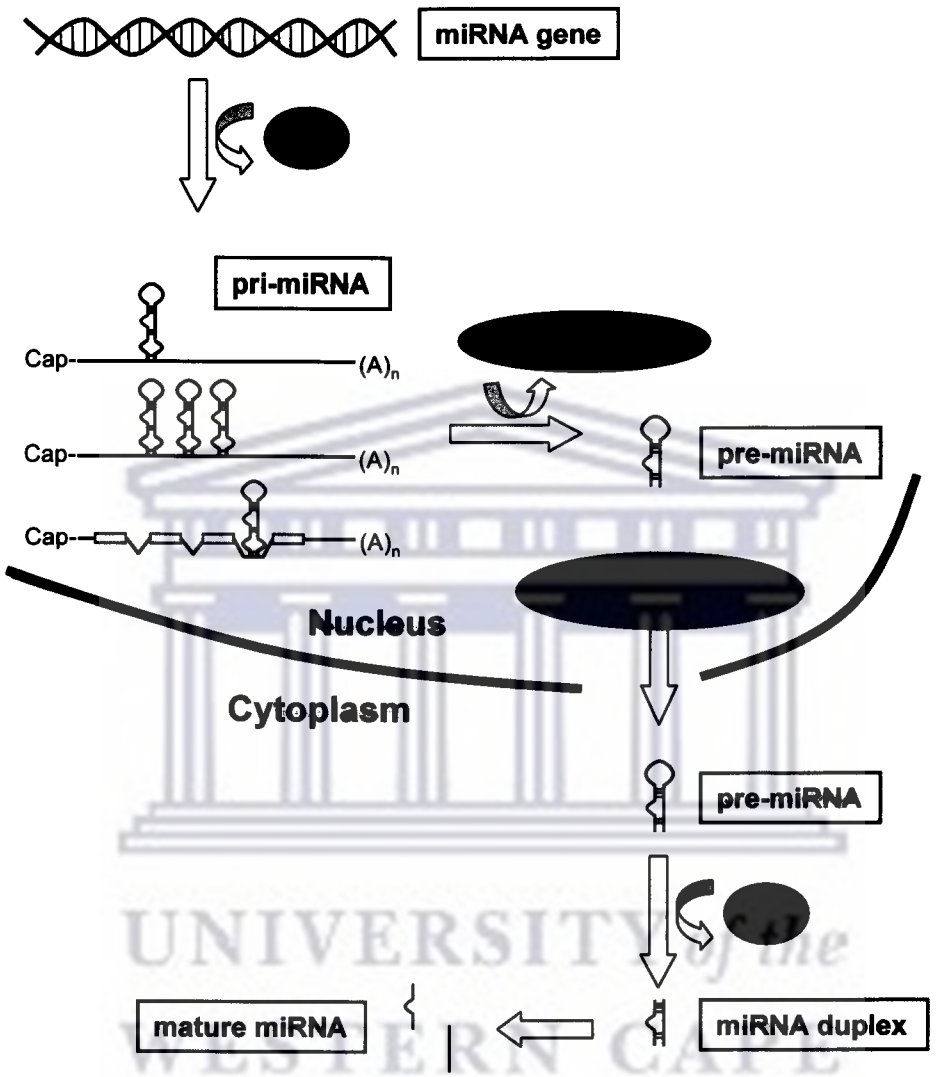


Figure 2. Overview of the miRNA biogenesis
The figure shows an overview of the human miRNA biogenesis. First the miRNA gene gets transcribed into pri-miRNA by RNA polymerase II. Shown are three examples of pri-miRNAs. The first miRNA has its own transcriptional unit. The second pri-miRNA contains a cluster of three miRNAs. The last pri-miRNA contains a miRNA that is transcribed together with a host gene. These pri-miRNAs are cleaved into 60~70nt pre-miRNAs by the microprocessor complex Drosha (RNase II endonuclease) and DGCR8, a double-stranded RNA binding protein. Pre-miRNAs are exported to the cytoplasm with the help of Exportin-5 and its co-factor RanGTP. Dicer, a RNase III endonuclease, cleaves ~22-nucleotides at the Drosha cleavage site yielding after strand separation mature miRNA.

2.3 Methods

All biological experiments to derive the expression data for miRNAs and TFs during PMA stimulation, were done in collaboration by scientists at the RIKEN Omics Science Center in Yokohama, Japan. The experimental data was submitted to all collaborators of the Genome Network Project, specifically FANTOM4, for utilisation. All experimental data derived throughout the project is made public with the publication of the main paper of the FANTOM4 collaboration [48].

2.3.1 miRNA Time-course Expression Data

The miRNA expression profiles were obtained using Agilent's Human miRNA microarrays as described in [75]. Three biological replicates were measured prior to PMA stimulation and post-PMA stimulation at nine time points ranging from one to 96 hours (1hr, 2hr, 4hr, 6hr, 12hr, 24hr, 48hr, 72hr, and 96hr). For the inclusion of a miRNA expression time-series in the analysis two criteria were derived:

i/ Expression of each miRNA should be denoted as "present" in at least one time point. Otherwise it was assumed that the expression series for the miRNA is insignificant.

ii/ For a miRNA, i/ must hold true in at least two of the three biological replicates.

The expression values of different biological replicates for a miRNA that satisfied the criteria above were averaged at each time point to generate one expression series per miRNA. Finally, each expression series was interpolated using piecewise cubic hermite interpolation [76,77] with half an hour steps. In this manner, 193 (0-96hrs) expression values for each individual miRNA expression series were obtained.

2.3.2 Identification of miRNAs Showing Differential Gene Expression

The $\log_2 fc$ was calculated by dividing each expression value of a miRNA by its expression value at zero hour (control) and taking the logarithm of base two ($\log_2$) of that ratio. A miRNA was considered to be influenced by the PMA stimulation in the differentiation process if:

i/ In at least one time point t its $\log_2 fc > 1$ or $\log_2 fc < -1$.

ii/ At any time point t where i/ holds true, the absolute difference d_t in expression e_t at time point t and the expression e_0 at zero hours must be greater than 0.1.

2.3.3 Transcription Factor Time-course Gene Expression Data

The TF expression profiles were obtained using qRT-PCR as described in [78]. Two biological replicates were measured prior to PMA stimulation and in nine time points post-PMA stimulation (1hr, 2hr, 4hr, 6hr, 12hr, 24hr, 48hr, 72hr, and 96hr). Primer design, RNA preparation, and cDNA synthesis were performed analogously to [48]. Normalization of the expression data of both replicates was done as described in [78,79].

All expression series for a TF that had available expression data within two biological replicates were averaged over the respective biological replicates to produce one series of expression values per TF. Finally, each expression series was interpolated in half an hour steps using piecewise cubic hermite interpolation. Thus, 193 (0-96hrs) expression values for each individual TF expression series were obtained.

2.3.4 Defining Promoter Regions of miRNAs

The definition of miRNA promoters was adopted from [69]. Each of the promoter regions had a score associated (as defined in [69]) that represents the confidence of dealing with a genuine regulatory region. All promoter regions with a score greater or equal to zero were extracted. The coordinates of the promoter regions were translated from the Human genome build 17 (hg17) to

the Human genome build 18 (hg18) [80] using the UCSC liftover program [81] (see Table 1).

2.3.5 Transcription Factor Binding Site Analysis of miRNA Promoter Regions

TFBSs were mapped to the promoter region of the miRNAs with the MATCH™ program [21] utilising 522 mammalian matrices of the TRANSFAC Professional Database (version 11.4) with their corresponding minimum false positive threshold profiles. Since TRANSFAC matrices are frequently associated with several TFs whose binding sites were used in building these matrices, each matrix was associated to all respective TFs (that have an Entrez Gene identifier associated). For example, several members of the JUN-FOS family (JUN, JUNB, JUND, FOS, FOSB, etc.) can be associated to matrix M00517. Binding sites of these TFs have been utilised to create this matrix. Thus, all of the TFs might be able to bind the TFBS predicted by the matrix.

2.3.6 Weighting Associations Using Pearson Correlation

For each of the predicted TF→miRNA associations, scores (*PCCs*) were calculated as an indicator of how reliable the predicted association is, and as a measure of the strength of the association within the context of monocytic differentiation. The expression data for TFs and mature miRNAs during monocytic differentiation were utilised to calculate the best time-lagged

correlation for a TF→miRNA association. The time-lagged expression correlation analysis calculates a *PCC* between the TF expression and the time-shifted mature miRNA expression at different time-delays in order to take the influence of the TF on the miRNA transcription over time into account. The method selects the time-delay that maximizes the absolute value of *PCC* between the expression of the TF and that of the mature miRNA. The associations between pre-miRNA and the mature miRNA were extracted using the miRBase sequence database (version 10.1) [55,56,82] (http://microrna.sanger.ac.uk/).

For each predicted TF→miRNA association, where the miRNA does not share the same promoter with other miRNAs (i.e. not in a cluster), the *PCC* was calculated as follows:

i/ Identify the time-shift s_t. This is the time-shift where the absolute value of the *PCC* between the expression of the TF and the respective mature miRNA is maximal. The *P*CC was calculated for time-shifts ranging from 0.5 hour to six hours in intervals of half an hour.

ii/ The *PCC* for the association was calculated as *PCC* of the expression of TF and mature miRNA at the time-shift s_t found in i/.

If a miRNA appears in a cluster with other miRNAs on the genome, then the predicted TF in the promoter of that cluster is associated to each of the

respective miRNAs. Since the cluster is transcribed as one primary transcript, it is assumed that a TF regulates each miRNA within the cluster with the same time-shift. Thus, one common time-shift s_t for the considered TF and all miRNAs within the cluster was calculated. The time-shift s_t was calculated as follows:

i/ The *PCC* of expression between the TF and each miRNA in the cluster was calculated for each considered time-shift (0.5 hour to six hours).

ii/ The average of all *PCC*s derived in i/ was calculated for each time-shift (0.5 hour to six hours). As a criterion for inclusion, the calculated *PCC*s for all associations should have the same sign.

iii/ If ii/ could not be calculated at any time-shift (due to the sign rule), it was not assumed that the TF *X* regulates any miRNA in that cluster and all *X*→miRNA associations of that cluster were discarded.

iv/ If not iii/, then the time-shift s_t was determined as the time-shift that maximizes the average calculated in ii/.

PCC of one TF→miRNA association where the miRNA is part of a cluster forms the *PCC* of expression of the TF and the respective mature miRNA at the determined time-shift s_t for the TF and the cluster. If a pre-miRNA is associated to more than one mature miRNA from its 5' and 3' arm, then the *PCC* was calculated independently for each mature miRNA and the maximum absolute *PCC* was chosen.

2.3.7 Target Predictions of miRNAs

The target gene predictions of human miRNAs were gathered from four public available databases for miRNA target predictions: microRNA.org version 4 [53] (http://www.microRNA.org), TargetScan version 4.2 [51] (http://www.targetscan.org/), miRBase version 5 [55,56], and EIMMO2 with a cut-off value greater than 0.5 [54] (http://www.mirz.unibas.ch/ElMMo2/). All target gene identifiers utilised in the respective databases were converted to Entrez Gene identifiers using BioMart [83] (http://www.ensembl.org/biomart/martview). If this was not possible the prediction was discarded. Only predictions that were present in at least three out of the four databases were considered.

2.3.8 Network Graphics and Pathway Analysis

All regulatory network graphics in the figures presented in the "Results and Discussion" were produced with the help of Cytoscape [84] and all pathway analyses were based on the Kyoto Encyclopaedia of Genes and Genomes (KEGG; [85]; http://www.genome.jp/kegg/) using the Database for Annotation, Visualization and Integrated Discovery (DAVID; [86]; http://david.abcc.ncifcrf.gov/).

2.4 Results and Discussion

First, the miRNA expression data was analysed to identify miRNAs that are differentially expressed due to the PMA stimulation. For these, promoter regions were extracted and TFBSs predicted. Subsequently, each predicted TF→miRNA association was scored using a time-lagged expression correlation analysis to get a measure of reliability for the predicted associations. Afterwards, TFs that are likely to play a central role in regulating miRNAs during the monocytic differentiation process were statistically identified. Finally, the predicted transcriptional regulation of several miRNAs and their potential influence on the differentiation process were investigated. Figure 3 gives an overview of the analysis steps.

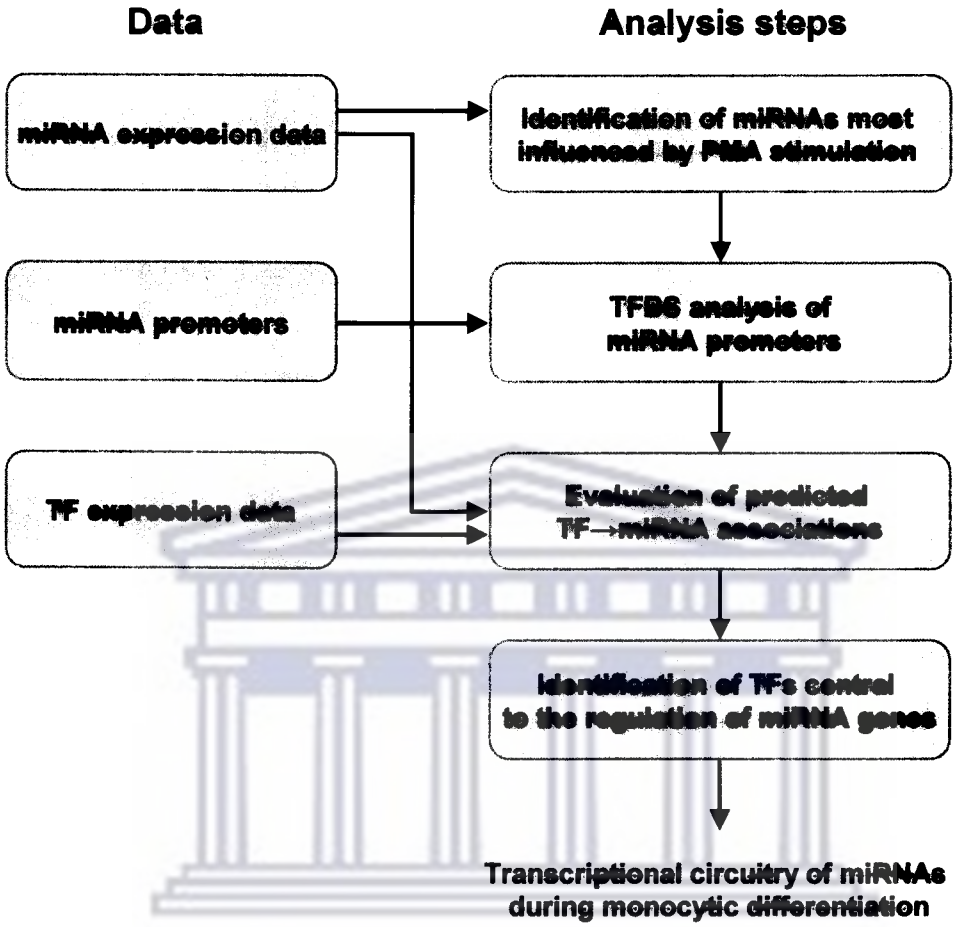


Figure 3. Overview of the analysis
The figure shows the analysis steps (blue/green boxes). In addition, the figure shows the data (red boxes) that have been utilised within individual analysis steps. In total five analysis steps have been conducted. First, the miRNA expression data was utilised to find the miRNAs that show differential expression throughout the differentiation process. Promoter regions for these miRNAs were extracted and TFBSs computationally mapped to these regulatory regions. The expression data from both miRNAs and TFs was used to score the predicted TF→miRNA associations. Subsequently, the TFs that are enriched in the set of associations with highest *PCC* between TF and miRNA expression were identified. Finally, for certain miRNAs their transcriptional regulations and impact on monocytic differentiation have been investigated.

2.4.1 Identification of miRNAs Most Influenced by the PMA Stimulation

Three biological replicates of miRNA expression data provided measured expression levels at nine time-points post-PMA stimuli and a zero hour control prior to PMA stimulation (see Methods). Two criteria had to be met for the inclusion of a miRNA expression time-series ('expression series' in further text) in the analysis:

i/ Expression of the miRNA had to be denoted as "present" in at least one time point, otherwise it was assumed that the expression series for the miRNA is invalid. In this manner, 155, 238, and 191 miRNAs and associated expression series for the first, second, and third replicate were identified.

ii/ For a miRNA, i/ must hold true in at least two of the three biological replicates.

The expression values of different biological replicates for a miRNA that satisfy the criteria were averaged at each time point to generate one expression series per miRNA. This resulted in expression series for 187 miRNAs (see Methods).

The $\log_2 fc$ (*fc* standing for fold-change relative to time point zero) for each of the 187 identified miRNAs at each measured time point was calculated (see Methods), to identify the set of miRNAs that show differential expression level changes due to PMA stimulation.. A miRNA was considered to be influenced by PMA stimulation if its $\log_2 fc > 1$ or $\log_2 fc < -1$ at any measured time point post-PMA stimulation (see Figure 4). A total of 81 miRNAs satisfied this criterion. The majority of the miRNA expression levels do not change significantly over time and are confined within the selected threshold values (see Figure 4). To determine which miRNAs deviated from the baseline expression level, the following steps were implemented:

i/ For each time point t where $\log_2 fc > 1$ or $\log_2 fc < -1$ was satisfied for a miRNA, the difference d_t of the expression e_t at time point t and its expression e_0 at the zero time point was calculated.

ii/ The miRNAs for which $\text{abs}(d_t) > 0.1$ in at least one time point were sub-selected.

A set of 53 miRNAs whose expression is most likely affected by the PMA stimulation met the implemented criteria.

The *fc* does not take the level of gene expression into account. It is important to note that miRNAs that may have very high expression levels and whose expression level only changes minimally over time might have a strong

biological impact, even though this is not reflected by variation in the expression levels. The approach utilised here, based on *fc* is not able to identify such cases. On the other hand, miRNAs with very low expression levels might have high *fc* values, which might suggest a strong biological impact, even though this may be arguable since the absolute changes in expression levels could be very small. Hence, a second threshold for the difference in expression values of 0.1 was introduced, even though there are no accepted standardised thresholds for such an analysis. Thus, all miRNAs whose *fc* based on their expression series suggests an impact through the PMA stimulation but were the absolute change of expression values is below 0.1 are not considered to be affected by the PMA stimulation.

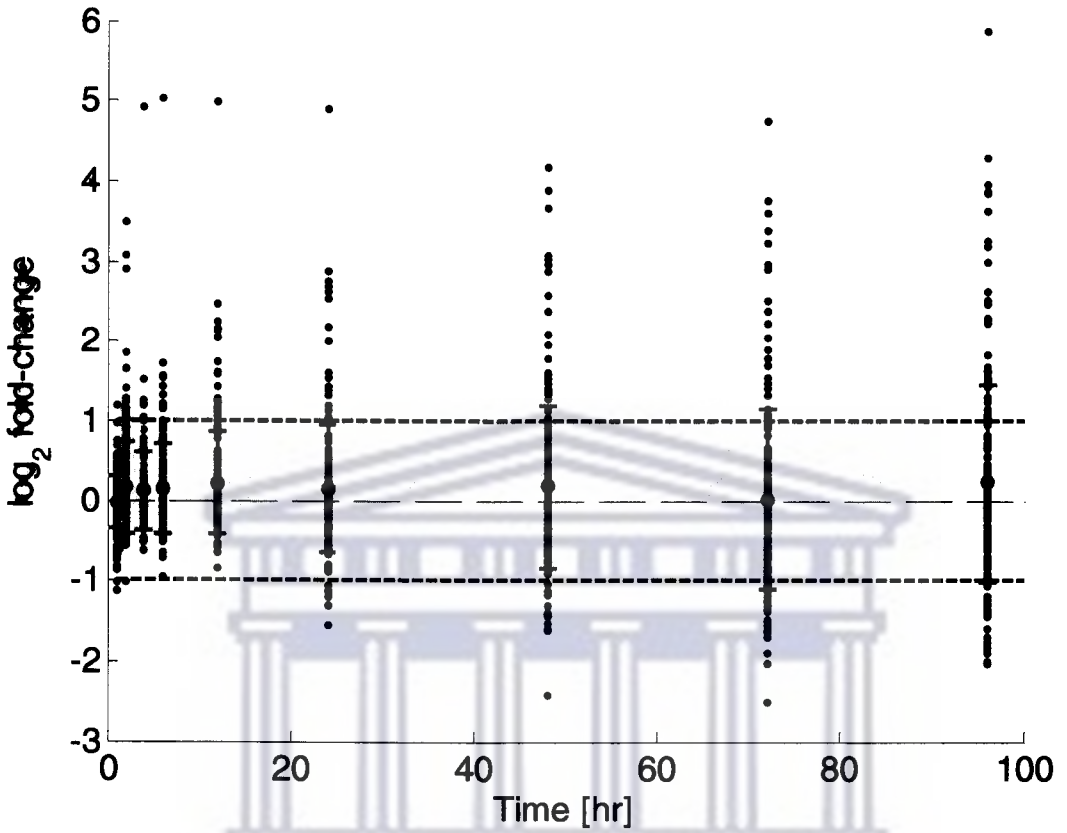


Figure 4. Selecting PMA induced miRNAs
For all measured time-points after PMA stimuli the $\log_2$ *fc* of the averaged expression set for all 187 selected mature miRNAs is presented (black dots). Each dot represents a $\log_2$ *fc* of a single miRNA at the considered time point relative to the zero time point. The red dashed lines mark the $\log_2$ *fc* of 1 and -1 that were utilised as a cut-off for miRNAs (see main text). The figure shows in addition the mean (blue dot) and the standard deviation of all $\log_2$ *fc* values from the 187 miRNAs at the considered time point (blue error bars). Grey dashed lines indicate individual miRNA expression series. The figure shows that the majority of the miRNA expression levels do not change significantly over time and are confined within the selected threshold values.

2.4.2 Transcription Factor Binding Site Analysis of miRNA Promoter Regions

Promoter regions of miRNAs are regions of DNA where TFs bind to regulate the transcription of miRNA genes into pri-miRNAs. A pri-miRNA can be associated with several promoter regions derived from different TSSs. The transcriptional control of TFs is towards the pri-miRNA that can be cleaved into several pre-miRNAs [87]. Thus, miRNAs that form such clusters are considered to be generally regulated in the same manner.

Marson *et al.* [69] defined promoter regions of miRNAs using TSSs that were determined using trimethylated histones. These form the basis for the promoter regions analysed in this study. For 34 of the 53 earlier identified mature miRNAs, 38 promoter regions for 37 associated miRNAs were extracted (see Methods and Table 1).

The TRANSFAC Professional database [16,17] was used to map TFBSs to the 38 promoters. TRANSFAC's 522 mammalian minimum false positive matrix profiles of binding sites were mapped to the promoter regions (see Methods). These matrices, which correspond to the predicted TFBSs, are associated with TFs that possibly bind these TFBSs (see Methods). By mapping the matrices to their corresponding TFs, 5,788 unique TF→miRNA associations for 673 TFs and 37 miRNAs were obtained.

Table 1. miRNA Promoter Regions

miRNA	Promoter regions (HG18)
hsa-mir-106b	chr7_99537263_99537463_-
hsa-mir-595	chr7_158073079_158073279_-
hsa-mir-21	chr17_55138283_55141202_+
hsa-mir-22	chr17_1566155_1566355_-
hsa-mir-23a	chr19_13807762_13808928_- chr19_13818427_13819944_-
hsa-mir-222	chrX_45497997_45498485_-
hsa-mir-181a-1	chr9_126460466_126460666_+
hsa-mir-181a-2	chr1_197173071_197173271_-
hsa-mir-19a	chr13_90797974_90798174_+
hsa-mir-503	chrX_133505836_133508763_-
hsa-mir-27a	chr19_13807762_13808928_- chr19_13818427_13819944_-
hsa-mir-34a	chr1_9164884_9165084_-
hsa-mir-221	chrX_45497997_45498485_-
hsa-mir-29a	chr7_130236779_130237964_-
hsa-mir-542	chrX_133505836_133508763_-
hsa-mir-20b	chrX_133131533_133136274_-
hsa-mir-29b-1 hsa-mir-17	chr7_130236779_130237964_- chr13_90797974_90798174_+
hsa-mir-132	chr17_1899412_1901670_-
hsa-mir-660	chrX_49573865_49574065_+
hsa-mir-9-1	chr1_154657781_154657981_- chr1_154665745_154665945_-
hsa-mir-9-3	chr15_87712233_87712433_+

hsa-mir-9-2	chr5_87997094_88000044_- chr5_88016256_88016456_-
hsa-mir-155	chr21_25867620_25868098_+ chr21_25856186_25856386_+
hsa-mir-210	chr11_558302_558502_-
hsa-mir-24-1	chr9_96528703_96528903_+ chr9_96806816_96807016_+
hsa-mir-425	chr3_49178672_49178872_- chr3_49030949_49031149_-
hsa-mir-424	chrX_133505836_133508763_-
hsa-mir-18b	chrX_133131533_133136274_-
hsa-mir-345	chr14_99840674_99844301_+
hsa-mir-18a	chr13_90797974_90798174_+
hsa-mir-137	chr1_98284260_98284460_-
hsa-mir-24-2	chr19_13807762_13808928_- chr19_13818427_13819944_-
hsa-mir-365-2	chr17_26909910_26910109_+ chr17_26910186_26913533_+
hsa-mir-365-1	chr16_14309182_14309534_+ chr16_14309748_14312532_+ chr16_14303600_14303800_+
hsa-mir-133b	chr6_52097779_52098124_+
hsa-mir-146a	chr5_159827731_159827931_+ chr5_159826351_159827231_+

The first column contains the miRNA identifier. The second column contains the chromosomal positions (format: chromosome_start_stop_strand) of the associated promoter regions. Note that several miRNAs can be associated with the same promoter region as they can be transcribed together as a cluster.

2.4.3 Evaluation of Predicted TF→miRNA Associations

Each predicted TF→miRNA association was evaluated and scored to obtain the most accurate picture of miRNA gene regulation during human monocytic differentiation. The result of this evaluation relates to the confidence that a TF→miRNA association is genuine. Evaluation was based on time-lagged expression correlation between the gene expression series of the TF and that of the mature miRNA (see Methods). Expression data for miRNAs and TFs was measured in human THP-1 cells prior PMA stimulus at one time point and post-PMA stimulus at non-equidistant time points up to 96 hours.

The expression series for each of the 34 mature miRNAs was interpolated using half an hour steps (see Methods and Appendix I X1). The TF qRT-PCR expression series were averaged over the two biological replicates at the same time points, and interpolated in concordance with the miRNA expression data. In this manner, expression series for 2,197 TFs were derived (see Methods).

The TF→miRNA associations were inferred from TFBS analysis of promoter regions of miRNA genes. From the predicted 5,788 TF→miRNA associations, all associations were discarded for which no expression data for the TF in the above mentioned averaged expression set exists. The Pearson correlation coefficient (*PCC*) was calculated for each TF→miRNA associations using time-lagged correlation analysis and the interpolated expression series for TFs

and mature miRNAs. Hence, a set of 1,989 TF→miRNA associations (see Appendix I X2) for 37 miRNAs and 258 TFs (see Appendix I X3), each associated with a *PCC* value were derived (see Methods). The number of TF→miRNA associations that have *PCC*s equal to or greater than selected thresholds is depicted in Figure 5A. As expected, the number of associations steadily decreases with increasingly stringent *PCC* thresholds.

Previous research has demonstrated that the regulatory effects of a TF on its target genes are not instantaneous but occur with a time-lag/shift [88-90]. Unfortunately, the correct time-shifts are undetermined. Hence, time-shifts in a range from 0.5 hours to six hours were incorporated to allow for a sufficient time-delay for the regulation by the TF to exert an effect on the transcription of its target miRNA gene. For each of the 1,989 TF→miRNA associations, the most favourable time-shift was calculated and with this, the time-lagged *PCC* of expression as the score for the association (see Methods). It is assumed that the value of the *PCC* relates to the confidence of an association being genuine and hence, plays an important role in the differentiation process. The higher the absolute value of the *PCC* for an association, the more reliable the association is to be. For each miRNA/miRNA-cluster and its regulating TFs, the maximum *PCC*s were calculated individually (see Methods). Other approaches considered all TFs that regulate a gene to extract a common time-shift for all TFs and the gene [73], or compute the best time-shift depending on known examples of regulation [71]. To date, very few experimentally verified

examples of TFs that regulate miRNAs are known, thus a model to introduce the "correct" time-shift cannot be inferred. Furthermore, certain miRNAs were predicted to be clustered and to share common promoter regions. Hence, a time-shift common to all miRNAs in a cluster was calculated for each of the associated TFs. As a criterion, common time-shifts were only taken into account if all *PCC*s between the TF and all miRNAs that form the cluster had the same sign (e.g. all positive or all negative) to avoid contradicting effects of the same TF on different miRNAs of the cluster. In the cases where there was sign disagreement in the *PCC*s the TF→miRNA associations were discarded.

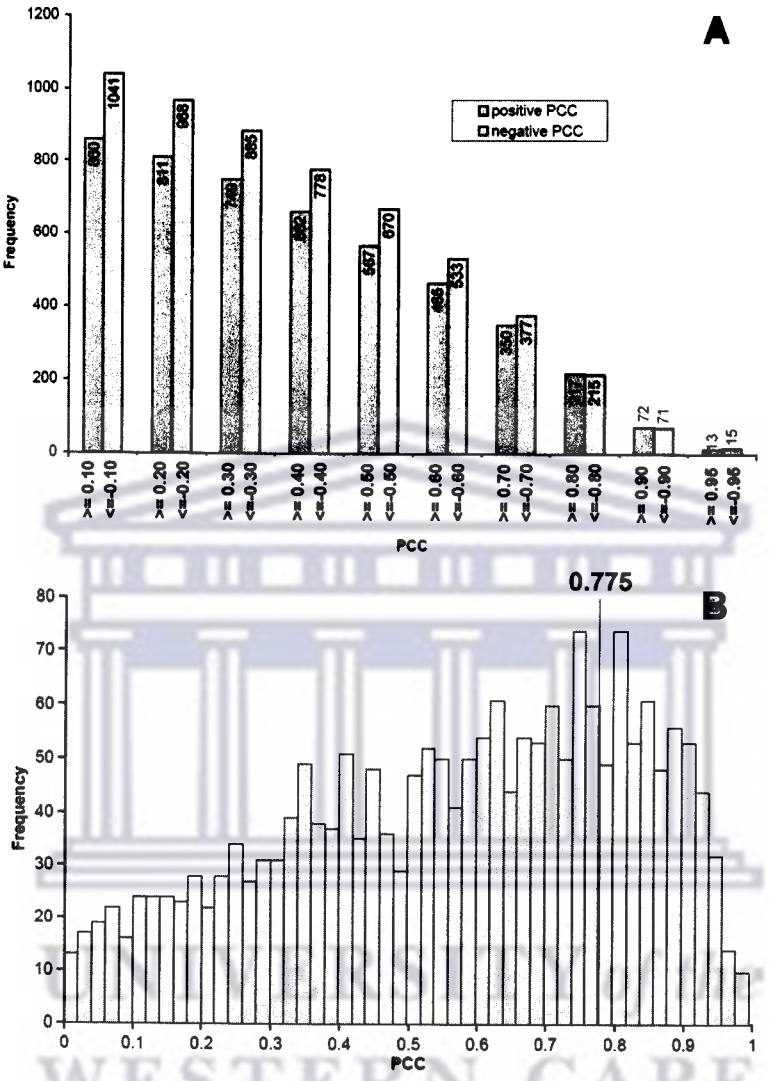


Figure 5. TF→miRNA Associations and their inferred PCCs

A/ Depicted is the number of TF→miRNA associations that have a score equal or greater than specific *PCC*s. The blue blocks indicate the number of associations that have a positive *PCC* greater or equal to the positive value indicated on the x-axis. The red blocks indicate the number of associations with a negative PCC smaller or equal to the negative value indicated on the x-axis. As expected, the number of associations steadily decreases with increasing absolute *PCC*. **B/** Presented is the distribution of the absolute value of the calculated *PCC*s for all 1,989 TF→miRNA associations. The red line indicates the cut-off value that was utilised to select the top quartile of the associations. The distribution is not normal distributed, but skewed towards higher *PCC*s resulting from the chosen method of time-shifts, which favours higher *PCC*s over lower ones.

2.4.4 Identification of Transcription Factors Central to the Regulation of miRNA Genes

The TFs being part of the TF→miRNA associations that have the highest absolute *PCC* were analysed to determine the TFs that have the most influence on miRNAs during the differentiation process. 1,989 TF→miRNA associations were ranked according to the absolute value of their corresponding *PCC*s. The upper quartile (with the highest absolute *PCC*s) was selected from the ranked associations. In this manner, 498 associations were selected, each with an absolute *PCC* greater than 0.775 (see Figure 5B). The 498 associations are formed by 111 unique TFs and 35 unique miRNAs. TFs that appear significantly more often in the upper quartile of associations are assumed to more likely play a central role in regulating miRNAs during the differentiation process. A one-sided Fisher's exact test was conducted to calculate the Bonferroni-corrected p-value for enrichment of each TF in the upper quartile subset of 498 associations, in contrast to the number of occurrences of the TF in the remaining set of associations (1,491). The correction factor utilised for the Bonferroni-correction is the number of unique TFs (258) in the complete set of all associations (1,989). 12 TFs were identified to be statistically significantly enriched in the set of 498 associations with a corrected p-value smaller than 0.01 (see Table 2).

Six of these 12 TFs (ATF2, E2F3, HOXA4, NFE2L1, SP3, and YY1) have not been previously described with regards to monocytic differentiation. The

remaining six TFs (CEBPB [91], CREB1 [92], ELK1 [93], NFE2L2 [94], RUNX1 [91], and USF2 [95]) are known to play a role within monocytic differentiation, but not explicitly as regulators of miRNAs in monocytic differentiation .

The approach implemented attempted to identify the most dominant TFs that putatively regulate miRNAs from the selected subset of the TF→miRNA associations with highest *PCCs*. The complete set of 1,989 TF→miRNA associations consists of many associations with a low *PCC* (see Figure 5). Associations with the highest *PCC*s were sub-selected, in order to be able to focus on associations that are most likely to be genuine. Simultaneously, to be able to deduce the general participants in the transcriptional regulation process of miRNAs, it is important not to focus on too few associations. Consequently, the upper quartile of TF→miRNA associations ranked, based on decreasing absolute values of *PCC*, were selected as a reasonable compromise between sensitivity and specificity.

Table 2. TFs Predicted to have a Central Role in regulating miRNAs

Gene Symbol	Gene ID	Hits in subset	Number of associations in subset	Total number of hits	Total number of associations	p-Value	p-Value (Bonferroni-corrected)
CREB1	1385	18	498	20	1989	1.33E-09	3.43E-07
ATF2	1386	15		17		6.56E-08	1.69E-05
SP3	6670	13		14		1.46E-07	3.76E-05
NFE2L2	4780	12		13		5.52E-07	1.42E-04
NFE2L1	4779	10		10		9.04E-07	2.33E-04
YY1	7528	10		11		7.72E-06	1.99E-03
CEBPB	1051	10		11		7.72E-06	1.99E-03
RUNX1	861	11		13		1.04E-05	2.69E-03
USF2	7392	9		10		2.85E-05	7.35E-03
E2F3	1871	13		18		3.18E-05	8.21E-03
ELK1	2002	10		12		3.59E-05	9.27E-03
HOXA4	3201	11		14		3.77E-05	9.74E-03

The TFs that have been identified through statistical analysis to be statistically enriched (corrected p-value < 0.01) within the upper quartile of predicted TF→ miRNA associations (see main text) are presented. The correction factor utilised for the Bonferroni-correction is the number of unique TFs in the complete set of predicted TF→miRNA associations (258).

2.4.5 Transcriptional Circuitry of miRNAs during Monocytic Differentiation

To shed light on a portion of the molecular underpinnings of monocytic differentiation, the TF→miRNA associations for miRNAs that have been described earlier to be affected by PMA stimulation are discussed. In this manner, it can be determined whether or not the findings in this analysis correspond to published scientific results and further introduce novel TF→miRNA associations. An overview of the regulatory effects of the TF subset (Table 2) on the miRNAs is depicted in Figure 6. Figure 6 shows each TF→miRNA association, from within the subset of the upper quartile of associations for the 12 TFs, in form of a coloured dot in a heat-map created using the TIGR Multiexperiment Viewer (version 4.3) (TMEV, [96,97]). Certain clusters of miRNAs that are regulated by the same set of TFs can be observed (see Figure 6). In the following results and discussion, the focus is mainly on the upper quartile of TF→miRNA associations and on the TFs illustrated in Figure 6 that were identified to be central to monocytic differentiation. For the sake of completeness, several TFs that are known to be regulators of certain miRNAs are discussed as well, even though they might not appear in the set of TF→miRNA associations with highest *PCCs*. Subsets of miRNAs that have literature-based support for their expression during PMA-induced differentiation are discussed.

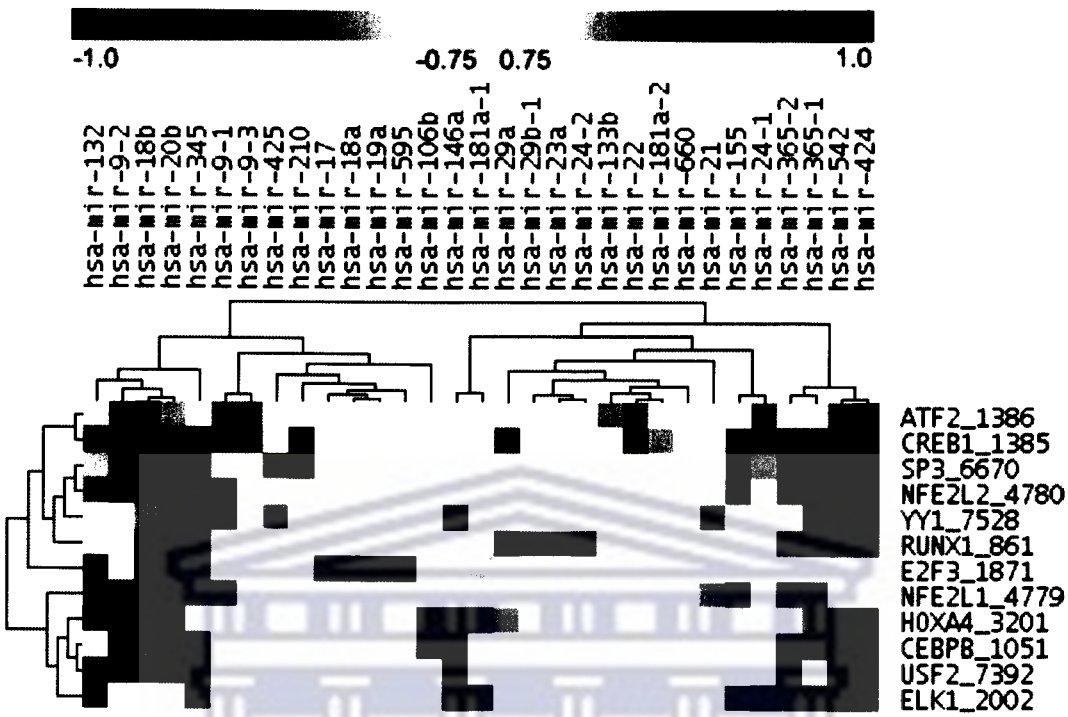


Figure 6. Overview of 12 TFs and their Regulatory Effect on miRNAs
The figure shows a heat-map, with miRNAs on the x-axis and TFs on the y-axis. The TF names on the y-axis are composed of the Entrez Gene symbol and Entrez Gene identifier, separated by "_". A coloured square indicates the value of the *PCC* between a TF and a mature miRNA where the TF has been predicted to regulate the corresponding miRNA. A red square indicates a negative *PCC* whereas a blue square indicates a positive *PCC*. The figure only shows associations where a TF identified to be central to monocytic differentiation takes part and only associations from the top quartile with highest *PCC*. A white dot in the figure does not necessarily indicate a non-association. A possible association would have a *PCC* that prevented its inclusion in the top associations and is thus not shown. The heat-map has been clustered using hierarchical clustering with average linkage and Euclidian distance as the distance measure.

2.4.5.1 miR-21

Fugita *et al.* demonstrated that mir-21 is expressed during PMA induced differentiation in the human promyelocytic leukaemia cell line, HL-60 [98]. The expression data utilised in the current analysis demonstrate that miR-21 is up-regulated during the differentiation process (see Figure 7C). The correlation data suggest that several of the 12 TFs (see Table 2), which were identified as being central to the monocytic differentiation process bind in the promoter region of miR-21 (YY1, NFE2L2, ATF2 and NFE2L1, see Figure 6). Additionally, the binding of TFs, AP-1/c-jun, and c-fos to the promoter region of mir-21 has been shown via ChIP in the HL-60 cell line four hours post-PMA induction [98]. The TFBS analysis suggests the binding of several members of the JUN-FOS family (JUN, JUNB, JUND, FOS, FOSB, FOSL1, and FOSL2) to the promoter region of mir-21, even though they do not appear in the upper quartile of TF→miRNA associations (see Appendix I X2). Expression data for the JUN family members displayed continued up-regulation for 96 hours, whereas FOS family members, with exception of FOSL1, were down-regulated after 4 hours (see Figure 7B). AP-1/c-jun form a complex with the JUN-FOS family members during transcription, and AP-1/c-jun is known to be activated by PMA-induction which is supported by the measured expression data (see Appendix I X3) [99]. Fugita *et al.* also demonstrated that AP-1 and SPI1 synergistically mediate the transcriptional process [98]. TFBS analysis predicts a SPI1 binding site in the promoter region of the mir-21 gene. The time-lagged

correlation analysis also demonstrates that SPI1 is highly correlated to miR-21 (*PCC* = 0.798; see Figures 7B and 7C).

miR-21 has been found to have an anti-apoptotic function and targets tumour suppressor genes, like PTEN in human hepatocellular cancer cells [100], tropomyosin 1 (TPM1), PDCD4, and maspin gene in the human breast cancer cell line, MDA-MB-231 [101]. The miR-21's predicted targets (see Methods) were found to be primarily involved in pathways such as TGF-β signalling pathway, MAPK signalling pathway and the JAK-STAT signalling pathway (see Figure 7A and Methods). The TGF-β signalling pathway and MAPK signalling pathway are primarily involved in differentiation, proliferation, apoptosis and developmental processes, while the JAK-STAT signalling pathway is involved in immune responses. Several TFs such as ATF2, FOS, JUN and JUND included in the predicted TF→mir-21 associations are involved in the MAPK signalling pathway (see Figure 7A).

Time-lagged expression correlation analysis demonstrates that NFE2L1 and SPI1 are highly correlated to miR-21, as opposed to YY1, NFE2L2, and ATF2, which have negative *PCC*s (see Figure 6). Besides JUN-FOS family members and SPI1 that are known to regulate miR-21, the results suggest a novel NFE2L1→miR-21 association, which seems to play an important role in monocytic differentiation (see Figure 7A).

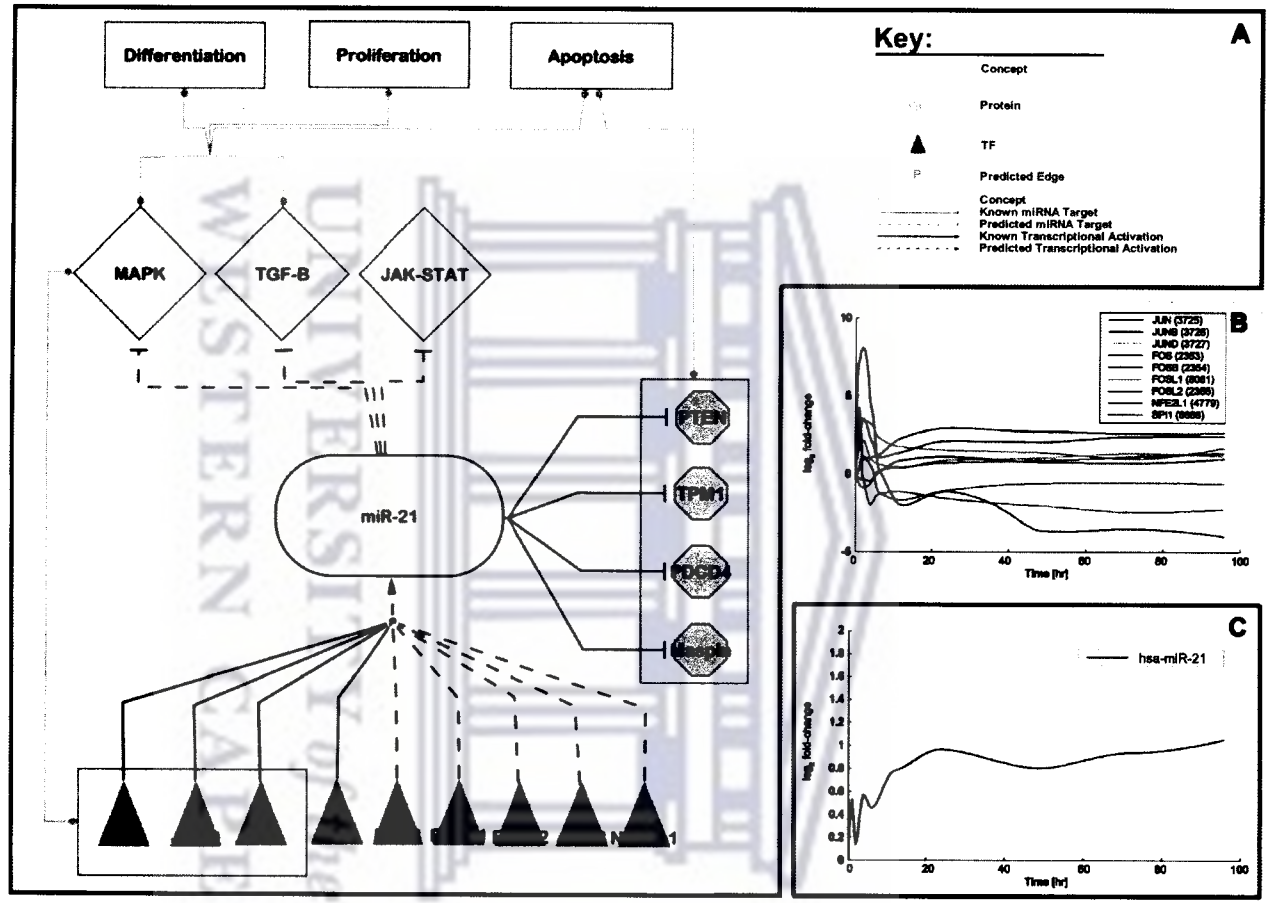


Figure 7. Involvement of miR-21 in monocytic differentiation.
A/ Depicted are the predicted regulations of miR-21 and its involvement in monocytic differentiation. **B**/ Shown is the $\log_2$ *fc* over time of the interpolated expression data of selected TFs that are predicted to regulate miR-21. **C**/ Shown is the $\log_2$ *fc* over time of the interpolated expression data for miR-21.

2.4.5.2 miR-424

Rosa *et al.* reported that mir-424 is expressed during PMA-induced differentiation [102]. Additionally, mir-424 is transcribed by SPI1 in CD34+ human cord blood cells and CEBPA (C/EBPα) blocks SPI1 induced dendritic cell development from CD34+ human cord blood cells by displacing the coactivator c-Jun [102,103]. The up-regulation of miR-424 (see Figure 8C) leads to the repression of NFIA which allows for the activation of differentiation specific genes such as M-CSFr (CSF1R) [102]. Furthermore, the pre-mir-424 is transcribed together with pre-mir-503 and pre-mir-542 as one transcript. These pre-miRNAs form the mature miRNAs miR-424, miR-503, miR-542-5p, and miR-542-3p. The TFBS analysis suggests that several of the 12 TFs (see Table 2), which were identified as being central to the considered differentiation process bind in the promoter region of miR-424 (RUNX1, E2F3, SP3, YY1, NFE2L2, CREB1, ATF2, USF2, ELK1, CEBPB and HOXA4; see Figure 6). As mir-424 and mir-542 are regulated by the same TFs, they form tight clusters in the heat-map (see Figure 6). However, mir-503, part of the same cluster and thus subject to the same regulations, is not present in Figure 6. This is a consequence of the expression data obtained for miR-503 causing the *PCC*s for the TF→miRNA associations to decrease and thus not being part of the top quartile of associations (see Appendix I X2). A SPI1 and CEBPA binding site were predicted in the promoter region of these clustered miRNAs, which corresponds to the findings reported by Rosa *et al.* [102]. SPI1 is positively correlated to miR-424 and CEBPA negatively correlated to miR-

424. In addition, both associations are not within the top quartile of associations with highest *PCCs*. Nevertheless, these observations indicate that SPI1 enhances the expression of the mir-424 cluster and might work in conjunction with other identified TFs to influence the miRNA's transcription.

The TF expression data (see Appendix I X3) confirms the down-regulation of the miR-424 target, NFIA shown by Rosa *et al.* [102]. However, NFIA is down-regulated (~-2.46-fold) three hours post-PMA induction, but recovers at 12 hours (see Appendix I X3). The predicted targets of miR-424 are involved in the same pathways as the targets of miR-21; the TGF-β signalling pathway, MAPK signalling pathway, and JAK-STAT signalling pathway, with additional pathways such as acute myeloid leukaemia and antigen processing and presentation, the p53 signalling pathway and SNARE interactions in vesicular transport (see Methods). Several TFs included in the predicted TF→mir-424 associations, are involved in the MAPK signalling pathway (ELK1, ATF2), acute myeloid leukaemia (E2F3, RUNX1) and antigen processing and presentation (CREB1) (see Figure 8A).

The time-lagged expression correlation analysis demonstrates that ELK1, USF2, CEBPB and HOXA4 are positively correlated to the expression of miR-424 (see Figure 6 and Figures 8B and 8C). Apart from the involvement of SPI1 in regulating mir-424 [102], the analysis suggests that ELK1, USF2, CEBPB,

and HOXA4 may be TFs likely responsible for the expression of mir-424 in monocytic differentiation (see Figure 8A).

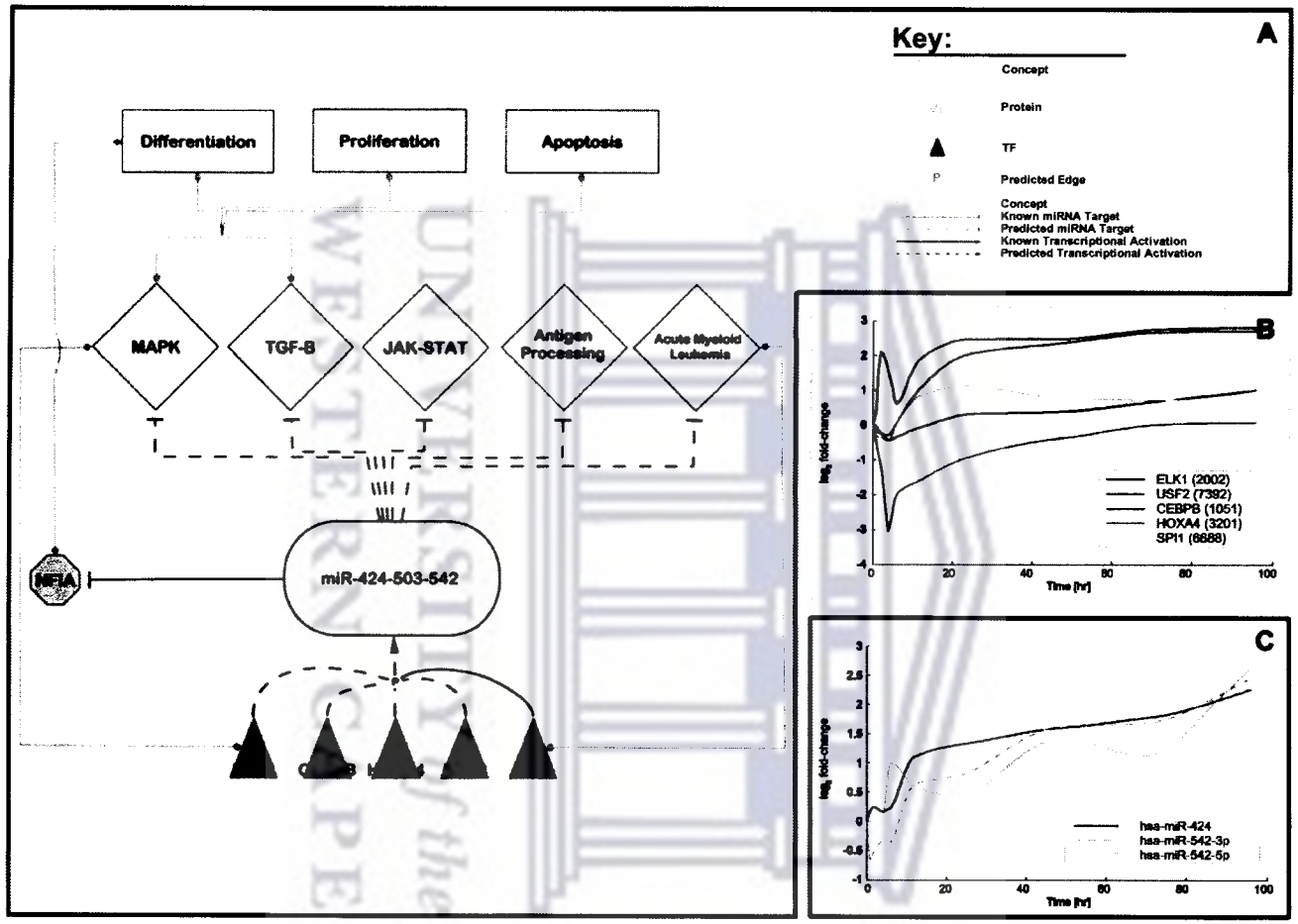


Figure 8. Involvement of miR-424 in monocytic differentiation.
A/ Depicted are the predicted regulations of miR-424/miR-542/miR-503 and their involvement in monocytic differentiation. **B**/ Shown is the $\log_2$ *fc* over time of the interpolated expression data of selected TFs that are predicted to regulate miR-424/miR-542. **C**/ Shown is the $\log_2$ *fc* over time of the interpolated expression data of miR-424, miR-542-3p, and miR-542-5p.

2.4.5.3 miR-155

Chen *et al.* reported that mir-155 is expressed during PMA-induced differentiation in the human promyelocytic leukaemia cell line, HL-60 [104]. The expression data also shows miR-155 to be up-regulated during the differentiation process (see Figure 9C). The TFBS analysis data suggest that several of the 12 TFs (see Table 2), which were identified as being central to the considered differentiation process bind in the promoter region of miR-155 (SP3, NFE2L2, CREB1, NFE2L1 and ELK1; see Figure 6). Zeller *et al.* demonstrated binding of MYC to the promoter region of mir-155 in human burkitt lymphoma cell line, P493-6 [105]. Also, Yin *et al.* demonstrated binding of FOSB and JUNB to the promoter region of mir-155 using ChIP in the human B-cell line [106]. miR-155 has been linked to Epstein-Barr viral (EBV) related diseases that are latent, during which only a subset of viral genes are transcribed with a set of EBV-encoded miRNAs. One such EBV gene is LMP1 which is a known oncogene that induces miR-155 expression in EBV-negative human B cells (DeFew cells) [107]. Gatto *et al.* demonstrated the positive expression of miR-155 in DeFew cells induced with PMA and that the promoter region has two NF-κB (NFKB1) binding sites [107]. Once again, several members of the JUN-FOS family were predicted to bind to the promoter region of mir-155, but neither MYC nor NF-κB. This may be a consequence of the extracted regulatory region for mir-155, being incomplete. The expression data demonstrated the up-regulation of JUN-FOS (see Figure 7B) family members and NF-κB, but a down-regulation of MYC (see

Appendix I X3). These observations indicate that JUN-FOS family members enhance the expression of the miR-155, even though the predicted associations are not within the top quartile of associations with highest *PCCs*.

MiR-155's predicted targets were found to be involved in the same pathways as the targets of miR-21 and miR-424; the TGF-β signalling pathway, MAPK signalling pathway, and JAK-STAT signalling pathway with additional pathways such as acute myeloid leukaemia and Wnt signalling pathway (see Figure 9A and Methods). Several TFs such as ATF2 and ELK1, included in the predicted TF→mir-155 associations, are involved in the MAPK signalling pathway and CREB1 was found to be involved in antigen processing and presentation (see Figure 9A and Methods).

The TFBS analysis and time-lagged expression correlation analysis demonstrated that of the 12 TFs (see Table 2), only NFE2L1 and ELK1 had TFBSs predicted within the promoter of miR-155 and were positively correlated to miR-155 (see Figure 6 and Figure 9B) and thus these findings propose that the NFE2L1→mir-155 and the ELK1→mir-155 associations are likely to be important to the monocytic differentiation process.

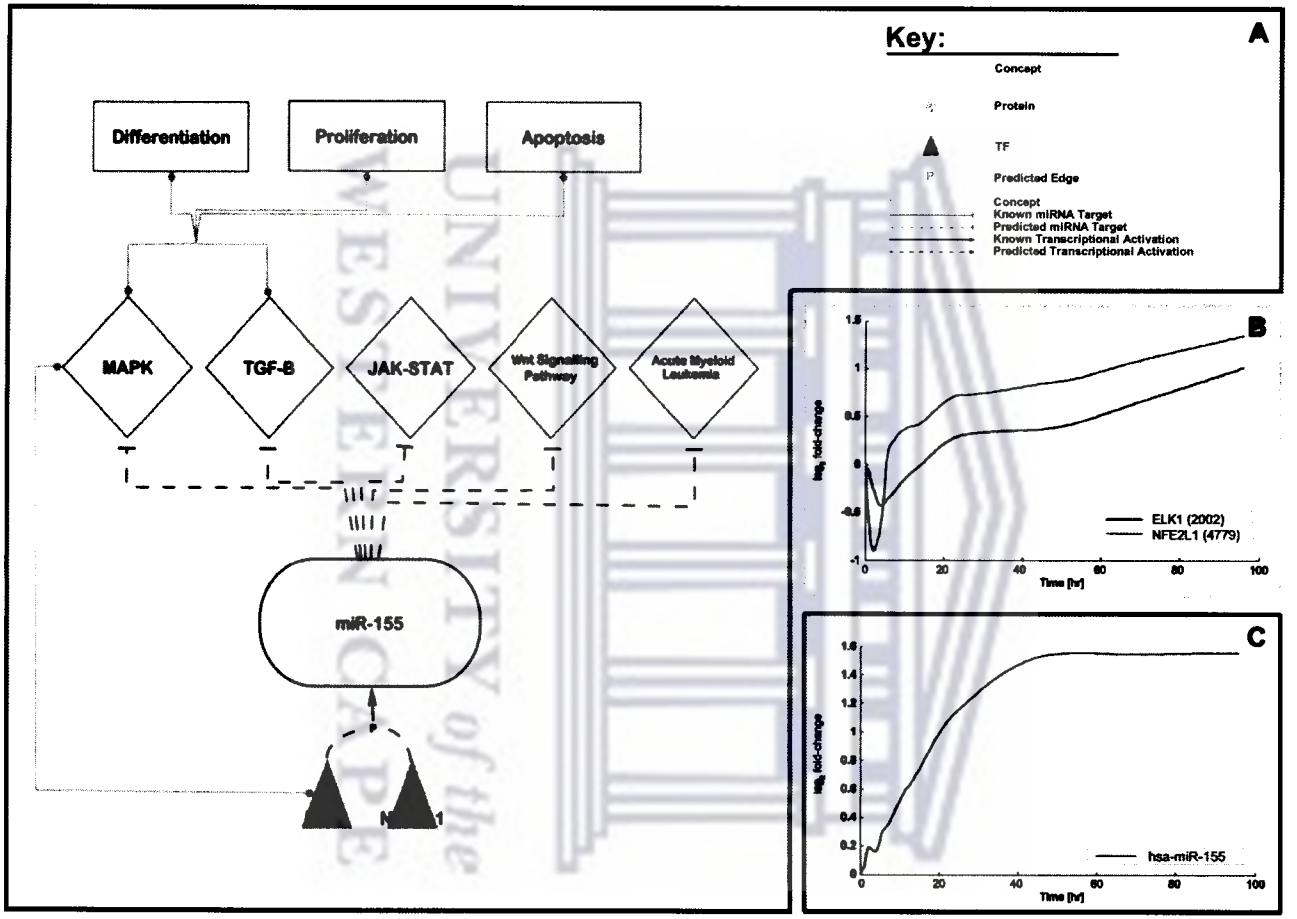


Figure 9. Involvement of miR-155 in monocytic differentiation.
A/ Depicted are the predicted regulations of miR-155 and its involvement in monocytic differentiation. **B/** Shown is the $\log_2$ *fc* over time of the interpolated expression data of selected TFs that are predicted to regulate miR-155. **C/** Shown is the $\log_2$ *fc* over time of the interpolated expression data of miR-155.

2.4.5.4 miR-17-92

Members of the miRNA cluster mir-17-92 are known to be down regulated in the HL-60 cell line after PMA stimulation [104]. The miRNA cluster on chromosome 13 contains several miRNAs (hsa-mir-17, hsa-mir-18a, hsa-mir-19a, hsa-mir-20a, hsa-mir-19b-1, and hsa-mir-92-1 (hsa-mir-92-1 was excluded from analysis, due to ambiguous nomenclature)) that are transcribed as a single transcript. The expression data demonstrates that members of miR-17-92 are indeed down regulated after PMA stimulation and furthermore, that the lowest *PCC* between the expression series of the miRNA cluster members is ~0.86, which supports the cluster membership. Even though the function of miR-17-92 is largely unknown, lymphomas that express these miRNAs at a high level have reduced apoptosis [108,109] and the miRNAs target multiple cell cycle regulators which promote G1→S phase transition [110]. Expression of miR-17-92 is high in proliferating cells and is positively regulated, in part, by MYC (c-Myc) [111]. E2F1, an activator of MYC, is itself a target of miR-17 and miR-20a [108] indicating that both MYC and E2F1 are under the control of a feedback loop. E2F3 has been experimentally shown to activate transcription of the miR-17-92 cluster [87,109]. A model has been proposed that miR-17-92 promotes cell proliferation by targeting pro-apoptotic E2F1 and thereby favouring proliferation through E2F3 mediated pathways [87]. Additionally, E2F3 is shown to be a predominant isoform that regulates miR-17-92 transcription [87]. Time-lagged expression correlation analysis indicates that after ranking *PCC*s of gene expression between miRNAs and putative TFs,

E2F3 is the only TF appearing significantly associated with mir-17-92 within the upper quartile of TF→miRNA associations (see Figure 6).

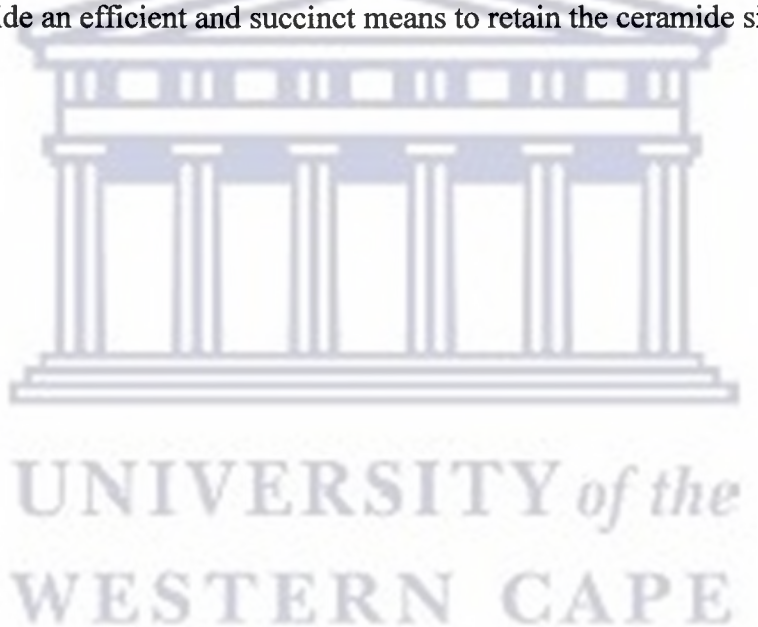
Amongst a small set of eight predicted regulators (E2F1, E2F3, E2F4, TFAP2A, TFAP2B, TFAP2C, TFDP1, SP1), TFDP1 is known to form a heterodimer with another putative TF, E2F1 [112], implicating TFDP1/E2F1 complex as a regulator of miR-17-92 transcription.

The putative regulation of miR-17-92 and its known effects in proliferation, differentiation and apoptotic pathways is depicted in Figure 10A. Specifically, E2F1 and E2F3 are predicted to regulate the miR-17-92 cluster. Figure 10B demonstrates that expression of miR-17-92 members are correlated to E2F3 with a minimum *PCC* of ~0.9. Conversely, miR-17-92 members are correlated with E2F1 by a maximum *PCC* of ~-0.65. A disproportionately high *PCC* of E2F3 gene expression to miR-17-92, as compared to other TFs, seems to support the claims made by Woods *et. al.* that E2F3 is indeed the predominant TF in this regulatory context [87]. In addition, Cloonan *et al.* demonstrated that the pri-miRNA is cell cycle regulated, which supports the claim that the cluster is under the control of E2F family members, which are master regulators of the cell cycle [110]. On inspection of the $\log_2 fc$ of the TF gene expression over time (see Figure 10C), it was observed that E2F3 is sharply up-regulated at 6 hours by ~2 fold, whilst its closely related and pro-apoptotic family member, E2F1, is down-regulated by a factor of ~5.7. After ~70 hours E2F3 and E2F1

gene expression levels return near to baseline corresponding to a progression towards a differentiated state before 96 hours post-PMA stimulation. Yet, regardless of the high *PCC* between E2F3 gene expression and the miR-17-92 cluster, the miRNA cluster is generally down-regulated (see Figure 10D). Acknowledging that the miRNA cluster targets and inhibits a well known RUNX1 (AML1) induced differentiation and proliferation pathway [113], these results strongly suggest that PMA stimulation disfavours both E2F1 induced proliferative and E2F1 induced apoptotic pathways. Whilst, equally, given that both ETS1 and ETS2, components of the RUNX1 differentiation and proliferation pathway, are up-regulated (see Appendix I X3), these results indicate that PMA-treated monocytes up-regulate members of differentiation pathways. In light of the above findings it can be hypothesized that since members of the AP-1 complex are concurrently up-regulated in the early stages after PMA stimulation, monocytic differentiation is mediated by the M-CSF receptor-ligand RAS signalling pathway and indirectly controlled by miR-17-92 through the E2F TF family members E2F1 and E2F3. Generally, this hypothesis seems to be plausible, since RUNX1 is also an inhibitor of miR-17-92 [113] indicating its dual role to both suppress transcription of the pro-proliferative miRNA cluster miR-17-92, and to mediate an M-CSF receptor differentiation pathway. Additionally, patterns of expression observed for miR-17-92 during monocytic differentiation are similar to previous analysis of miR-17-92 expression levels during lung development [114], supporting the general involvement of miR-17-92 amongst differentiation pathways.

TFAP2A (AP-2) and SP1 are two TFs predicted to regulate the miR-17-92 cluster and are notably up-regulated along with the cluster in the first 20 hours post-PMA stimulation. TFAP2A and SP1 are known to activate transcription of an enzyme involved in the sphingolipid metabolism consisting of several metabolites which affect cellular proliferation [115]. TFAP2A and SP1 transcribe sphingomyelin phosphodiesterase 1 (SMPD1) during monocytic differentiation in THP-1 cells after PMA stimuli [115]. SMPD1 is required for the cleavage of sphingomyelin to phosphocholine and ceramide. As ceramide is a known inhibitor of proliferation [116], it seems reasonable that TFs of SMPD1 are up-regulated during differentiation. However, ceramide is also a substrate for several other enzymes whose products have not been implicated in proliferation, apoptosis or differentiation. Interestingly, miR-19a and miR-19b (part of the miR-17-92 cluster), were predicted to target sphingosine kinase 2 (SPHK2) mRNA in four independent databases (see Methods). SPHK2 is an enzyme that metabolizes downstream ceramide products. In the sphingolipid metabolism, SPHK2 has two functions. First, it catalyses the production of sphingosine 1-phosphate from sphingosine, which is produced from ceramides; and second, it catalyses the production of sphinganine 1-phosphate from sphinganine [116]. Sphinganine and sphinganine 1-phosphate have been shown to inhibit and promote cell growth, respectively [116]. Thus, the predicted targeting and down-regulation of SPHK2 by miR-19a and miR-19b in the first 20 hours post-PMA stimulation could prevent the metabolism of two anti-

proliferative metabolites simultaneously, thereby inhibiting proliferation. PMA stimulation is known to block proliferation of THP-1 cells up to 24hrs [46]. Hence, an additional regulatory effect of TFAP2A and SP1 on the sphingolipid metabolism via the miRNA cluster miR-17-92 is proposed. TFAP2A/SP1 mediated transcription of SMPD1 alone might not be enough to maintain an anti-proliferative ceramide signal, as ceramide is metabolized by other factors. On the other hand, TFAP2A/SP1 co-transcription of miRNAs targeting SPHK2 could provide an efficient and succinct means to retain the ceramide signal.



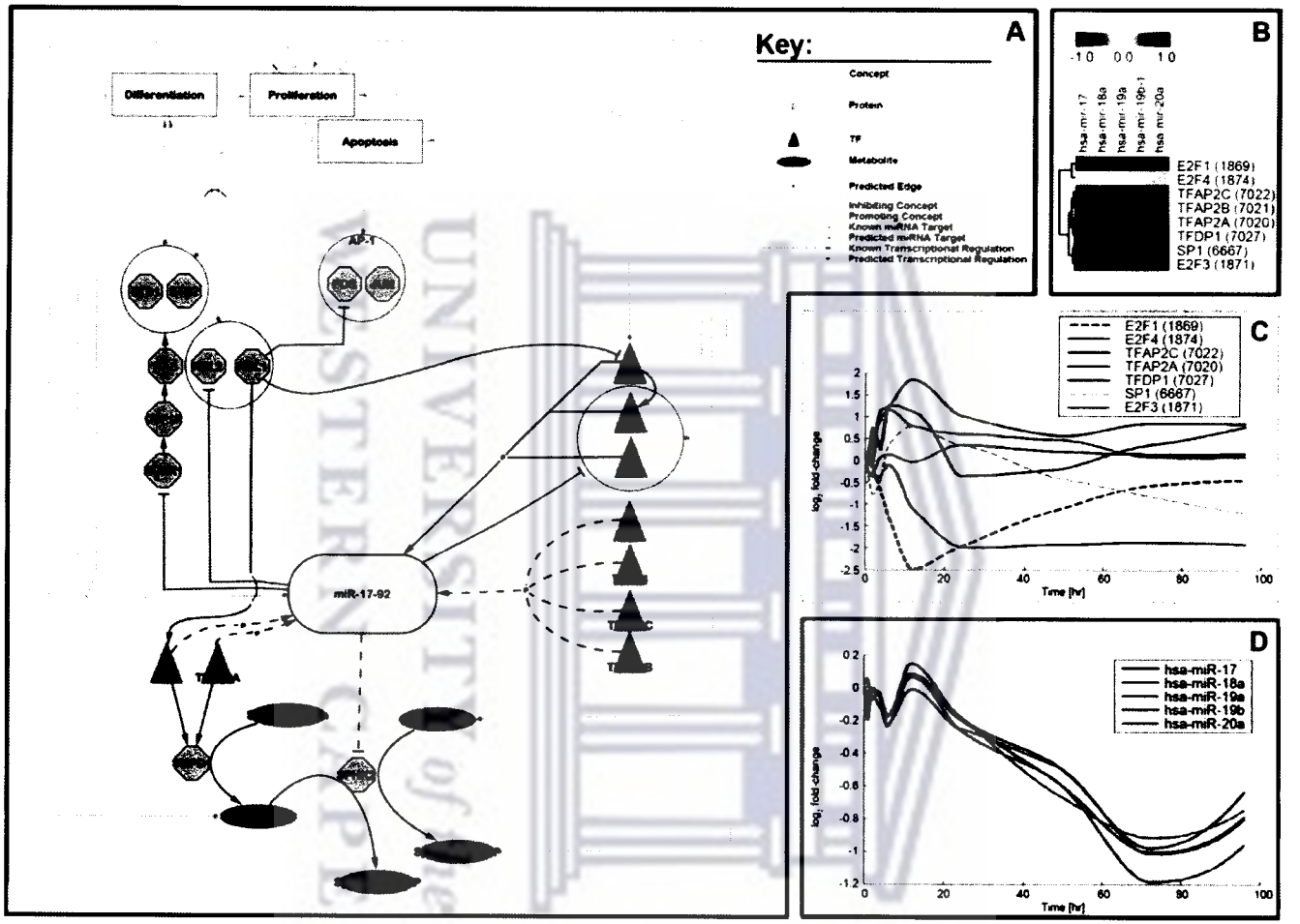


Figure 10. Involvement of miR-17-92 in monocytic differentiation.
A/ Depicted are the predicted regulations of miR-17-92 and their involvement in monocytic differentiation. **B/** Shown is a heat-map representation of the TFs that are predicted to regulate the miR-17-92 cluster. A coloured dot indicates the value of the *PCC* in expression between a TF and a miRNA where the TF has been predicted to regulate the miRNA. **C/** Shown is the $\log_2$ *fc* over time of the interpolated expression data of selected TFs that are predicted to regulate miR-17-92. **D/** Shown is the $\log_2$ *fc* over time of the interpolated expression data of miR-17-20a

2.5 Conclusions

The regulatory machinery that potentially affects transcription of miRNA genes during monocytic differentiation was computationally analysed. The methodology included the extraction of promoter regions for miRNA genes defined by trimethylated histones, computational prediction of TFBSs to establish TF→miRNA associations, and the use of time-course expression data for TFs and miRNAs measured during monocytic differentiation to assess reliability of the predicted TF→miRNA associations via time-lagged expression correlation analysis.

Several TFs (CEBPB, CREB1, ELK1, NFE2L2, RUNX1, and USF2), which are known to play a role in monocytic differentiation, were identified to have an important influence on the regulation of miRNAs as well. In addition, several other TFs (NFE2L1, E2F3, ATF2, HOXA4, SP3, and YY1) were proposed to have a central role in the regulation of miRNA transcription during the differentiation process. For several miRNAs (miR-21, miR-155, miR-424, and miR-17-92) it was shown how their predicted transcriptional regulation could impact the monocytic differentiation process.

The process of identifying a complete list of TF→miRNA associations is hampered by the correct definition of promoter/regulatory regions being an unresolved issue that has a great impact on all studies that deal with gene

regulation. A recent set of promoters defined, based on the observation that histones are generally trimethylated at lysine 4 residues at TSSs of genes, was used. Due to the promoter definition employed by Marson *et al.* it was not possible to extract regulatory regions for several miRNAs. Furthermore, the subset of promoter regions defined by Marson *et al.* that were used in this analysis range in length between 200 and ~4,700 bp with 60 percent of the utilised promoter regions being of length below 202 bp. Consequently, the promoter set defined by Marson *et al.* allows for analysis of regulatory elements primarily proximal to the TSS. Nevertheless, it has been well documented [117,118] that proximal regulatory elements such as the TATA box play an important role in type II polymerase gene transcription. However, the promoter set utilised in this study, though possibly incomplete, represents one of the first described sets of regulatory regions for miRNA genes.

It is important to note that the transcriptional circuitry described in the analysis is specific towards monocytic differentiation expression data, as several of TF→miRNA associations were discarded due to missing/incomplete expression data for either TF or miRNA. Furthermore, the expression based approach is limited as mature miRNAs are not the direct product of the TF-mediated regulation, but can undergo post-transcriptional regulation on the pri- and pre-miRNA level [119]. Hence, it is possible that miRNAs that are transcribed together as one primary transcript, show different expression profiles on the mature miRNA level. The three main reasons that constrained

the set of TF→miRNA associations determined in this study are as follows: 1/ An incomplete promoter set for miRNA genes. 2/ An incomplete/inaccurate motif set for the prediction of TFBSs. 3/ An incomplete expression set for TFs and miRNAs. Each of the reasons impacts on the accuracy of the predicted TF→miRNA associations.

Nevertheless, this analysis provides the first large-scale insights into the transcriptional circuitry of miRNA genes in monocytic differentiation. Taken together, the results suggest important regulatory functions of several TFs on the transcriptional regulation of miRNAs. The regulatory networks discussed here form only the starting point for an in-depth analysis of the regulatory mechanisms involved. The predicted TF→miRNA associations and their corresponding *PCC*s can provide the basis for a more detailed experimental analysis of miRNA regulation during monocytic differentiation.

2.6 Acknowledgments

In this analysis the experimental data was provided by Harukazu Suzuki, Yoshihide Hayashizaki, Alistair Forrest, and Mutsumi Kanamori from the RIKEN Omics Science Center. Special thanks to Cameron R. MacPherson, Magbubah Essack, Ulf Schaefer, and Mandeep Kaur for helpful discussions about the biological interpretation of the results.

2.7 Appendix I

X1 – Interpolated expression data for 34 mature miRNA

The file consists of 195 columns. The first column contains the mature miRNA identifier. The second column contains the associated pre-miRNA identifier(s). Column 3-194 contain the interpolated expression values ranging in half an hour steps from 0 to 96 hours.

X2 – Predicted TF→miRNA associations and their inferred *PCC* values

The file consists of three columns. The first column contains the TF. An identifier consists of the Gene Symbol separated by an underscore with the Entrez Gene id. The second column contains the miRNA identifier that forms an association with the TF of the first column. The third column contains the inferred *PCC* for the association, which is based on the expression data of the TF and the mature miRNA associated to the miRNA(s). In total the file contains 1,989 TF→miRNA associations.

X3 – Interpolated expression data for 258 TFs.

The file consists of interpolated expression data for 258 TFs that are present in the predicted TF→miRNA associations. Furthermore, the file consists of 194 columns. The first column is the TF identifier (Entrez Gene Id). Column 2-194 contain the interpolated expression values ranging in half an hour steps from 0 to 96 hours.

Chapter 3

Predicting Human Transcription Factor Interactions from Primary Structure

3.1 Abstract

Combinatorial control is an important mechanism within the scope of transcriptional gene regulation. Physical interactions between TFs play a crucial role in this process. Knowledge about these interactions is scarce and thus predicting these interactions will help in better understanding the complex machinery involved in gene regulation. The present study attempts to develop a system that is able to predict if two TF interact. It is based on primary sequence information of the participating TFs alone to minimise the data acquisition overhead. Amino acid properties were utilised to construct simple representations of TF pairs. A support vector machine was trained on known examples of TF interactions to create a model that is able to classify these TF pairs. Cross-validation experiments demonstrated a prediction accuracy of 80.1 %. Feature selection techniques led to a high reduction in the computational resources necessary for model selection. Even though the system for TF interaction prediction is of a simplistic nature, its performance is comparable to much more complicated approaches for predicting protein-protein and TF interactions.

3.2 Introduction

The transcriptional regulatory machinery that acts on the transcription of genes is complex and not yet completely understood. TFs are proteins that regulate a gene's transcription by binding in regulatory regions on genomic DNA [4]. They are found in the nucleus of cells where they often work cooperatively to enhance or repress the transcription of various genes [6,7]. This cooperative functioning can be achieved through the physical interaction of TFs [6,7]. To better understand the elaborate transcriptional machinery that acts within the nucleus of cells, it is essential to know about these interactions.

The combinatorial regulation of genes has been studied extensively [78,120-124]. Here, groups of TFs were identified that work cooperatively to facilitate their role on the transcription of genes or gene groups. The combinatorial regulations described do not necessarily entail the physical interaction of the participating TFs. However, to better understand the underlying mechanisms that play a role in gene regulation, knowledge of TF interactions is of great importance.

Protein-protein interaction (PPI) prediction gained a lot of attention over the last decade. Various methods and tools exist that predict such interacting proteins [125-129]. These methods make use of manifold properties of proteins and combinations thereof, such as functional categorisation and gene ontology annotations [130], primary structure [131-135], secondary, tertiary structure,

and protein domain information [126,127,129,136-139], ortholog-based and phylogenetic-based profiles [140,141], gene expression and other experimental data [142], text mining [128,143], etc.

Predicting TF interactions can be seen as a subclass of the PPI prediction problem that, unfortunately, is more complex. Members of TF families are often, due to duplication, sequence-wise very similar to each other [144], which makes a sequence based prediction of interactions difficult. Furthermore, TFs are located in the same cell compartment, the nucleus, making it impossible to utilise such discriminative factors as cellular localisation as distinguishing attributes. Finally, information about known TF interactions is scarce as compared to PPIs.

Former approaches for deciphering the combinatorial control of TFs have included co-expression analysis [145], thermodynamic models based on time-course microarray data [146], relationships of TFBSs [147,148], or phylogenetic footprinting and combinations of the afore mentioned methods [149]. To aid future studies of combinatorial gene regulation, the present study aims at predicting TF interactions computationally. As with computational PPI predictions, a representation for an interacting TF pair has to be found with a multitude of possibilities being available as mentioned before. Common to all of these is that often it is difficult to acquire such information or the information is not readily available at all. To circumvent these obstacles, the

approach utilised here is based on protein sequence information alone. The present analysis shows that even with minimal prior knowledge about TFs, it is possible to achieve comparable prediction performance to more complicated approaches. This analysis utilises amino acid (AA) properties for TF primary sequences and combines these into a representation for TF pairs. The artificial intelligence system employed to classify these interactions is a support vector machine (SVM) [150,151]. The SVM learns, based on examples, a model that classifies positive and negative TF interactions. A brute-force grid search was employed to find the SVM parameter combination with which a trained model achieves best prediction performance. Different feature selection techniques were tested to minimise the computational resources necessary for model selection. A 10-fold cross-validation (CV) showed that the system presented here performs with an accuracy of 80.10%, while having a precision of 89.30 %, a recall of 68.89%, and a specificity of 91.39% respectively. The advantage of the here presented methodology over other more complicated methods lies in its simplicity, which could be easily extended to include more complex data, but yielded a comparable prediction performance nevertheless.

3.3 Methods

3.3.1 Interacting Transcription Factors

TRANSFAC Professional database (version 11.4) [16,17] contains information about TFs, TF families, DNA binding site, motifs for binding site prediction,

etc. In addition, it contains information about interacting TFs. In total the database contains 2,291 human TF interactions (based on the interaction of TRANSFAC entities) mined from scientific literature. From these, all interactions were sub-selected where each participating TF has AA sequence information available in the database. In this manner, a total of 338 positive TF interaction pairs (TRANSFAC entities) were extracted.

Negative examples of TF interactions were randomly chosen by associating two TRANSFAC TF entities that have AA sequence information available. In total, 1,184 TRANSFAC TFs have an associated AA sequence. Three different classes of negative TF pairs were identified:

i/ Absolute negatives: None of the two TFs that form the pair is part of the TFs forming the positive interactions.

ii/ Partial negatives: One TF that forms the pair is part of the TFs from the positive interactions.

iii/ PPI negatives: Both TFs that form the pair are part of the TFs from the positive interactions but the pair itself is not within the positive group of interactions.

3.3.2 Feature Representation and Feature Vectors

The AAIndex database [152] (www.genome.jp/aaindex/) contains biochemical and physicochemical properties for AAs collected from scientific literature. In

total the database contains 544 AA properties. All properties that were available for all 20 AAs were selected. This reduced the number of properties that conformed with this condition to 531 (see Appendix II Y1). The feature vector for a TF pair consists of two concatenated feature representations for each participating TF. The feature representation F_t for a single TF t consists of 531 features f_p, each representing the average of one of the 531 AA properties p over its AA sequence, $F_t = (f_{p1},\ldots f_{p531})$. An individual feature f_p for AA property p is calculated as:

$$f_p = \frac{\sum_{i=1}^{n} p_j}{n},$$

where n equals the protein sequence length, i the *ith* AA, and p_j the value of AA property p for AA j. Thus, each TF, disregarding the length of its AA sequence is represented with the same length of feature representation. If an AA in the sequence is either "X" or "U", then the AA was disregarded from the averaging process.

To represent a pair of interacting TFs, the representations for individual TFs got concatenated into one vector, consisting of 1,062 features (531 features from each TF comprising the interacting pair). In order to avoid multiple different representations of the same TF interaction pair caused by symmetry of the interaction, the following condition for concatenation had to be met. Consider a TF interaction *A-B*, where *A* and *B* are two TFs, then the first

representation is always the one for the TF with the smaller molecular weight. Lets assume that in the interaction above *B* is the TF with the smaller molecular weight, then the interactions between *A* and *B* is always expressed as *B-A*. In this manner, the resulting feature vector for a TF interaction is always unique.

3.3.3 Support Vector Machines

The artificial intelligence system utilised here for the classification task is a SVM [150,151]. The inputs into a SVM are vectors of features of arbitrary length (for one problem each vector must have the same length). A feature can either be nominal (yes vs. no; present vs. not present) or continuous (real numbers). The most common classification task is binary, which entails that the SVM classifies a feature vector into one of two classes (e.g. positive vs. negative). To be able to train a SVM model, it is necessary that the SVM "learns" on known examples of preferable both classes. Hence, the SVM belongs to the class of statistical supervised learning methods (e.g. decision trees [153], random forest [154] , etc.), as opposed to the class of unsupervised learning methods (e.g. clustering, independent component analysis [155], etc.).

Each of the vectors represents a point in a high-dimensional space (the number of dimensions equals the number of features). The two classes of vectors form clouds in the high-dimensional space, e.g. positive and negative TF interactions (see Figure 11). The SVM tries to calculate a representation for a hyperplane

that separates the two classes. In an ideal case the two clouds would be clearly separated from each other so that such hyperplane is easily found. In real case scenarios the clouds almost always overlap, thus making the construction of such a hyperplane difficult. A new vector that is projected into the space is mapped on one of the two sides of the hyperplane and thus classified into one of the two classes. The representation of the hyperplane which the SVM "learns" consists of the so called support vectors (see Figure 11) which are vectors from the set of vectors used to train the SVM model. Support vectors are the vectors with minimal distance to the theoretical hyperplane in the region where the vectors of the two classes are closest to each other. The aim is to maximise the margin of the support vectors to the hyperplane while minimising the training error. The parameter c controls the trade-off between margin and the error.

The hyperplane presented in Figure 11 that is used to separate the two classes is linear. For most problems this is generally not applicable, e.g. overlapping clouds. The kernel technique aims to overcome this problem by mapping the input data into an even higher dimensional space, where a linear separation is possible [156]. The Gaussian radial basis function kernel (rbf-kernel) is used in this analysis. It is defined as:

$$K(x_i, x_j) = \exp(\frac{-|x_i - x_j|^2}{\sigma^2}),$$

where x_i and x_j are two vectors where one of them is a support vector and σ is an adjustable parameter that determines the area of influence of the support

vector over the data space. Larger values of σ reduce the number of support vectors, since each support vector covers more data space.

The SVM implementation used in the present analysis is SVM^{light} [157] (http://svmlight.joachims.org/) which is free for academic use. Once trained, the SVM model can be utilised to classify new vectors. The parameter *j* can be used to apply different weights for penalties for wrongly classified positive and negative examples. Once a new vector is submitted to a SVM model, the decision function of the SVM outputs a value that indicates if the new classified vector belongs to the positive or negative class. The threshold for the decision function value *th* determines how to classify a decision function value, meaning that values greater than *th* are classified as positives and values less than *th* classified as negatives. By default this threshold, *th* equals zero.

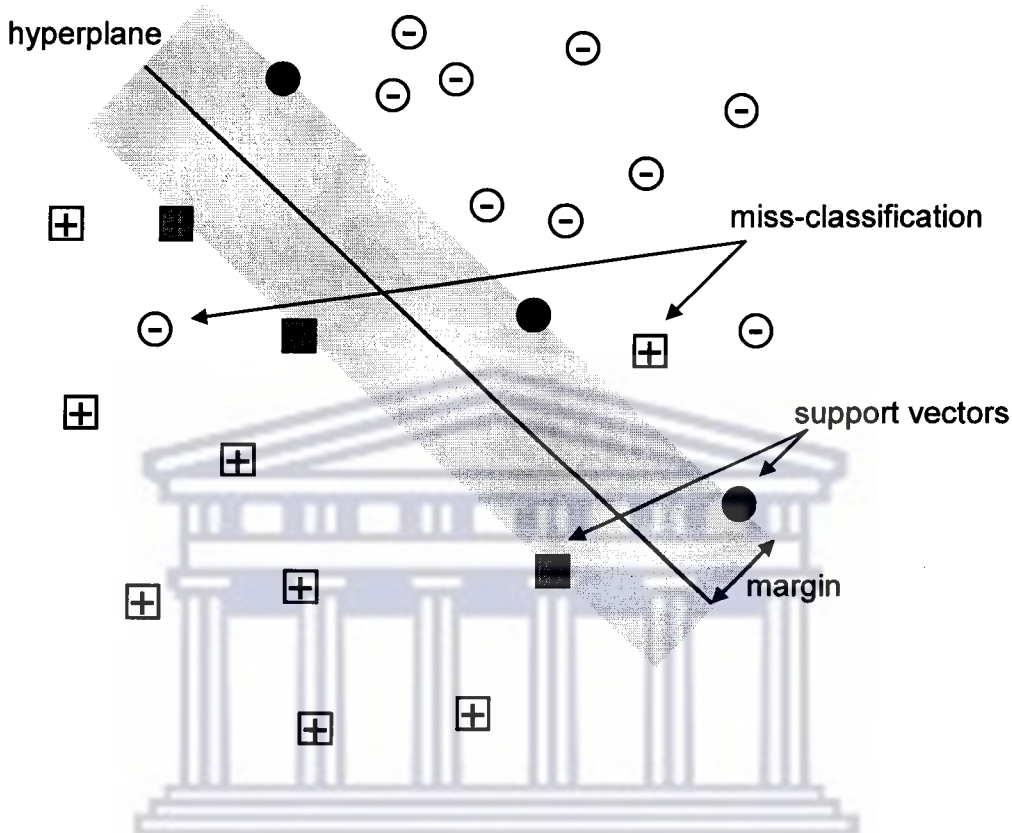


Figure 11. Schematic of a SVM Classification
The figure illustrates positive and negative examples (represented through plus signs and minus signs) in two dimensional space. The SVM learned the representation of a hyperplane, here illustrated through a grey rectangle that best separates the two classes of examples from each other. The examples that lie on the edge of the hyperplane are the so called support vectors (the actual representation learned by the SVM).

3.3.4 Min-Max Scaling

In general it is good practice to scale the values for each separate feature before presenting the feature vectors to a SVM. The scaling technique employed here is commonly known as min-max scaling:

i/ For all values v_f of feature f over all examples find minimum value $v_{f\text{-}min}$ and maximum value $v_{f\text{-}max}$

ii/ For an individual value w_f and feature f, calculate the new scaled value $w_{f\text{-new}}$ as: $w_{f-new} = \frac{w_f - v_{f-\min}}{v_{f-\max} - v_{f-\min}}$

This results in a scaling for each value for one feature between zero and one. When a model is utilised that was trained on the scaled training set, T_{train} to classify examples of a test set, T_{test}, then the values of T_{test} have to be scaled before classification according to the minimum and maximum values for each feature found while scaling T_{train}. Thus, the scaled values of T_{test} do not necessarily lie in the range of zero and one but are scaled according to T_{train}, which is important for the classification task.

3.3.5 Model Optimisation and Performance Evaluation

Several performance measures are used to judge the performance of a classification system that is based on machine learning. Considering the

confusion matrix presented in Table 3, the precision (positive predictive value) is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

The precision is the percentage of real positives on all predicted positives. The recall (sensitivity) is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The recall is the percentage of predicted real positives on all positives in the set. The specificity is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The specificity denotes the percentage of predicted real negatives on all negatives in the set. The accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

The accuracy is the percentage of true prediction on all predictions.

The F-measure is the harmonic mean between precision and recall and is defined as:

F-measure = (2* precision * recall) / (precision + recall)

Several parameters of the SVM can be tweaked to enhance the performance of the classification. The parameters changed in this study in order to enhance the performance are the trade-off factor c, the cost factor j, the σ parameter of the rbf-kernel function, and the threshold for the decision function *th* (see above).

The method utilised for model selection was a brute-force grid search over the parameter space. The parameters for c, j, and σ were tested on a exponentially growing sequence ($c = 2^{-5}, 2^{-4},\ldots, 2^{3}$; $j = 2^{-5}, 2^{-4},\ldots, 2^{12}$; $\sigma = 2^{-15}, 2^{-14}, \ldots, 2^{3}$). The threshold for the decision function, th was varied from -0.99 to 0.99 in steps of 0.01. In total, 612,522 different parameter combinations were tested. Each parameter combination was tested with a 10-fold CV and the performance measures calculated as the average over the ten CV runs.

Table 3: Confusion Matrix

		Actual class	
		Positive	**Negative**
Predicted class	**Positive**	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

The table indicates the nomenclature for an outcome of a prediction in perspective to the actual value.

3.3.6 Feature Selection Based on the Mahalanobis Distance

During feature selection, the task was to find the best combination of features with which help it is possible to achieve similar performance for the classification task and to speed up the model selection.

Given two vectors with random variables x and y sampled from the same distribution, and the covariance matrix S, the Mahalanobis distance [158] is defined as:

$$d(x,y)=\sqrt{(x-y)^T S^{-1}(x-y)}$$

The same Mahalanobis distance can be used to calculate the distance between two matrices. Here each matrix contains vectors of one of the two classes. The aim is to find the subset of features that maximises the distance between the two classes. An iterative process was conducted:

i/ Find feature f_1 that maximises the Mahalanobis distance between the two classes.

ii/ Delete feature f_1 from further investigation and put f_1 into feature group g.

iii/ Find next feature f_n that maximises together with features in g the Mahalanobis distance between the two classes.

iv/ Put f_n into feature group g and delete f_n from matrices.

v/ Redo steps iii/ and iv/ until no improvement in the Mahalanobis distance can be achieved.

The result is a group of features g that maximises the distance between the two matrices and thus between the two classes.

3.3.7 Feature Selection Based on t-Statistic

The second approach for feature selection ranks the features according to the t-statistic, which is based on Student's t-test [159,160]. The underlying assumption is that the values of each feature for each separate class are sampled from a normal distribution with equal variances. The t-statistic assesses the quality of each feature to separate members of the positive and negative class, while comparing the mean values of a specific feature in both classes. Each feature is evaluated independently and is assigned a t-value. The higher the t-value for a feature, the better the feature is suited to separate the two classes. The t-statistic is calculated for each feature. Sorting the list of features based on their calculated t-values, results in a ranked list of features (ranked according to their suitability to separate the two classes). The formula utilised to calculate the t-value t is:

$$t = \frac{|\mu_x - \mu_y|}{\sqrt{\frac{\frac{1}{n_x} + \frac{1}{n_y}}{n_x + n_y - 2} * \left(\sum_{i=1}^{n_x} (x_i - \mu_x)^2 + \sum_{i=1}^{n_y} (y_i - \mu_y)^2 \right)}},$$

where x and y are two vectors of values of the considered feature of the two classes X and Y (positive and negative), n_x is the number of elements in x and n_y is the number of elements in y. μ_x and μ_y are the mean values of x and y. With the list of features ranked according to the t-value in decreasing order, it is possible to create feature vectors of any length n (n <= 1062) based on the n features with highest t-value.

3.4 Results

In the following, four experiments were conducted. First, the performance evaluation of the methodology utilising the complete feature set is presented. Subsequently, the best models were tested on independent data sets of TF interactions and on data with randomised class labels. Finally, feature selection techniques were applied to reduce the feature space and speed up the model selection and performance evaluation procedure.

3.4.1 Performance Evaluation Using the Complete Feature Set

Examples of positive TF interactions were selected from the TRANSFAC database (see Methods). In total, it was possible to extract 338 TF interactions. The number of TFs (TRANSFAC entities) comprising these interactions is 212. The distribution of sequence lengths of this set of TFs is presented in Figure 12B. The mean and standard deviation of the sequence lengths is 546.54 AAs and 400.08 AAs, respectively.

The model utilised for the classification task should be able to distinguish between positive and negative TF interactions. Thus, it is necessary to present negative examples to the artificial intelligence system as well. No database exists where information about non-interacting TFs is stored. Thus, a set of TF interactions that do not appear in the positive set of TF interactions had to be randomly sampled. Three groups of negative interactions were identified (see

Methods). 112 TF interactions from each group were sampled at random, resulting in a set of 336 negative TF interactions. These are comprised out of 493 different TFs (TRANSFAC entities). Figure 12C shows the length distribution of the AA sequences of all TFs comprising the randomly sampled negative set of TF interactions. The mean and standard deviation of the sequence length in number of AAs is 520.17 and 438.67, respectively. The distribution of sequence lengths of the combined set of 566 unique TFs from the positive and negative set of TF interactions is shown in Figure 12A. The mean and standard deviation of the combined unique set of TFs is 523.41 AAs and 438.25 AAs, respectively.

For the complete set of 674 positive and negative TF interactions, all 1,062 features were extracted and the feature vectors created (see Methods). The set of 674 positive and negative TF interaction feature vectors were randomly split into ten groups, preserving the same ratio of positives and negatives within each group. The model selection and performance evaluation was done utilising a 10-fold CV, where in each step the respective training set was scaled (min-max scaling, see Methods) and the respective testing set scaled accordingly to the training set (using the maximum and minimum values from the respective training set).

The precision versus the recall of all tested parameter combinations is presented in Figure 13A. The receiver operating characteristic (ROC) curve for

the same parameter combinations is presented in Figure 13B. The best achievable accuracy was 80.10%. The best F-measure was found to be 79.25%.

The model selection and performance evaluation process ran on a Linux Pentium 4 core duo machine with a 1.8 GHz CPU and 2GB of memory in ~92 hours. 612,522 different parameter combinations were tested and evaluated.

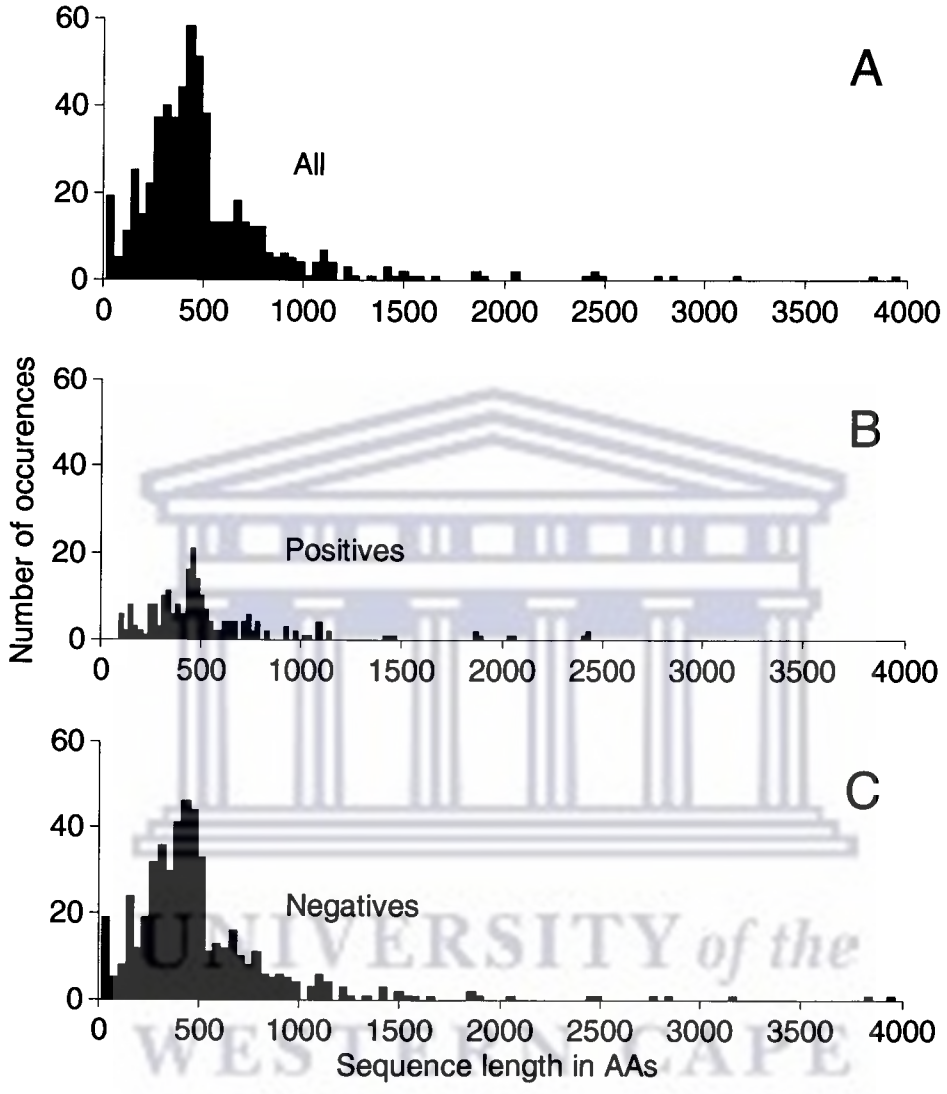


Figure 12. Histogram of Sequence Length Distribution
A/ Presented is a histogram with 100 bins of the length of all sequences in the complete set of 674 TF interactions. 566 different protein sequences are shown in the figure. The average sequence length is 523.41 AAs and the standard deviation is 438.25 AAs. **B/** Presented is a histogram with 100 bins of the length of all sequences in the positive set of 338 TF interactions. 212 different protein sequences are shown in the figure. The average sequence length is 546.54 AAs and the standard deviation is 400.08 AAs. **C/** Presented is a histogram with 100 bins of the length of all sequences in the negative set of 336 TF interactions. 493 different protein sequences are shown in the figure. The average sequence length is 520.17 AAs and the standard deviation is 438.67 AAs.

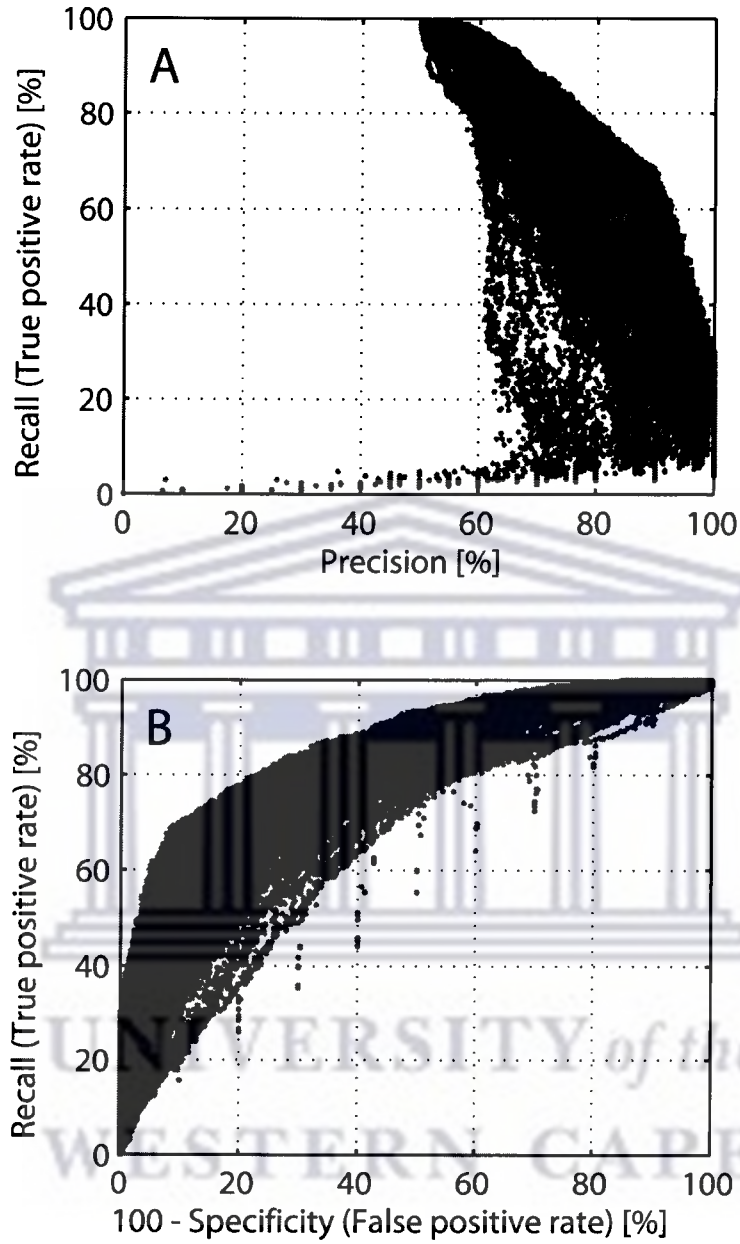


Figure 13. Performance Results of the CV with the Complete Feature Set
A/ Depicted is the precision versus the recall of all 612,522 different parameter combinations tested during the 10-fold CV on the TRANSFAC derived data with all available features. The typical trade-off between precision and recall is evident. **B/** Presented is the ROC-curve. Each dot represents again a performance result of one parameter combination tested during the 10-fold CV on the TRANSFAC derived data with all available features. Here the false positive rate is plotted versus the true positive rate.

3.4.2 Performance Evaluation on Independent Data Sets

After evaluating the performance of the methodology using CV, it is of interest to see how the method performs on independent sets of TF interactions. Several databases provide data about PPIs, such as the Human Protein Reference Database (HPRD; [161]; http://www.hprd.org/), the Biomolecular Interaction Network Database (BIND; [162,163]; http://www.bind.ca/), the Molecular INTeraction database (MINT; [164]; http://mint.bio.uniroma2.it/mint/Welcome.do), the IntAct database [165,166] (http;//www.ebi.ac.uk/intact/), and the Database of Interacting Proteins (DIP; [167,168]; http://dip.doe-mbi.ucla.edu/dip/Main.cgi). A recent collection of TF interactions from these sources was downloaded [78]. From these, all TF interactions from the public sources described above were extracted. Each TF within these interactions is represented by the Entrez gene identifier of the corresponding gene. The protein sequences for all Entrez gene identifiers were assigned using Ensemble's BioMart system [83]. A TF represented by a gene identifier might have multiple protein sequences assigned to itself, due to the nature of gene-protein relations. This led to various feature representations for single TFs (see Methods). Hence, the feature vector for a TF interaction, which consists of the concatenated feature representations of two TFs, is not always unique.

In the following, a TF interaction is represented by concatenating each possible combination (similar to a cross-product) of the feature representation of each protein sequence associated to the first and second TF. This led to several different feature vectors for the same TF interaction, due to the multiple assigned protein sequences per TF. The positive interactions were created for each source separately and for one common set of all unique interactions. The number of interactions and number of feature vectors for each group is presented in Table 4. The number of unique TFs that are part of all interactions is 1,344. The distribution of the sequence lengths of all 1,344 TFs is presented in Figure 14. 1,907 sequences are associated to the 1,344 TFs. The average sequence length in number of AAs is 607.14 and the standard deviation is 491.39 respectively.

Negative examples were sampled at random. Here, two TFs from the set of 1,344 unique TFs were selected at random. If they were not known to interact, feature representations of their protein sequences were created, associated to each other (again similar to a cross-product), and the feature vectors created as described above. Negative feature vector examples were sampled for each independent set in a similar amount as positive feature vector examples are available for the respective independent set (see Table 4). Thus, each independent set consists of similar amounts of positive and negative TF interaction feature vector examples.

Table 4. Number of Positive and Negative TF Interactions and Unique Feature Vectors for Each Independent Set of Interactions

Set	Number of positive interactions	Number of positive feature vectors	Number of negative interactions	Number of negative feature vectors
BIND	668	1528	838	1528
DIP	204	574	304	574
INTACT	631	1417	689	1417
MINT	839	1935	1024	1934
HPRD	4907	11190	5609	11189
ALL	5213	11944	6986	11944

The table presents the number of unique TF interaction and unique feature vector representations for each of the independent set of interactions. Presented are the numbers for the positive interactions and the randomly sampled negative interactions. The sets "ALL" consist of all unique TF interactions over all utilised database sources.

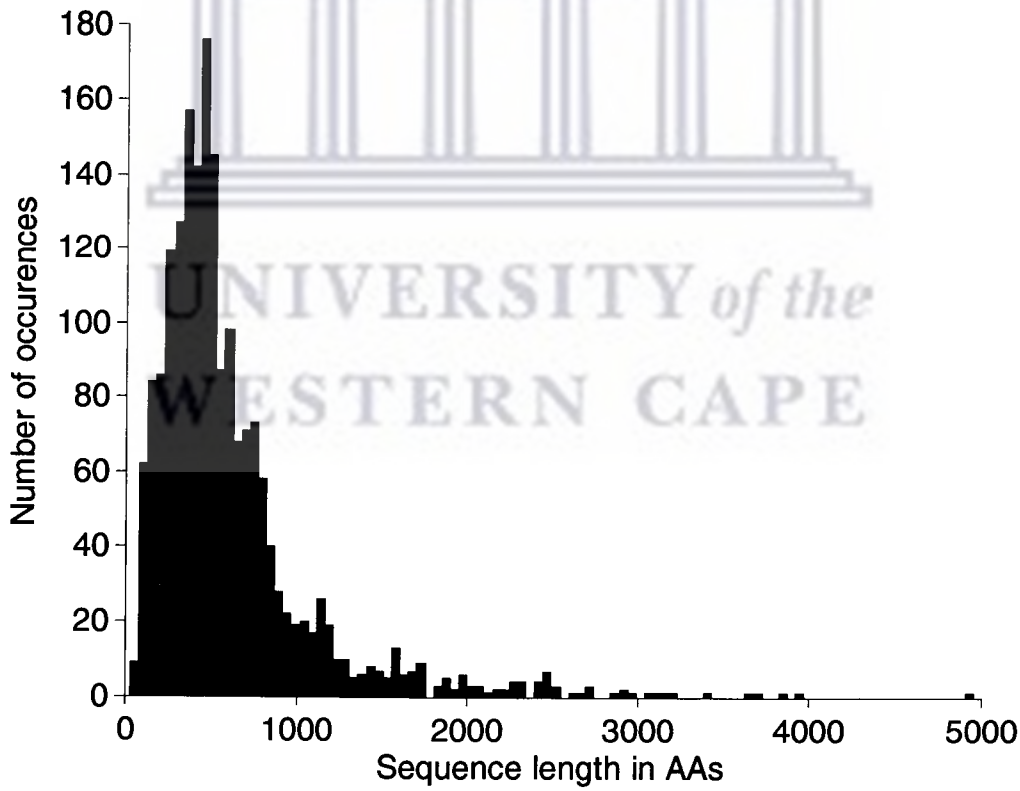


Figure 14. Histogram of the Sequence Lengths of TFs in the Independent Sets of TF Interactions

Presented is a histogram with 100 bins of the length of all 1,907 sequences of the 1,344 TFs from the independent sets of TF interactions. The average sequence length is 607.14 AAs and the standard deviation is 491.39 AAs respectively.

In order to classify the new data sets, a model is necessary that performs the classification task. Two strategies to derive such a model were applied. The first utilises the parameter combinations found during the CV that yielded best performance results in terms of accuracy and F-measure. In this manner, two different parameter combinations were chosen (see Table 5). The complete set of vectors (674) derived from TRANSFAC was scaled. Each individual independent test set was scaled accordingly to the former scaling of the TRANSFAC vectors. Two models were created utilising the scaled TRANSFAC vectors and employed to classify the scaled independent sets.

The second strategy involved a different model selection and training methodology. The complete TRANSFAC feature vector set of positive and negative interactions was randomly split into two equal sized sets of 337 TF interactions, while preserving the ratios of positive to negative interactions in either subset. One of the subsets was denoted as the "Training set" while the other was denoted as the "Testing set". The "Training set" of feature vectors was once again scaled. Afterwards, scaling of the "Testing set" was performed accordingly to the "Training set". A grid search with the same value ranges as denoted above for the SVM parameters was performed. The scaled "Training set" was utilised for creating the model with the chosen parameter combination and the scaled "Testing set" was employed to evaluate the model. In this manner all parameter combinations were tested. Once again two parameter combinations were chosen for model development. The first parameter set

according to the best accuracy and the second according to the best F-measure on the "Testing set" (see Table 5). The results in form of precision versus recall of all tested parameter combinations are presented in Figure 15A. The respective ROC-curve is presented in Figure 15B.

To summarise, four different models derived from the TRANSFAC TF interactions were selected for further investigation of their performance on the independent sets of TF interactions:

1. The model that achieved the best accuracy during 10-fold CV
2. The model that achieved the best F-measure during 10-fold CV
3. The model that achieved the best accuracy on the "Testing set" with the Training-Testing setting
4. The model that achieved the best F-measure on the "Testing set" with the Training-Testing setting

To create the model 1 and 2 all scaled TRANSFAC vectors were utilised, whereas for models 3 and 4 only the scaled "Training set" of the second strategy explained above was used. The performance of each model on the independent sets of TF interactions are presented in Table 6.

Table 5. Selected Models, their Parameter Combinations, and Performance on their Respective Test Data

Model	c	j	σ	th	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-measure [%]	TP	FP	TN	FN
1	8	8	0.25	0.18	89.30	68.89	91.39	**80.10**	77.35	23.3	2.9	30.7	10.5
2	4	1	0.25	-0.16	75.55	83.71	72.31	78.02	**79.25**	28.3	9.3	24.3	5.5
3	2	8	0.125	-0.03	74.73	82.25	72.02	**77.15**	78.31	139	47	121	30
4	1	8	0.125	-0.03	71.36	86.98	64.88	75.96	**78.40**	147	59	109	22

Presented are the four chosen models for application on the independent sets of TF interactions. Highlighted numbers, indicate the method for selecting the respective model (highest accuracy versus highest F-measure). The first two models were chosen from the 10-fold CV runs, whereas the model 3 and 4 have been selected utilising a simple training and testing set methodology (see main text). All performance measures for model 1 and 2 represent averages over the CV runs. This is the reason why the numbers for TP, FP, TN, and FN are floating point numbers.

Table 6. Prediction Performance of the Four Models on the Independent Sets of TF Interactions

Model	Independent set	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-measure [%]
1	BIND	80.14	29.58	92.67	61.13	43.21
	DIP	82.97	26.31	94.60	60.45	39.95
	INTACT	70.68	13.27	94.50	53.88	22.34
	MINT	74.50	20.98	92.81	56.89	32.74
	HPRD	79.69	24.48	93.76	59.12	37.45
	ALL	78.77	23.89	93.56	58.72	36.66
2	BIND	67.31	48.10	76.64	62.37	56.11
	DIP	70.59	48.08	79.97	64.02	57.20
	INTACT	59.68	31.55	78.69	55.12	41.27
	MINT	64.21	40.05	77.66	58.85	49.33
	HPRD	67.46	45.33	78.13	61.73	54.22
	ALL	67.00	44.63	78.01	61.32	53.58
3	BIND	65.43	53.27	71.86	62.57	58.73
	DIP	67.16	55.92	72.65	64.29	61.03
	INTACT	54.83	34.86	71.28	53.07	42.62
	MINT	59.94	41.45	72.29	56.86	49.01
	HPRD	63.21	48.26	71.91	60.08	54.73
	ALL	62.92	47.70	71.89	59.79	54.26
4	BIND	61.88	62.37	61.58	61.98	62.13
	DIP	61.91	64.29	60.45	62.37	63.08
	INTACT	53.66	46.51	59.85	53.18	49.83
	MINT	58.91	52.30	63.50	57.90	55.41
	HPRD	60.42	58.21	61.86	60.03	59.29
	ALL	60.11	57.54	61.81	59.67	58.80

Presented are for all four models the performance results on each independent set of TF interactions.

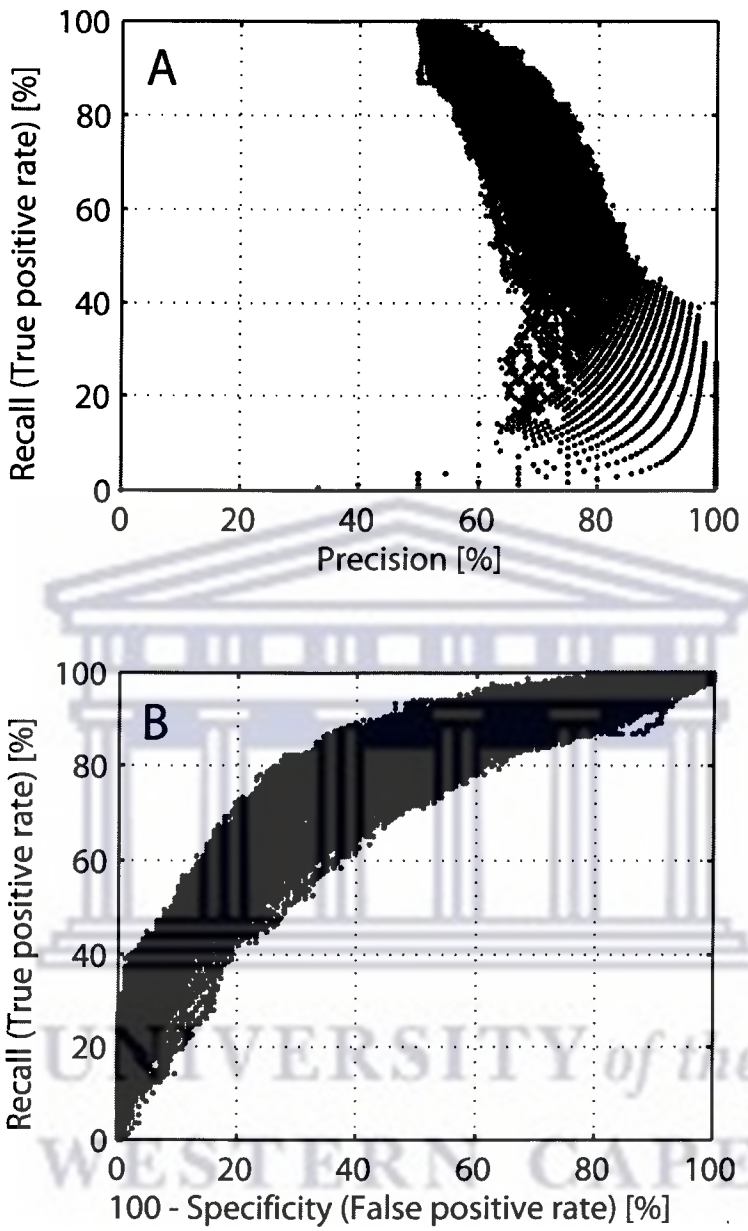


Figure 15. Performance Results of the CV during Model Selection for the Independent Test Sets

A/ Depicted is the precision versus the recall of all 612,522 different parameter combinations tested with the Training-Testing setting on the TRANSFAC derived data with all available features. The typical trade-off between precision and recall is evident. **B/** Presented is the ROC-curve. Each dot represents again a performance result of one parameter combination tested with the Training-Testing setting on the TRANSFAC derived data with all available features. Here the false positive rate is plotted versus the true positive rate.

3.4.3 Performance Evaluation with Randomized Class Labels

To evaluate the chosen models further, and to see if they do not represent only artefacts learned from random data, an additional evaluation step was performed. First the models selected utilising the CV were evaluated. The complete scaled TRANSFAC data set was used to create the models 1 and 2 from the previous section. The same TRANSFAC data was chosen for testing but the labels (positive and negative interaction) beforehand randomly shuffled. This was performed 100,000 times and the average performance over the 100,000 prediction steps calculated. The calculated average performance is presented in Table 7. In the following the models 3 and 4 from the previous section were evaluated. Here, the scaled "Training data" was employed to create the models. These were tested on the scaled "Testing data" with permutated labels. Here, 100,000 different permutations were tested as well and the average performance calculated (see Table 7).

Table 7. Prediction Performance with Randomized Class Labels

Model	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-Measure [%]
1	50.14	50.14	49.84	49.99	50.14
2	50.14	49.99	49.99	49.99	50.06
3	50.15	55.19	44.81	50.01	52.55
4	50.15	61.13	38.87	50.03	55.10

The table presents the performance of the four selected models on their individual test data with randomized class labels. The class labels have been randomized 100.000 times and the average performance measures are depicted. As expected is the performance around 50%.

3.4.4 Feature Selection Based on the Mahalanobis Distance

A feature selection was performed to find a smaller group of features that is able to distinguish the positive and negative TF interactions in a similar fashion as when utilising the complete set of available features. After randomly splitting the whole set of TRANSFAC derived feature vectors into half, the feature selection was performed on one half ("feature selection set"), and evaluated on the other half ("evaluation set"). Thus, the feature selection was run on 337 vectors (see Methods) and produced a subset of 34 features (see Table 8). The 34 features were extracted for each vector of the "evaluation set". Furthermore, the "evaluation set" was split into ten groups for CV purposes, preserving the ratio of positives to negatives in each group. Each CV-training group was scaled and the respective CV-testing groups scaled respectively. Afterwards, the model selection and performance evaluation was done with a 10-fold CV, utilising the same brute-force grid search as mentioned earlier. Thus, the performance of the sub-selected features was evaluated with a 10-fold CV on the "evaluation set" that was not utilised to select the features in the first place.

The precision versus the recall of all tested parameter combinations is presented in Figure 16A. The ROC curve for the same parameter combinations is presented in Figure 16B. The best achievable accuracy was 76.53%. The best F-measure was found to be 75.51%.

The model selection process with the reduced feature set and only half the TRANSFAC data ran on a Linux Pentium 4 core duo machine with a 1.8 GHz CPU and 2GB of memory in ~7 hours. 612,522 different parameter combinations were tested and evaluated.

Table 8. Features Selected using the Mahalanobis Distance

TF	AAIndex identifier	AAIndex description
TF1	ARGP820102	Signal sequence helical potential
TF1	AURR980103	Normalized positional residue frequency at helix termini N"
TF1	CHOP780206	Normalized frequency of N-terminal non helical region
TF1	CHOP780211	Normalized frequency of C-terminal non beta region
TF1	FAUJ880112	Negative charge
TF1	GARJ730101	Partition coefficient
TF1	GEIM800103	Alpha-helix indices for beta-proteins
TF1	GUOD860101	Retention coefficient at pH 2
TF1	KANM800102	Average relative probability of beta-sheet
TF1	MEIH800103	Average side chain orientation angle
TF1	NAKH900106	Normalized composition from animal
TF1	RACS770101	Average reduced distance for C-alpha
TF1	RICJ880106	Relative preference value at N3
TF1	WIMW960101	Free energies of transfer of AcWI-X-LL peptides from bilayer interface to water
TF1	WOLR810101	Hydration potential
TF2	AURR980102	Normalized positional residue frequency at helix termini N'"
TF2	BHAR880101	Average flexibility indices
TF2	CEDJ970105	Composition of amino acids in nuclear proteins (percent)
TF2	CHOP780204	Normalized frequency of N-terminal helix
TF2	CHOP780214	Frequency of the 3rd residue in turn
TF2	ISOY800105	Normalized relative frequency of bend S
TF2	NAKH920101	AA composition of CYT of single-spanning proteins
TF2	OOBM850102	Optimized propensity to form reverse turn
TF2	OOBM850103	Optimized transfer energy parameter
TF2	PALJ810113	Normalized frequency of turn in all-alpha class
TF2	PONP800106	Surrounding hydrophobicity in turn
TF2	PRAM820103	Correlation coefficient in regression analysis
TF2	QIAN880123	Weights for beta-sheet at the window position of 3
TF2	RACS820107	Average relative fractional occurrence in A0(i-1)
TF2	RICJ880103	Relative preference value at N-cap
TF2	RICJ880107	Relative preference value at N4
TF2	TANS770106	Normalized frequency of chain reversal D
TF2	TANS770109	Normalized frequency of coil
TF2	WERD780103	Free energy change of alpha(Ri) to alpha(Rh)

The table presents the 34 sub-selected features derived through the approach that was based on the Mahalanobis distance (see Methods). The first column indicated the position in the features vector. Each feature appears twice in the feature vector representation for a TF interaction (once for each TF comprising the interaction). If the entry in the column indicates "TF1" ("TF2"), then the feature was selected from the first (second) 531 features of the first (second) TF. The second column contains the AAIndex identifier for the AA property. The third column contains a short description of the property taken from the AAIndex database.

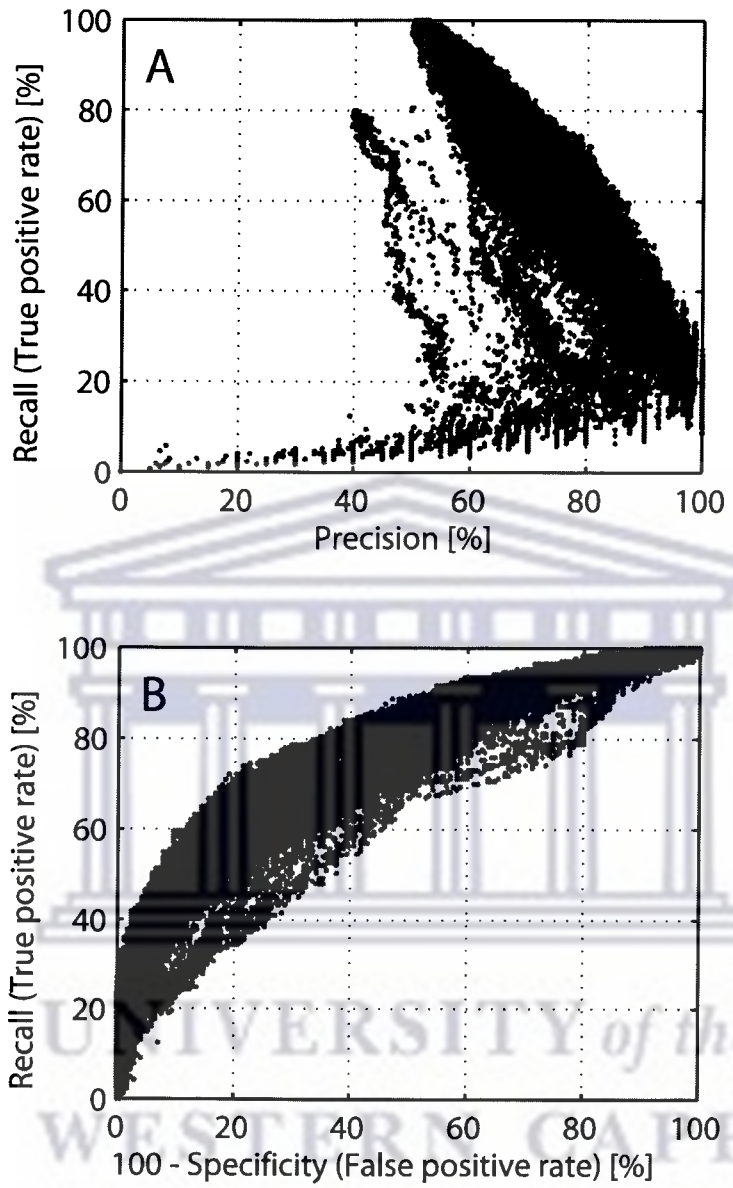


Figure 16. Performance Results of the CV with Selected Features Extracted through Mahalanobis Distance

A/ Depicted is the precision versus the recall of all 612,522 different parameter combinations tested during the 10-fold CV on the TRANSFAC derived "evaluation set" with 34 features, selected through the Mahalanobis distance. The typical trade-off between precision and recall is evident. **B/** Presented is the ROC-curve. Each dot represents again a performance result of one parameter combination tested during the 10-fold CV on the TRANSFAC derived "evaluation set" with 34 features, selected through the Mahalanobis distance. Here the false positive rate is plotted versus the true positive rate.

3.4.5 Feature Selection Based on t-Statistic

The second feature selection approach utilised the t-statistic (see Methods) to calculate for each of the 1,062 features the t-value. Again, the complete TRANSFAC set of TF interactions was randomly split into half. The feature selection was performed on one half of the data ("feature selection set") and the model selection and performance evaluation on the other half of the data ("evaluation set"). The ordered t-values of all 1,062 features are presented in Figure 17A. The distribution of t-values is shown in Figure 17B. The top 25 features with highest t-value where selected from the ranked list of features (see Table 9). The 25 features for each of the 337 feature vectors of the "evaluation set" were sub-selected. The vectors were subsequently divided into 10 equal sized groups for CV purposes while preserving the ratio of positive to negative examples in each set. Scaling of each CV-training and CV-testing set was performed as described above. Finally, the brute-force grid search with a 10-fold CV was run. The results in form of the precision vs. recall plot and the ROC-curve of the grid search are presented in Figure 18A and Figure 18B. The best accuracy and F-measure achieved during CV was 72.15% and 73.01%, respectively. The CV with the sub-selected features and half the TRANSFAC data ran on a Linux Pentium 4 core duo machine with a 1.8 GHz CPU and 2GB of memory in ~8 hours. 612,522 different parameter combinations were tested and evaluated.

Table 9. Features Selected using t-statistic

TF	t-Value	AAIndex identifier	AAIndex description
TF2	6.5983	RACS820107	Average relative fractional occurrence in A0(i-1)
TF2	6.4810	OOBM850102	Optimized propensity to form reverse turn
TF2	6.1823	VINM940103	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbours
TF2	6.1205	ZASB820101	Dependence of partition coefficient on ionic strength
TF2	5.9698	PALJ810112	Normalized frequency of beta-sheet in alpha/beta class
TF2	5.8427	BAEK050101	Linker index
TF1	5.7012	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water
TF2	5.6225	CASG920101	Hydrophobicity scale from native protein structures
TF1	5.4939	ZASB820101	Dependence of partition coefficient on ionic strength
TF1	5.4921	NAKH920103	AA composition of EXT of single-spanning proteins
TF2	5.4850	SUYM030101	Linker propensity index
TF1	5.4611	FUKS010104	Surface composition of amino acids in nuclear proteins (percent)
TF2	5.3822	SNEP660103	Principal component III
TF2	5.3736	NADH010105	Hydropathy scale based on self-information values in the two-state model (25% accessibility)
TF2	5.3367	GEOR030104	Linker propensity from 3-linker dataset
TF1	5.2824	SUYM030101	Linker propensity index
TF2	5.2719	GUYH850102	Apparent partition energies calculated from Wertz-Scheraga index
TF1	5.1449	ROBB760111	Information measure for C-terminal turn
TF2	5.1321	WERD780101	Propensity to be buried inside
TF1	5.1304	FUKS010112	Entire chain composition of amino acids in nuclear proteins (percent)
TF1	5.1222	CASG920101	Hydrophobicity scale from native protein structures
TF2	5.0781	QIAN880122	Weights for beta-sheet at the window position of 2
TF2	5.0686	GARJ730101	Partition coefficient
TF1	5.0659	VINM940101	Normalized flexibility parameters (B-values), average
TF1	5.0147	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbours

The table presents the 25 sub-selected features derived through the approach that was based on the t-statistic (see Methods). The first column indicated the position in the features vector. Each feature appears twice in the feature vector representation for a TF interaction (once for each TF comprising the interaction). If the entry in the column indicates "TF1" ("TF2"), then the feature was selected from the first (second) 531 features of the first (second) TF. The second column contains the calculated t-value for the feature. The third column contains the AAIndex identifier for the AA property. The fourth column contains a short description of the property taken from the AAIndex database.

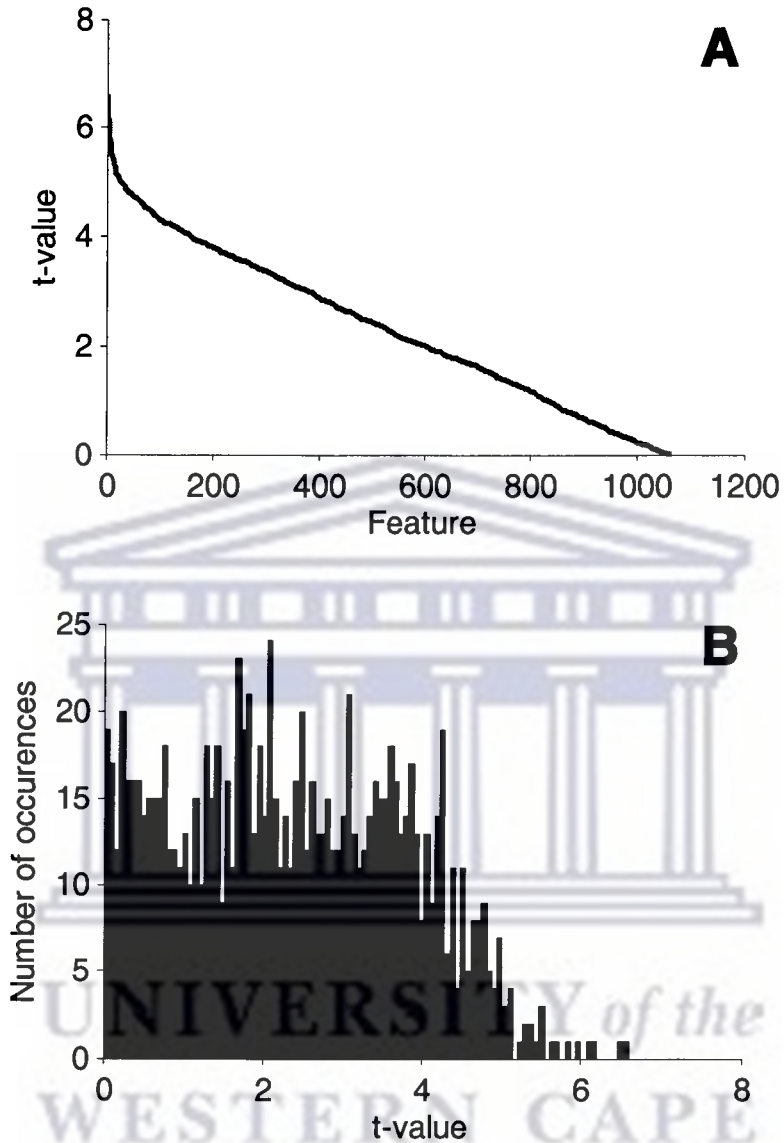


Figure 17. t-statistic Results
A/ Shown is the calculated t-value for each feature, ordered according to the t-value. The figure shows that only a few features have a high t-value. **B/** Shown is the distribution of t-values in form of a histogram with 100 bins.

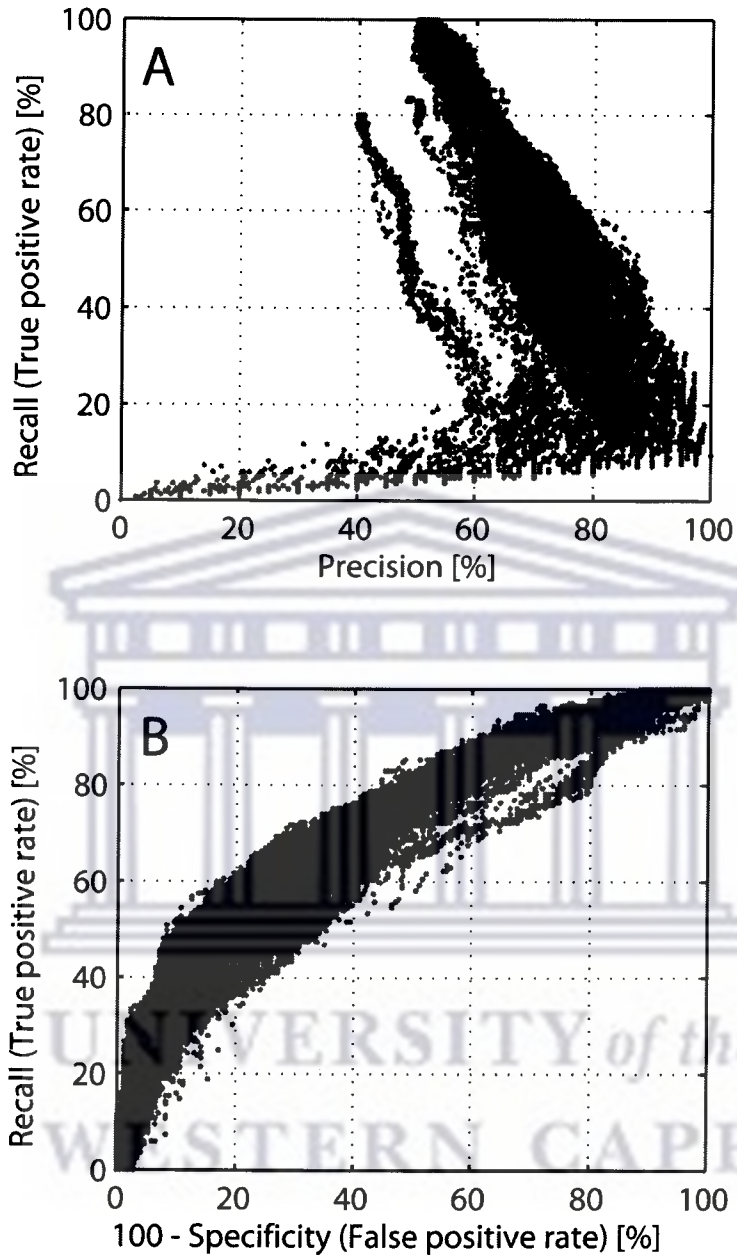


Figure 18. Performance Results of the CV with Selected Features Extracted through t-statistic

A/ Depicted is the precision versus the recall of all 612,522 different parameter combinations tested during the 10-fold CV on the TRANSFAC derived "evaluation set" with 25 features, selected through t-statistic. The typical trade-off between precision and recall is evident. **B/** Presented is the ROC-curve. Each dot represents again a performance result of one parameter combination tested during the 10-fold CV on the TRANSFAC derived "evaluation set" with 25 features, selected through t-statistic. Here the false positive rate is plotted versus the true positive rate.

3.4.6 Performance Evaluation on Independent Sets with Mahalanobis Distance Sub-Selected Features

Because the evaluation of the feature selection was done on the "evaluation set" (see above), the model creation for testing the independent sets was done on this scaled "evaluation set". The complete "evaluation set" with only the 34 selected features, was scaled and the SVM models created. The models that were chosen for evaluation on the independent sets were selected from the CV results of the "evaluation set" (see above). Two models were chosen; the one that achieved the best accuracy ("Model 1") and the one that achieved the best F-measure ("Model 2") during CV (see Table 10). The 34 features from Table 8 were sub-selected for each feature vector of each independent set of TF interactions (see Table 4) and scaled according to the "evaluation set". Afterwards the two models, created with the parameters presented in Table 10, were utilised to classify the TF interactions from each independent set. The results are presented in Table 11.

Table 10. Mahalanobis Distance: Selected Models, their Parameter Combinations, and Performance on their Respective Test Data

Model	c	j	σ	th	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-measure [%]	TP	FP	TN	FN
1	4	1	4	0.10	80.24	72.10	80.99	**76.53**	75.22	12.2	3.2	13.6	4.7
2	2	4	2	0.17	74.09	78.05	70.74	74.46	**75.51**	13.2	4.9	11.9	3.7

Presented are the two chosen models for application on the independent sets of TF interactions. Highlighted numbers indicate the method for selecting the respective model (highest accuracy and highest F-measure). The two models were chosen from the 10-fold CV runs. All performance measures represent averages over the CV runs.

Table 11. Prediction Performance on Independent Data Set with the Features Selected through Mahalanobis Distance

Model	Independent set	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-measure [%]
1	BIND	66.37	33.84	82.85	58.34	44.82
	DIP	61.42	34.67	78.22	56.45	44.32
	INTACT	62.44	29.22	82.43	55.82	39.81
	MINT	64.87	31.68	82.83	57.25	42.57
	HPRD	66.41	32.69	83.47	58.08	43.81
	ALL	66.07	32.63	83.25	57.94	43.68
2	BIND	61.36	43.13	72.84	57.98	50.65
	DIP	60.22	48.26	68.12	58.19	53.58
	INTACT	59.18	40.72	71.91	56.32	48.24
	MINT	61.47	41.14	74.20	57.66	49.29
	HPRD	62.41	42.98	74.11	58.54	50.90
	ALL	62.17	42.81	73.95	58.38	50.70

Presented are for the two selected models the performance results on each independent set of TF interactions. Only the 34 features selected through the Mahalanobis distance were utilised from each independent set of interactions.

3.4.7 Performance Evaluation on Independent Sets with t-Statistic Sub-Selected Features

Two models were chosen from the CV run for further investigation on the independent sets of TF interactions. The models were once again selected based on best average accuracy and F-measure during CV on the t-statistic "evaluation set" (see Table 12). The 25 features, selected based on t-statistic (see Table 9), were extracted for each feature vector of the "evaluation set" of the TRANSFAC data. The complete "evaluation set" was scaled using min-max scaling and the two SVM models created utilising the parameters shown in Table 12. The same 25 features were extracted for each feature vector of the independent sets of TF interactions (see Table 4). Each of the independent sets was scaled according to the "evaluation set" of TF interactions used for creating the models. Subsequently the two models were used to classify each feature vector of the independent sets. The results of the classification are presented in Table 13.

Table 12. t-statistic: Selected Models, their Parameter Combinations, and Performance on their Respective Test Data

Model	c	j	q	th	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-measure [%]	TP	FP	TN	FN
1	4	0.5	8	0.2	83.56	56.88	87.57	**72.15**	67.08	9.6	2.1	14.7	7.3
2	8	1	4	-0.42	64.13	85.29	51.25	68.30	**73.01**	14.4	8.2	8.6	2.5

Presented are the two chosen models for application on the independent sets of TF interactions. Highlighted numbers indicate the method for selecting the respective model (highest accuracy versus highest F-measure). The two models were chosen from the 10-fold CV runs. All performance measures represent averages over the CV runs. This is the reason why the numbers for TP, FP, TN, and FN are floating point numbers.

Table 13. Prediction Performance on Independent Data Set with the Features Selected through t-statistic

Model	Independent set	Precision [%]	Recall [%]	Specificity [%]	Accuracy [%]	F-measure [%]
1	BIND	71.49	32.98	86.85	59.92	45.14
	DIP	72.22	33.97	86.93	60.45	46.21
	INTACT	60.63	18.91	87.72	53.32	28.83
	MINT	64.83	24.29	86.82	55.54	35.34
	HPRD	68.09	27.53	87.09	57.31	39.21
	ALL	67.69	27.33	86.96	57.14	38.94
2	BIND	59.11	62.83	56.55	59.69	60.91
	DIP	58.48	69.69	50.52	60.11	63.59
	INTACT	56.29	60.34	53.14	56.74	58.24
	MINT	57.31	60.00	55.27	57.64	58.62
	HPRD	56.51	61.77	52.45	57.11	59.02
	ALL	56.92	61.78	53.25	57.51	59.25

Presented are for the two selected models the performance results on each independent set of TF interactions. Only the 25 features selected through the t-statistic were utilised from each independent set of interactions.

3.5 Discussion

Despite the variety of experimental techniques to verify PPIs, e.g. protein chips [169], two-hybrid based methods [170], etc., the information gained through these methods covers up to now only a fraction of the PPIs involved in biological processes. Information about experimentally verified TF interactions, which form a subclass of PPIs, is scarce.

Approaches for the deciphering of combinatorial gene regulation included co-expression analysis [145], thermodynamic models based on time-course microarray data [146], or relationships of TFBSs [147,148]. The corporative regulation of genes by multiple TFs requires often the physical interaction of these TFs [6,7]. In order to support future studies that deal with combinatorial gene regulation, the present study implements a computational approach for predicting if specific TFs interact.

The task of predicting TF interactions is comparable to the task of predicting PPIs. Most methods for predicting PPIs need a great deal of information to represent protein pairs and to predict protein synergisms. Often this information is difficult to acquire or is sometimes not readily available. Prediction of PPIs has been done before solemnly from sequence information [131-134] to circumvent the obstacles of the requisition of the multitude of data. Bock *et al.* made use of k-mers of AAs to infer PPIs by AA properties

[131] and Shen *et al.* by k-mer frequencies [133]. The former method made use of a small selected set of AA properties and an undefined method for reducing the feature space of the TF interaction representation to avoid the problem of having vectors of different length, due to different protein sequence lengths. The latter method only focuses on frequencies of AA triads that have been classified into groups of AAs with similar properties, and a newly proposed kernel method to circumvent the problem of symmetry of feature vectors (Protein1-Protein2 equals Protein2-Protein1). Pitre *et al.* utilised the PAM120 similarity matrix to compare and score short AA sequences of individual partners of a hypothetical interaction with the sequences of proteins that are known to interact [132]. Guo *et al.* used a fixed set of seven distinct physicochemical properties to construct feature vectors based on auto covariance and thus circumvent the problem of vectors that differ in length [134]. On the other hand, they did not address the problem of symmetry in protein pairs. Van Dijk *et al.* focused on specific TF families and utilised short motif sequences found in sequences of TFs to predict specific TF interactions with the help of a random forest feature selection approach [135].

Most of these methods for predicting PPIs achieved a prediction accuracy of around 80%. Shen *et al.* achieved 83.9% accuracy for the prediction of human PPI. Van Dijk *et al.* predicted interaction between specific TF families and achieved for different families varying prediction accuracy ranging from 60-90%. The PPI prediction method by Guo *et al.* achieved ~88% in terms of

accuracy, but was solely applied to yeast data [134] and had a very large training base available.

The method implemented here utilises only primary protein sequences to build a representation for a TF interaction pair without any additional prior knowledge to minimise the complexity in feature vector generation. The technique applied here for representing features is based on an averaging scheme of AA properties. It takes the protein sequence as a whole into account and does not put preference onto certain parts of the sequence. Even though the methodology is simple, it might obscure certain domain specific properties, particularly as most parts of the protein sequence are not necessary for the interaction and thus their influence in the averaged values might hamper the performance. Encouragingly though, the distributions of sequence lengths in the utilised positive and negative data sets are similar (see Figures 12 and 14), which leads to the assumption that such an effect would affect both sets in the same manner.

The artificial intelligence system used for the classification of the TF interactions is based on a SVM. SVMs have been extensively utilised in various tasks in computational biology. Examples are, but are not limited to, analysing DNA microarray data [171-173], prediction of protein localisation [174-176], protein secondary structure prediction [177,178], biomedical text mining [179,180], functional gene classification [181], etc.

The performance of the SVM model applied in the present study achieved a prediction accuracy of 80.10% on the TRANSFAC data set as evaluated by 10-fold CV (see Figure 13). When selecting an appropriate model, one can observe the typical trade-off between precision and recall (see Figure 13A). Generally, the performance of the model with the highest achieved accuracy or F-measure is chosen and reported. Different tasks require that e.g. the recall of the model should be a hundred percent, so as to not lose any positives that are within the set. Other tasks, on the other hand, require a precision of one hundred percent so as to be absolutely sure about a positive predicted element. These models are not necessarily the ones with highest accuracy or F-measure. However, here the models with highest accuracy and F-measure were reported (see Table 5) to allow for a comparison with the former approaches for PPI and TF interaction prediction. The performance of SVMs depends heavily on the selected SVM parameters utilised to train the model. The approach of selecting these parameters with a brute-force grid search, as applied in the current study, is common practice but unfortunately does not ensure that the best parameter set is found and in addition requires significant computational resources. Other, more elaborate, methods for selecting parameters, including genetic algorithms [182], or simulated annealing techniques [183]. etc., are much more difficult to apply to problems where the parameter space is restricted as in the present case (e.g. no negative values allowed for SVM parameters c and j; different value ranges per parameter). In addition, the number of parameters that can be

adjusted within a SVM setting is relatively small. For example, the application of a genetic algorithm would involve the adjustment of genetic algorithm specific parameters, such as the mutation rate, crossover probability, number of individuals, etc., which introduces another layer of complexity.

In order to additionally evaluate the performance of the methodology, it was applied to independent data sets. The main obstacle here was the creation of feature vectors for the independent sets of positive and random negative TF interactions. The positive TF interactions were extracted from several databases, while individual TFs were represented through their Entrez gene identifiers. The method employed here utilises the primary protein sequences of the interacting TFs to represent an interaction. The representation of a TF through its Entrez gene identifier hampered the assignment of the protein sequences to the TFs. A gene identifier can be associated to several different protein sequences for the same gene, stemming from e.g. different isoforms or splice products. Even though the protein sequences extracted for one TF's gene identifier might be very similar to each other, it is by no means obvious which of the sequences does take part in the reported interaction. Here, the process denoted all possible sequence combinations of two TFs as positive. This problem had a negative effect on the performance of the applied models. Not all sequence combinations for a positive TF interaction are interacting. The process of automatic protein sequence assignment to the gene identifiers of the TFs is not able to ascertain which of the possible sequence combinations is the

one reported to interact. The performance of the models on the independent sets of TF interactions was thus impacted for these reasons, resulting in a considerably lower prediction performance than that shown through CV on the TRANSFAC data (see Table 5 and Table 6). The selection of an appropriate model for the classification task is equally crucial. Here, a general approach was applied to select the models for classifying the independent sets. One approach took the model with highest accuracy; the other approach chose the model with highest F-measure. Either approach selected individual models with specific characteristics in terms of precision, recall, and specificity (see Table 5).

Class label randomisation was performed on the data used for model evaluation, to test if the selected models are problem specific and do not stem from random artefacts. Given a sufficient amount of such permuted data, one can estimate the performance of the created models on random data and evaluate their prediction performances on the actual data. The expectation is that the models perform worse on the random labelled data than on the original one. This expectation held true for all 4 selected models (see Table 7). When looking at specific performance measures of the models under consideration, one can estimate their performance on the data with randomised class labels. These expectations are not fulfilled for model 1 and 2. The expected specificity for model 1 is 61 % with an expected recall of 39%. The discrepancy in numbers comes from the different data utilised for the model selection and the

randomisation (CV folds vs. complete set of vectors, see Results). Even though the results in Table 7 represent exactly the expectations for model 3 and 4, one can observe a small bias towards the recall in these models. The models 3 and 4, as presented in Table 5, predict more interactions as positives as opposed to negatives. These ratios are reflected in the results when randomisation of class labels was applied. Nevertheless, the results show that all models perform on random data badly, which confirms that the models were not randomly selected but that they incorporate specific properties of the data utilised.

SVMs are known to handle large feature sets well. In order to minimize the computational resources required and to improve the speed of computation, a feature selection step makes sense. The aim is to find a subset of features which performs the classification task in a similar fashion as when performing the analysis using all features. Such a sub-selected group of features has two advantages. First, by reducing the number of features, the whole system becomes much faster during the training, testing, and model selection process. The second advantage is based on the assumption that most features do not hold much information to distinguish the groups of positives and negatives and, thus, a model based on a subset of features becomes more reliable.

Two distinct feature selection algorithms, based on the Mahalanobis distance and on t-statistic, were employed. To gain a fair evaluation of the performance of the sub-selected set of features, it is important that vectors that were utilised

for selecting the features do not take part in the evaluation step. Thus, in both feature selection trials, only half of the data (337 random feature vectors) were used to select a subset of features. The sub-selected features were then evaluated using the other half of the data set. Using the Mahalanobis distance, a set of 34 features was sub-selected by the algorithm (see Methods). After creation of feature vectors with only the sub-selected 34 features (see Table 8), a CV was performed to evaluate the performance of models with varying SVM parameter combinations (see Figure 16). The best achievable accuracy and F-measure with the reduced feature set is 76.53% and 75.51%, respectively. The interpretation of the selected feature set (see Table 8) is difficult in so far as only the combination of these features led to their selection. The focus on individual features within the set might lead to misinterpretations.

Features selected based on t-statistic are associated with a calculated t-value that represents a feature's value in being able to separate the two classes of positive and negative interactions. The calculation of the t-value is independent for each feature, as opposed to the feature selection using the Mahalanobis distance. In Figures 17A and 17B it can be observed that only few features have high t-values. However, the selection of a feature subset can be done in multiple ways. Here, all features were selected that had a t-value of over six. This selection was arbitrary. The only requirement that should be met was that the subset of features should be appropriately small to achieve a highly reduced computation time for the model selection and performance evaluation.

Nevertheless, a different cut-off value based on the t-value for the inclusion of a feature into the subset of features might result in a different performance. A total of 25 features were selected (see Table 9), the respective feature vectors created, and the performance during a 10-fold CV evaluated (see Figure 18). The best achieved accuracy and F-measure is 72.15% and 73.01%, respectively.

Two models were created for each, the Mahalanobis distance based feature subset and the t-statistic based feature subset, using the SVM parameter combinations that led to highest accuracy and F-measure as shown through CV (see Table 10 and Table 12). These models were used to classify the TF interactions of the independent sets and their performance was calculated (see Table 11 and Table 13). The same reasons as discussed before (when utilising all features on the independent sets of TF interactions) led to a huge reduction in all performance measures, with the exception of specificity.

The interpretation of the selected features in both cases using Mahalanobis distance and t-statistic is difficult. All AA properties have influence on the conformation of the protein structure, which in turn affects the proteins ability to interact with other proteins. Interestingly, in the t-statistic subset of features the feature "linker propensity" is often represented. Linker region are regions in the sequence that link protein domains. The multiple occurrences of these

linker features could suggest that TFs with multiple domains are more likely to interact.

Nevertheless, the reduction in performance due to the feature selection, the doubts about their general applicability, and the characteristic of SVMs which handles large feature sets well, suggests a classification with the complete feature set for further application.

Three major problems were identified that occur while predicting either PPIs or TF interactions from sequence data alone:

i/ Symmetry problem of representing pairs of interacting proteins.

ii/ Different feature-vector lengths, due to different protein sequence lengths.

iii/ Missing negative set of protein interactions for training an appropriate model.

The first two problems deal with the representation of features, while the last one affects the artificial intelligence system employed for classification. The present approach utilised a stringent methodology for the representation of a pair of TFs, thus not having a symmetrical effect while creating TF pairs (see Methods). Protein sequences of TFs vary in their length (see Figures 12 and 14). In this study, a representation that is not dependent on the length of the

AA sequence of a TF was implemented (see Methods). The nature of the feature representation approach utilised ensures that a feature vector representation for any pair of TFs is always of the same size. The problem that still exists is the rare number of training examples, positive as well as negative interaction pairs. The lack of datasets of non-interacting TFs is a huge disadvantage. The same obstacles as in the PPI prediction task are evident [184]. Just tuning parameters of machine learning algorithms is not sufficient to compensate for missing real negative examples, which are necessary to train the classification system properly. One common practice is to choose random negatives [133,134,185,186] and/or negative interaction partners that are not functional in the same cellular compartment [184,187]. The latter approach is not applicable in the case of predicting TF interactions, due to the localisation of TFs in the nucleus where they are functional. The random selection of negative examples in the present study, tried to cover cases of negative TF pairs with varying prediction complexity (see Methods). The random selection of negatives TF pairs has its limitations. The performance of the system as presented here might not reflect the real performance, because the as negative denoted interactions can be contaminated with positives that are neither yet experimentally verified nor contained in the TRANSFAC database. This has an influence on all performance measures, which might be in reality higher. In particular, an in-depth experimental investigation into the group of false positive predictions might be of interest, because these would contain possible new true interactions that are not yet known. Nevertheless, the absence of a

real negative set of non-interacting gene products, has been shown to be the bottleneck in all studies that dealt with either PPI or TF interaction prediction. In addition, during the task of predicting TF interactions, one has to deal with a relatively small set of positive interactions available for training as opposed to the larger number of positive examples in PPI prediction task. This is an additional shortcoming for developing models for predicting TF interactions with higher performance.

It is known that not all residues in a protein are equally important, some are important for function and binding while others can be exchanged without such a loss of function [188]. The parts of a protein that interact with another protein are normally very short (often between 3 and 8 residues) [189]. The present study focused on the complete AA sequence of the TF. Further studies could incorporate methods for predicting the importance of certain AA residues in the sequence [190], e.g. through conservation analysis for protein-protein interfaces [191,192], utilise different classification methods, and investigate more elaborate parameter selection algorithms.

Nevertheless, even though the task of predicting TF interactions is more difficult than the prediction of general PPIs and the representation for a TF pair utilised in the present analysis is much simpler, the method applied here is able to achieve comparable prediction performance results when utilising the complete feature set. The advantage of the method lies in its simplicity of

feature representation. Even though the approach takes much less prior knowledge, such as sequence motifs, domains, a specific set of AA properties, gene expression data, etc., into account it is able to perform the classification task with a reasonable performance.

3.6 Conclusions

The present study investigated the feasibility of computationally predicting interactions between TFs using only their protein sequences. The methodology presented here excels through its simplicity. The advantages over more complex methods for representing PPI or TF interactions in the form of feature vectors for classification through an artificial intelligence system are evident. *A priori* information about single TFs is kept to a minimum. Only primary protein sequence information, the AA sequence, is utilised. The features representation takes available AA properties into account and addresses the problems of symmetry and different sequence lengths in a manner that ensures correctness and simplicity. The major shortcoming of the method is the non-existing negative set of interactions for training an appropriate model. The employed SVM was able to classify positive and random negative TF interaction examples with an accuracy of 80.10% as shown through CV, which is comparable to earlier employed methods for predicting PPIs. The results presented here indicate the potential for further studies that deal with the prediction of TF interactions solely from protein sequences and might impact on studies in combinatorial gene regulation in general.

3.7 Appendix II

Y1 – 531 AA properties utilised for creating the feature vectors.

The table shows in alphabetical order the AA indices utilised in the analysis. The first column holds the AAIndex identifier. The second column holds a short description taken from AAIndex. Columns 3-5 contain the reference of the properties in the form of authors, title of publication, and journal.

Chapter 4

Conclusions

An understanding of the machinery that controls the transcription of genes is of great interest. Studies on transcriptional regulation deal with the identification and investigation of regulatory events that lead to the transcription of a gene or gene group. In the present study, two distinct biological questions within the general scope of transcriptional regulation have been investigated.

The first analysis dealt with the deciphering of regulatory events that lead to the distinct expression of miRNA genes within human monocytic differentiation. Human THP-1 cells were treated with PMA to stimulate differentiation of the cells to macrophage-like cells. Here, computational TFBS predictions were combined with time-course gene expression data for miRNAs and TFs to deduce the underlying regulatory mechanisms. This resulted in three major findings. First of all, a global map was derived of the regulatory machinery that acts upon miRNA genes during monocytic differentiation. Secondly, a set of specific miRNAs that are known to be influenced by PMA stimuli were investigated in detail. Here, information from available scientific literature was combined with the predicted transcriptional regulation to deduce how the regulation of the miRNAs affects monocytic differentiation. Finally, several TFs could be identified that seem to have a central role in regulating miRNA genes during the differentiation process.

The analysis is the first of its kind to computationally elucidate, on a large-scale, how miRNAs transcription is regulated during human monocytic differentiation. It uses expression data to find secondary support for the computational predictions. The study makes valuable predicted suggestions for further validation through biological experiments. Furthermore, it defines a starting point for wet lab scientists that are interested in further studying the regulatory circuitry of miRNAs and its impact on monocytic differentiation.

Three general problems could be identified during the analysis that would impact further studies on miRNA gene regulation. First of all, promoter regions for miRNA genes are not well defined and incomplete. The promoter set used in the present analysis is one of the first characterised sets of miRNA promoter regions [69]. In general, the promoters utilised here are very short and can only be considered as partial promoter regions [5]. Other regulatory regions such as proximal and distal promoter regions, enhancers, or silencers may also impact the regulation of any gene, including miRNA genes [4]. The second problem that affects all computational studies on transcriptional regulation is the incomplete set of PWMs/PFMs for the actual prediction of TFBSs. Many known TFs are not represented through a PWM/PFM, because there are either no or too few experimentally validated binding sites of the TF known for the introduction of such a matrix. On the other hand, even the PWMs/PFMs that are currently available are, in many cases, only comprised of incomplete

information on the biologically utilised binding sites in living organisms. Thus, the computational predictions can only be as accurate as the imperfect PWMs/PFMs allow for. The third problem is more specific towards the present study. Expression data for both TFs and miRNAs was used to evaluate the predicted regulations. To be most accurate, the inclusion of an expression time-course series was handled in a strict manner. Hence, several TFs and miRNAs had to be excluded from the analysis, because of inconsistencies in the measured data. This had an effect on the completeness of the analysis with respect to the miRNAs and their regulating TFs.

Follow up studies could include the biological experimental validation of the top predicted associations. This would show to what extent the methodology is able to deduce the regulations computationally. Furthermore, it would show where the methodology could be further improved to enhance future computational studies of this type. With the experimental discovery of more detailed promoter regions for miRNAs, the same kind of computational analysis could be repeated to get an even more detailed picture of the TFs involved in regulating miRNA genes.

Studies on transcriptional regulation frequently involve the analysis of the cooperative functioning of TFs in regulating a gene's transcription. This cooperative function is often mediated through the physical interaction of the TFs [6,7]. In the third chapter, a study is presented whose focus is on

predicting these interactions computationally. Previous studies on computational PPI prediction incorporated various properties of TFs to represent an interaction of two TFs in a format that can be further processed by a computer. The information about the TFs is often difficult to acquire or not available at all. To circumvent these problems, previous studies concentrated on predicting PPIs from sequence data alone [131-134]. An approach for predicting TF interactions based only on sequence information is presented in Chapter 3. Here, available properties of AAs were combined to represent a protein sequence in the form of a vector. An interaction of two TFs is represented by combining two of these vectors. These are fed into an artificial intelligence system to create a computational model which enables the classification into interacting and non-interacting TF pairs. The outcome of this analysis indicates that such a computational classification is possible but with certain restrictions in performance.

As mentioned previously, the major problem that influences all studies on the prediction of PPIs or TF interactions is the missing set of negative examples. Such a set would be invaluable in training an appropriate model and would significantly enhance the performance of the system. Another problem is the relatively small set of known TF interactions. A sufficiently large set of examples is necessary to create a model with good generalisation properties.

Besides enhancing the prediction methodology, future subsequent studies could include a combination of the method for TF interaction prediction with a TFBS analysis of promoter regions of genes. It would be of interest to see how TFBSs of TFs that are known to interact or are predicted to interact, are distributed in regulatory regions. In addition, new TF interactions might be identified that are important for the control of certain genes.

However, the results of the analysis affect further studies on combinatorial gene regulation. It suggests that the utilisation of this kind of analysis in the scope of combinatorial gene regulation would make valuable contributions towards the elucidation of the machinery that controls the transcription of genes.

References

1. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2008, **36:**D475-D479.

2. Stein L: **Genome annotation: from sequence to biology.** *Nat Rev Genet* 2001, **2:**493-503.

3. Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227:**561-563.

4. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes.** *Annu Rev Genet* 2000, **34:**77-137.

5. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72:**449-479.

6. Lemon B, Tjian R: **Orchestrated response: a symphony of transcription factors for gene control.** *Genes Dev* 2000, **14:**2551-2569.

7. Remenyi A, Scholer HR, Wilmanns M: **Combinatorial control of gene expression.** *Nat Struct Mol Biol* 2004, **11:**812-815.

8. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61:**1053-1095.

9. Choo Y, Klug A: **Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions.** *Proc Natl Acad Sci U S A* 1994, **91:**11168-11172.

10. Tuerk C, Gold L: **Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.** *Science (New York, N Y)* 1990, **249:**505-510.

11. Choo Y, Klug A: **Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage.** *Proc Natl Acad Sci U S A* 1994, **91:**11163-11167.

12. Horak CE, Snyder M: **ChIP-chip: a genomic approach for identifying transcription factor binding sites.** *Methods Enzymol* 2002, **350:**469-483.

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

14. Hoeglund A, Kohlbacher O: **From sequence to structure and back again: approaches for predicting protein-DNA binding.** *Proteome Sci* 2004, **2:**3.

15. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5:**276-287.

16. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K et al.: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34:**D108-D110.

17. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, OhnhÃ¤user R et al.: **The TRANSFAC system on gene expression regulation.** *Nucleic Acids Res* 2001, **29:**281-283.

18. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-D94.

19. Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV: **enoLOGOS: a versatile web tool for energy normalized sequence logos.** *Nucleic Acids Res* 2005, **33:**W389-W392.

20. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.

21. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCHTM: a tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.

22. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23:**4878-4884.

23. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21:**2933-2942.

24. Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison.** *Nucleic Acids Res* 2004, **32:**W249-W252.

25. Elnitski L, Jin VX, Farnham PJ, Jones SJM: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.** *Genome Res* 2006, **16:**1455-1464.

26. Hannenhalli S, Wang LS: **Enhanced position weight matrices using mixture models.** *Bioinformatics* 2005, **21:**i204-i212.

27. Hannenhalli S: **Eukaryotic transcription factor binding sites--modeling and integrative search methods.** *Bioinformatics* 2008, **24:**1325-1331.

28. Jensen ST, Liu XS, Zhou Q, Liu JS: **Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective.** *Stat Sci* 2004, **19:**188-204.

29. Kolchanov NA, Merkulova TI, Ignatieva EV, Ananko EA, Oshchepkov DY, Levitsky VG, Vasiliev GV, Klimova NV, Merkulov VM, Charles Hodgman T: **Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes.** *Brief Bioinform* 2007, **8:**266-274.

30. Levitsky V, Ignatieva E, Ananko E, Turnaev I, Merkulova T, Kolchanov N, Hodgman T: **Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions.** *BMC Bioinformatics* 2007, **8**.

31. Merkulova TI, Oshchepkov DY, Ignatieva EV, Ananko EA, Levitsky VG, Vasiliev GV, Klimova NV, Merkulov VM, Kolchanov NA: **Bioinformatical and experimental approaches to investigation of transcription factor binding sites in vertebrate genes.** *Biochemistry (Mosc)* 2007, **72:**1187-1193.

32. Siggia ED: **Computational methods for transcriptional regulation.** *Curr Opin Genet Dev* 2005, **15:**214-221.

33. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.

34. Vavouri T, Elgar G: **Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both.** *Curr Opin Genet Dev* 2005, **15:**395-402.

35. Werner T: **Models for prediction and recognition of eukaryotic promoters.** *Mamm Genome* 1999, **10:**168-175.

36. Zhou Q, Liu JS: **Extracting sequence features to predict protein-DNA interactions: a comparative study.** *Nucleic Acids Res* 2008, **36:**4137-4148.

37. Fickett JW, Wasserman WW: **Discovery and modeling of transcriptional regulatory regions.** *Curr Opin Biotechnol* 2000, **11:**19-24.

38. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ et al.: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-144.

39. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75:**843-854.

40. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116:**281-297.

41. Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.** *Molecular Cell* 2007, **Vol 27:**91-105.

42. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II.** *EMBO J* 2004, **23:**4051-4060.

43. van Furth R, Cohn ZA, Hirsch JG, Humphrey JH, Spector WG, Langevoort HL: **The mononuclear phagocyte system: a new classification of macrophages, monocytes, and their precursor cells.** *Bull World Health Organ* 1972, **46:**845-852.

44. Tsuchiya S, Yamabe M, Yamaguchi Y, Kobayashi Y, Konno T, Tada K: **Establishment and characterization of a human acute monocytic leukemia cell line (THP-1).** *Int J Cancer* 1980, **26:**171-176.

45. Auwerx J: **The human leukemia cell line, THP-1: a multifacetted model for the study of monocyte-macrophage differentiation.** *Experientia* 1991, **47:**22-31.

46. Traore K, Trush MA, George M, Spannhake EW, Anderson W, Asseffa A: **Signal transduction of phorbol 12-myristate 13-acetate (PMA)-**

induced growth inhibition of human monocytic leukemia THP-1 cells is reactive oxygen dependent. *Leuk Res* 2005, **29:**863-879.

47. Martinez FO, Gordon S, Locati M, Mantovani A: **Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression.** *J Immunol* 2006, **177:**7303-7311.

48. Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJ et al.: **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nat Genet* 2009.

49. Lee CT, Risom T, Strauss WM: **MicroRNAs in mammalian development.** *Birth Defects Res C Embryo Today* 2006, **78:**129-139.

50. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA Targets.** *PLoS Biology* 2004, **2.**

51. Lewis BP, Shih Ih, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115:**787-798.

52. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120:**15-20.

53. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression.** *Nucl Acids Res* 2008, **36:**D149-D153.

54. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathway analysis.** *BMC Bioinformatics* 2007, **8:**69.

55. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucl Acids Res* 2006, **34:**D140-D144.

56. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucl Acids Res* 2008, **36:**D154-D158.

57. Bracht J, Hunter S, Eachus R, Weeks P, Pasquinelli AE: **Trans-splicing and polyadenylation of let-7 microRNA primary transcripts.** *RNA* 2004, **10:**1586-1594.

58. Cai X, Hagedorn CH, Cullen BR: **Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.** *RNA* 2004, **10:**1957-1966.

59. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ: **Processing of primary microRNAs by the Microprocessor complex.** *Nature* 2004, **432:**231-235.

60. Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R: **The Microprocessor complex mediates the genesis of microRNAs.** *Nature* 2004, **432:**235-240.

61. Bohnsack MT, Czaplinski K, Gorlich D: **Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs.** *RNA* 2004, **10:**185-191.

62. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, RÃ¥dmark O, Kim S et al.: **The nuclear RNase III Drosha initiates microRNA processing.** *Nature* 2003, **425:**415-419.

63. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs.** *Science (New York, N Y)* 2001, **294:**853-858.

64. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.** *Science (New York, N Y)* 2001, **294:**858-862.

65. Lee Y, Jeon K, Lee JT, Kim S, Kim VN: **MicroRNA maturation: stepwise processing and subcellular localization.** *EMBO J* 2002, **21:**4663-4670.

66. Xie Z, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC: **Expression of Arabidopsis MIRNA Genes.** *Plant Physiol* 2005, **138:**2145-2154.

67. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129:**823-837.

68. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA: **A chromatin landmark and transcription initiation at most promoters in human cells.** *Cell* 2007, **130:**77-88.

69. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J et al.: **Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells.** *Cell* 2008, **134:**521-533.

70. Arkin A, Shen P, Ross J: **A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements.** *Science* 1997, **277:**1275-1279.

71. Shi Y, Mitchell T, Bar-Joseph Z: **Inferring pairwise regulatory relationships from multiple time series datasets.** *Bioinformatics* 2007, **23:**755-763.

72. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression Analysis of Human Genes Across Many Microarray Data Sets.** *Genome Res* 2004, **14:**1085-1094.

73. Redestig H, Weicht D, Selbig J, Hannah M: **Transcription factor target prediction using multiple short expression time series from Arabidopsis thaliana.** *BMC Bioinformatics* 2007, **8:**454.

74. Schmitt WA, Raab RM, Stephanopoulos G: **Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data.** *Genome Res* 2004, **14:**1654-1663.

75. Forrest ARR et al.: **FANTOM4: A network of induced microRNAs promotes myeloid differentiation.** Unpublished.

76. Fritsch FN, Carlson RE: **Monotone Piecewise Cubic Interpolation.** *SIAM J Numerical Analysis* 1980, **17:**238-246.

77. Kahaner DK, Moler C, Nash SG: *Numerical Methods and Software.* Prentice-Hall; 1988.

78. Ravasi T, Katayama S, Bajic VB, Tan K, Schmeier S, Kanamori-Katayama M, Bertin N et al.: **An atlas of combinatorial transcriptional regulation in mouse and man.** Unpublished.

79. Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, Hume D, Quackenbush J: **Data-driven normalization strategies for high-throughput quantitative RT-PCR.** *BMC Bioinformatics* 2009, **10**.

80. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

81. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al.: **The UCSC Genome Browser Database: update 2006.** *Nucl Acids Res* 2006, **34:**D590-D598.

82. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32:**D109-D111.

83. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21:**3439-3440.

84. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.

85. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T et al.: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-D484.

86. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**3.

87. Woods K, Thomson JM, Hammond SM: **Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors.** *J Biol Chem* 2007, **282:**2130-2134.

88. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M: **Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions.** *J Mol Biol* 2001, **314:**1053-1066.

89. Wu WS, Li WH, Chen BS: **Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data.** *BMC Bioinformatics* 2007, **8:**188.

90. Yu H, Luscombe NM, Qian J, Gerstein M: **Genomic analysis of gene expression relationships in transcriptional regulatory networks.** *Trends Genet* 2003, **19:**422-427.

91. Valledor AF, BorrÃ s FE, Cullell-Young M, Celada A: **Transcription factors that regulate monocyte/macrophage differentiation.** *J Leukoc Biol* 1998, **63:**405-417.

92. Sawka-Verhelle D, Escoubet-Lozach L, Fong AL, Hester KD, Herzig S, Lebrun P, Glass CK: **PE-1/METS, an antiproliferative Ets repressor factor, is induced by CREB-1/CREM-1 during macrophage differentiation.** *J Biol Chem* 2004, **279:**17772-17784.

93. Li C, Yu Y, Wang Y, Liu L, Zhang M, Sugano S, Wang Z, Chang Z: **Both ERK and JNK are required for enhancement of MD-2 gene expression during differentiation of HL-60 cells.** *Biol Cell* 2008, **100:**365-375.

94. Gavin IM, Glesne D, Zhao Y, Kubera C, Huberman E: **Spermine acts as a negative regulator of macrophage differentiation in human myeloid leukemia cells.** *Cancer Res* 2004, **64:**7432-7438.

95. Chen N, Szentirmay MN, Pawar SA, Sirito M, Wang J, Wang Z, Zhai Q, Yang HX, Peehl DM, Ware JL et al.: **Tumor-suppression function of transcription factor USF2 in prostate carcinogenesis.** *Oncogene* 2006, **25:**579-587.

96. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M et al.: **TM4: a free, open-source system for microarray data management and analysis.** *BioTechniques* 2003, **34:**374-378.

97. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J: **TM4 microarray software suite.** *Methods Enzymol* 2006, **411:**134-193.

98. Fujita S, Ito T, Mizutani T, Minoguchi S, Yamamichi N, Sakurai K, Iba H: **miR-21 Gene expression triggered by AP-1 is sustained through a double-negative feedback mechanism.** *J Mol Biol* 2008, **378:**492-504.

99. Mollinedo F, Gajate C, Tugores A, Flores I, Naranjo JR: **Differences in expression of transcription factor AP-1 in human promyelocytic HL-60 cells during differentiation towards macrophages versus granulocytes.** *Biochem J* 1993, **294 (Pt 1):**137-144.

100. Meng F, Henson R, Wehbe-Janek H, Ghoshal K, Jacob ST, Patel T: **MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer.** *Gastroenterology* 2007, **133:**647-658.

101. Zhu S, Wu H, Wu F, Nie D, Sheng S, Mo YY: **MicroRNA-21 targets tumor suppressor genes in invasion and metastasis.** *Cell Res* 2008, **18:**350-359.

102. Rosa A, Ballarino M, Sorrentino A, Sthandier O, De Angelis FG, Marchioni M, Masella B, Guarini A, Fatica A, Peschle C et al.: **The interplay between the master transcription factor PU.1 and miR-424 regulates human monocyte/macrophage differentiation.** *Proc Natl Acad Sci U S A* 2007, **104:**19849-19854.

103. Reddy VA, Iwama A, Iotzova G, Schulz M, Elsasser A, Vangala RK, Tenen DG, Hiddemann W, Behre G: **Granulocyte inducer C/EBPalpha inactivates the myeloid master regulator PU.1: possible role in lineage commitment decisions.** *Blood* 2002, **100:**483-490.

104. Chen A, Luo M, Yuan G, Yu J, Deng T, Zhang L, Zhou Y, Mitchelson K, Cheng J: **Complementary analysis of microRNA and mRNA expression during phorbol 12-myristate 13-acetate (TPA)-induced differentiation of HL-60 cells.** *Biotechnol Lett* 2008, **30:**2045-2052.

105. Zeller KI, Zhao X, Lee CWH, Chiu KP, Yao F, Yustein JT, Ooi HS, Orlov YL, Shahab A, Yong HC et al.: **Global mapping of c-Myc binding sites and target gene networks in human B cells.** *Proc Natl Acad Sci U S A* 2006, **103:**17834-17839.

106. Yin Q, Wang X, McBride J, Fewell C, Flemington E: **B-cell receptor activation induces BIC/miR-155 expression through a conserved AP-1 element.** *J Biol Chem* 2008, **283:**2654-2662.

107. Gatto G, Rossi A, Rossi D, Kroening S, Bonatti S, Mallardo M: **Epstein-Barr virus latent membrane protein 1 trans-activates miR-155 transcription through the NF-{kappa}B pathway.** *Nucleic Acids Res* 2008.

108. He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ et al.: **A microRNA polycistron as a potential human oncogene.** *Nature* 2005, **435:**828-833.

109. Sylvestre Y, De Guire V, Querido E, Mukhopadhyay UK, Bourdeau VÃ, Major Fo, Ferbeyre G, Chartrand P: **An E2F/miR-20a autoregulatory feedback loop.** *J Biol Chem* 2007, **282:**2135-2143.

110. Cloonan N, Brown M, Steptoe A, Wani S, Chan W, Forrest A, Kolle G, Gabrielli B, Grimmond S: **The miR-17-5p microRNA is a key regulator of the G1/S phase cell cycle transition.** *Genome Biol* 2008, **9:**R127.

111. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT: **c-Myc-regulated microRNAs modulate E2F1 expression.** *Nature* 2005, **435:**839-843.

112. Helin K, Wu CL, Fattaey AR, Lees JA, Dynlacht BD, Ngwu C, Harlow E: **Heterodimerization of the transcription factors E2F-1 and DP-1 leads to cooperative trans-activation.** *Genes Dev* 1993, **7:**1850-1861.

113. Fontana L, Pelosi E, Greco P, Racanicchi S, Testa U, Liuzzi F, Croce CM, Brunetti E, Grignani F, Peschle C: **MicroRNAs 17-5p-20a-106a control monocytopoiesis through AML1 targeting and M-CSF receptor upregulation.** *Nat Cell Biol* 2007, **9:**775-787.

114. Lu Y, Thomson JM, Wong HYF, Hammond SM, Hogan BLM: **Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells.** *Developmental Biology* 2007, **310:**442-453.

115. Langmann T, Buechler C, Ries S, Schaeffler A, Aslanidis C, Schuierer M, Weiler M, Sandhoff K, de Jong PJ, Schmitz G: **Transcription factors Sp1 and AP-2 mediate induction of acid sphingomyelinase during monocytic differentiation.** *J Lipid Res* 1999, **40:**870-880.

116. Merrill AH: **De Novo Sphingolipid Biosynthesis: A Necessary, but Dangerous, Pathway.** *J Biol Chem* 2002, **277:**25843-25846.

117. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, EngstrÃ¶m PrG, Frith MC et al.: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.

118. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA: **Mammalian RNA polymerase II core promoters: insights from genome-wide studies.** *Nat Rev Genet* 2007, **8:**424-436.

119. Obernosterer G, Leuschner PJ, Alenius M, Martinez J: **Post-transcriptional regulation of microRNA expression.** *RNA* 2006, **12:**1161-1167.

120. GuhaThakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17:**608-621.

121. Banerjee N, Zhang MQ: **Identifying cooperativity among transcription factors controlling the cell cycle in yeast.** *Nucleic Acids Res* 2003, **31:**7024-7031.

122. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ: **Identifying combinatorial regulation of transcription factors and binding motifs.** *Genome Biol* 2004, **5:**R56.

123. Hu Z, Hu B, Collins JF: **Prediction of synergistic transcription factors by function conservation.** *Genome Biol* 2007, **8:**R257.

124. Wang J: **A new framework for identifying combinatorial regulation of transcription factors: a case study of the yeast cell cycle.** *J Biomed Inform* 2007, **40:**707-725.

125. Browne F, Wang H, Zheng H, Azuaje F: **GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction.** *Source Code Biol Med* 2009, **4:**2.

126. Aloy P, Russell RB: **InterPreTS: protein interaction prediction through tertiary structure.** *Bioinformatics* 2003, **19:**161-162.

127. McDowall MD, Scott MS, Barton GJ: **PIPs: human protein-protein interaction prediction database.** *Nucleic Acids Res* 2009, **37:**D651-D656.

128. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K et al.: **PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine.** *BMC Bioinformatics* 2003, **4:**11.

129. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A: **PRISM: protein interactions by structural matching.** *Nucleic Acids Res* 2005, **33:**W331-W336.

130. Wu X, Zhu L, Guo J, Zhang DY, Lin K: **Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations.** *Nucleic Acids Res* 2006, **34:**2137-2150.

131. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure.** *Bioinformatics* 2001, **17:**455-460.

132. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N et al.: **PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.** *BMC Bioinformatics* 2006, **7:**365.

133. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: **Predicting protein-protein interactions based only on sequences information.** *Proc Natl Acad Sci U S A* 2007, **104:**4337-4341.

134. Guo Y, Yu L, Wen Z, Li M: **Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences.** *Nucleic Acids Res* 2008, **36:**3025-3030.

135. van Dijk AD, ter Braak CJ, Immink RG, Angenent GC, van Ham RC: **Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control.** *Bioinformatics* 2008, **24:**26-33.

136. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology.** *Proc Natl Acad Sci U S A* 2002, **99:**5896-5901.

137. Li XL, Tan SH, Ng SK: **Improving domain-based protein interaction prediction using biologically significant negative datasets.** *Int J Data Min Bioinform* 2006, **1:**138-149.

138. Hoskins J, Lovell S, Blundell TL: **An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements.** *Protein Sci* 2006, **15:**1017-1029.

139. Guharoy M, Chakrabarti P: **Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions.** *Bioinformatics* 2007, **23:**1909-1918.

140. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.

141. Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, Kao CY, Huang CY: **Ortholog-based protein-protein interaction prediction and its application to inter-species interactions.** *BMC Bioinformatics* 2008, **9 Suppl 12:**S11.

142. Ramani AK, Li Z, Hart GT, Carlson MW, Boutz DR, Marcotte EM: **A map of human protein interactions derived from co-expression of human mRNAs and their orthologs.** *Mol Syst Biol* 2008, **4:**180.

143. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet* 2004, **36:**664.

144. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36:**492-496.

145. Yu X, Lin J, Zack DJ, Qian J: **Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues.** *Nucleic Acids Res* 2006, **34:**4925-4936.

146. Chen CC, Zhu XG, Zhong S: **Selection of thermodynamic models for combinatorial control of multiple transcription factors in early differentiation of embryonic stem cells.** *BMC Genomics* 2008, **9 Suppl 1:**S18.

147. Hannenhalli S, Levy S: **Predicting transcription factor synergism.** *Nucleic Acids Res* 2002, **30:**4278-4284.

148. Yu X, Lin J, Masuda T, Esumi N, Zack DJ, Qian J: **Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae.** *Nucleic Acids Res* 2006, **34:**917-927.

149. Zhu Z, Shendure J, Church GM: **Discovering functional transcription-factor combinations in the human cell cycle.** *Genome Res* 2005, **15:**848-855.

150. Vapnik VN: *The Nature of Statistical Learning Theory.* New York: Springer-Verlag; 1995.

151. Schölkopf B, Burges CJ, Smola AJ: *Advances in Kernel Methods: Support Vector Learning.* Cambridge: MIT Press; 1998.

152. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: **AAindex: amino acid index database, progress report 2008.** *Nucleic Acids Res* 2008, **36:**D202-D205.

153. Yuan Y, Shaw MJ: **Induction of fuzzy decision trees.** *Fuzzy Sets and Systems* 1995, **69:**125-139.

154. Ho TK: **Random Decision Forests.** In *Proceedings of the 3rd International Conference on Document Analysis and Recognition:* 1995:278-282.

155. Comon P: **Independent component analysis, A new concept?** *Signal Processing* 1994, **36:**287-314.

156. Aizerman A, Braverman EM, Rozoner LI: **Theoretical foundations of the potential function method in pattern recognition learning.** *Automation and Remote Control* 1964, **25:**821-837.

157. Joachims T: **Text Categorization with Support Vector Machines: Learning with Many Relevant Features.** In *Proceedings of ECML-98, 10th European Conference on Machine Learning:* Edited by Nédellec C, Rouveirol C. Heidelberg: Springer-Verlag; 1998:137-142.

158. Mahalanobis PC: **On the generalised distance in statistics.** In *Proceedings of the National Institute of Sciences of India:* 1936:49-55.

159. Gosset WS: **The probable error of a mean.** *Biometrika* 1908, **6:**1-25.

160. Ewens WJ, Grant GR: *Statistical Methods in Bioinformatics.* New York: Springer-Verlag; 2001.

161. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M et al.: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13:**2363-2371.

162. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: **BIND--The Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2001, **29:**242-245.

163. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E et al.: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33:**D418-D424.

164. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513:**135-140.

165. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A et al.: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2004, **32:**D452-D455.

166. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R et al.: **IntAct--open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35:**D561-D565.

167. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32:**D449-D451.

168. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins.** *Nucleic Acids Res* 2000, **28:**289-291.

169. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T et al.: **Global analysis of protein activities using proteome chips.** *Science* 2001, **293:**2101-2105.

170. Fields S, Song O: **A novel genetic system to detect protein-protein interactions.** *Nature* 1989, **340:**245-246.

171. Lee Y, Lee CK: **Classification of multiple cancer types by multicategory support vector machines using gene expression data.** *Bioinformatics* 2003, **19:**1132-1139.

172. Schramm A, Schulte JH, Klein-Hitpass L, Havers W, Sieverts H, Berwanger B, Christiansen H, Warnat P, Brors B, Eils J et al.: **Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling.** *Oncogene* 2005, **24:**7902-7912.

173. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, Biankin AV et al.: **An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin.** *Cancer Res* 2005, **65:**4031-4040.

174. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.

175. Chen YL, Li QZ: **Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition.** *J Theor Biol* 2007, **248:**377-381.

176. Habib T, Zhang C, Yang JY, Yang MQ, Deng Y: **Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition.** *BMC Genomics* 2008, **9 Suppl 1:**S16.

177. Hua S, Sun Z: **A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach.** *J Mol Biol* 2001, **308:**397-407.

178. Wang LH, Liu J, Li YF, Zhou HB: **Predicting protein secondary structure by a support vector machine based on a new coding scheme.** *Genome Inform* 2004, **15:**181-190.

179. Hakenberg J, Schmeier S, Kowald A, Klipp E, Leser U: **Finding kinetic parameters using text mining.** *OMICS* 2004, **8:**131-152.

180. Hakenberg J, Bickel S, Plake C, Brefeld U, Zahn H, Faulstich L, Leser U, Scheffer T: **Systematic feature evaluation for gene name recognition.** *BMC Bioinformatics* 2005, **6 Suppl 1:**S9.

181. Pavlidis P, Weston J, Cai J, Noble WS: **Learning gene functional classifications from multiple data types.** *J Comput Biol* 2002, **9:**401-411.

182. Forrest S: **Genetic algorithms: principles of natural selection applied to computation.** *Science* 1993, **261:**872-878.

183. Kirkpatrick S, Gelatt CD, Jr., Vecchi MP: **Optimization by Simulated Annealing.** *Science* 1983, **220:**671-680.

184. Jansen R, Gerstein M: **Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction.** *Curr Opin Microbiol* 2004, **7:**535-545.

185. Chen XW, Liu M: **Prediction of protein-protein interactions using random decision forest framework.** *Bioinformatics* 2005, **21:**4394-4400.

186. Lo SL, Cai CZ, Chen YZ, Chung MC: **Effect of training datasets on support vector machine prediction of protein-protein interactions.** *Proteomics* 2005, **5:**876-884.

187. Ben Hur A, Noble WS: **Choosing negative examples for the prediction of protein-protein interactions.** *BMC Bioinformatics* 2006, **7 Suppl 1:**S2.

188. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42:**108-124.

189. Kim WK, Henschel A, Winter C, Schroeder M: **The many faces of protein-protein interactions: A compendium of interface geometry.** *PLoS Comput Biol* 2006, **2:**e124.

190. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics* 2007, **23:**1875-1882.

191. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13:**190-202.

192. Guharoy M, Chakrabarti P: **Conservation and relative importance of residues across protein-protein interfaces.** *Proc Natl Acad Sci U S A* 2005, **102:**15447-15452.