COMPUTATIONAL DETECTION OF TRANSCRIPTIONAL REGULATORS OF PROTEIN COMPLEXES IN APOPTOSIS

by

Fred Yefang Peng

B.Sc., Botany, Sun Yat-sen University, 1987

M.Sc., Plant Physiology, Sun Yat-sen University, 1990

Ph.D., Sun Yat-sen University, 1993

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the

Department of

Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Fall 2004

APPROVAL

Name: Fred Yefang Peng

Degree: Master of Science

Title of thesis: Computational Detection of Transcriptional Regulators of Protein Complexes in Apoptosis

Examining Committee:

Chair: Dr. Nancy Hawkins, Assistant Professor
Department of Molecular Biology and Biochemistry

Dr. Frederic Pio
Senior Supervisor, Assistant Professor
Department of Molecular Biology and Biochemistry

Dr. David Baillie
Supervisor, Professor
Department of Molecular Biology and Biochemistry

Dr. Steven Jones
Supervisor, Head of Bioinformatics
Genome Sciences Centre, BC Cancer Research Centre

Dr. Sharon Gorski
Supervisor, Scientist
Genome Sciences Centre, BC Cancer Research Centre

Dr. Peter Unrau
Internal Examiner, Assistant Professor
Department of Molecular Biology and Biochemistry

Date of Approval: September 23, 2004

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

ABSTRACT

Apoptosis is a type of cell death mediated by different signaling pathways involving protein-protein interactions that eventually activate caspases, a family of proteases capable of degrading cellular proteins. In this study we identify genes that belong to 16 protein families known to be involved in apoptosis in 5 vertebrate genomes: human, mouse, rat, *Danio* zebrafish, and *Fugu* pufferfish. It is shown that most apoptotic pathways are conserved in these vertebrate genomes, whereas key genes of the Fas-mediated extrinsic pathway have not yet been identified in zebrafish and pufferfish genomes. Sequence alignment indicates that the upstream regions are less conserved than the corresponding transcript sequences and the sequence identity further declines after masking out the repetitive elements in the upstream sequences. These data are critical for phylogenetic footprinting studies of apoptosis genes in vertebrate genomes.

Based on 366 known protein-protein interactions covering 168 (~72%) human apoptosis genes, we assemble a protein interaction network. To facilitate human visualization and potentially help biologists in apoptosis research, a two-layer protein interaction network is built for each human apoptosis gene. Several known apoptotic complexes, such as apoptosome, DISC, inflammasome and TNFR1 complex, are all visualized in the two-layer interaction networks. We hypothesize that these two-layer protein interaction networks may help infer other multi-protein complexes in apoptosis.

Furthermore, we computationally identify putative transcription factor (TF) binding sites upstream of apoptosis genes, and use protein-protein interactions in conjunction with phylogenetic footprinting information for prediction filtering. Our results suggest that protein-protein interaction data could complement sequence conservation to reduce

false positive predictions. The TF classes STAT, bHSH, paired, SMAD, and bZIP have most predicted binding sites upstream of human apoptosis genes. From the computational analysis and known transcription factor binding sites, we construct a regulatory network for each main apoptotic signaling pathways. These *in silico* networks demonstrate that some transcription factors might regulate several genes involved in the same pathway. Lastly, to make these data available for apoptosis research, we develop a database-driven web site and its URL is http://apoptosis.mbb.sfu.ca/main.php.

ACKNOWLEDGEMENTS

The author is very grateful for the financial support provided by the Canadian Institute for Health Research (CIHR), Michael Smith Foundation for Health Research (MSFHR) and Alfred P. Sloan Foundation Strategic Training Program in Bioinformatics for Health Research, offered through a partnership between the Genome Sciences Centre, Simon Fraser University and the University of British Columbia. This work has also been supported by the Natural Science and Engineering Council of Canada (NSERC) and British Columbia Advanced Systems Institute (ASI).

I wish to thank my senior supervisor, Dr Frederic Pio, for guiding this project in the past year, as well as members of my supervisory committee, Dr David Baillie, Dr Steven Jones and Dr Sharon Gorski for their valuable advice.

I am also indebted to Dr Duncan Napier, the systems and instrumentation consultant in the Department of Molecular Biology and Biochemistry, for his system administrative support on the database server and web server.

Last but not least, I want to thank my family for their support during my study in this bioinformatics program in the past two years.

TABLE OF CONTENTS

APPROVAL ii

ABSTRACT iii

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES ix

LIST OF ABBREVIATIONS x

CHAPTER ONE: INTRODUCTION 1

1.1 Apoptosis and programmed cell death 1

1.2 Apoptotic signaling pathways 2

1.3 Computational identification of transcription factor binding sites 5

1.4 Aim of the thesis 7

CHAPTER TWO: METHODS 10

2.1 Identification of apoptosis genes in mammalian genomes 10

2.2 Identification of apoptosis genes in zebrafish and pufferfish genomes 10

2.3 Retrieval of upstream sequences 11

2.4 Known transcription factors and their binding sites 11

2.5 The known apoptotic protein-protein interactions in humans 12

2.6 Data storage 12

2.7 Identification of conserved regions 12

2.8 TFBS identifications 13

2.9 Statistical analysis 14

2.10 Calculating the expected number of occurrences for each TF class 16

2.11 Predictive performance testing 16

2.12 Construction of apoptotic regulatory networks 21

2.13 Development of the data-driven web site 21

CHAPTER THREE: RESULTS **22**
3.1 Apoptotic signaling pathways in vertebrate genomes 22
3.2 Sequence similarities of apoptosis genes and upstream regions 25
3.3 Human protein-protein interaction networks in apoptosis 27
3.4 TFBS prediction using both phylogeny and interaction data 30
3.5 Distribution of binding sites for each TF class 32
3.6 *In silico* construction of human apoptotic regulatory networks 35
3.7 A data-driven website for apoptosis research 40

CHAPTER FOUR: DISCUSSION AND CONCLUSION **43**
4.1 Usage of protein-protein interaction data for TFBS prediction 43
4.2 Human protein-protein interaction networks and regulatory networks in apoptosis 46

REFERENCES **49**
WEB SITE REFERENCES 56

APPENDICES **57**
Appendix A: The RefSeq accession numbers of apoptosis genes identified in the mammalian genomes. 57
Appendix B: The Ensembl IDs of apoptosis genes identified in the *Danio* and *Fugu* genomes. 66
Appendix C: The human apoptotic protein-protein interaction network 72
Appendix D: An example position frequency matrix (PFM) for each TF class in Fig. 6, its average matrix length and expected number of occurrences in upstream sequences 73

LIST OF FIGURES

Figure 1. The TNFR1-mediated extrinsic apoptotic signaling pathway and related pathways 3

Figure 2. The phylogenetic relationship of the 5 vertebrate species 9

Figure 3. The flow chart of the TFBS prediction approach. 14

Figure 4. Sequence identity of apoptosis transcripts and their upstream regions. 27

Figure 5. A two-layer protein interaction network of human TNFR1 (TNFRSF1A). 30

Figure 6. Average number of predicted transcription factor binding sites in the upstream regions. 34

Figure 7. Distribution of predicted binding sites in different upstream regions. 36

Figure 8. Computationally identified regulatory network for the TNFR1-mediated extrinsic apoptotic-signaling pathway 37

Figure 9. A web-based two-layer protein interaction network for the human Apaf1. 41

Figure 10. The search results of putative transcription factors and binding sites shared by human Apaf1 and Caspase-9. 42

LIST OF TABLES

Table 1. The test data set consisting of 15 genes and 54 experimentally determined transcription-factor binding sites for 26 distinct transcription factors. 18

Table 2. Functions of caspases and homologous genes identified in the 5 vertebrate genomes. 23

Table 3. Key genes involved in the four major apoptotic pathways and homologous genes identified in the 5 vertebrate genomes. 25

Table 4. Highly interacting genes in the protein-protein interaction network of human apoptotic pathways. 28

Table 5. TFBS prediction sensitivity (SN) and specificity (SP) using protein-protein interaction data and phylogenetic footprinting information. 32

Table 6. Potential importance of transcription factors in the regulatory networks of three major apoptotic pathways. 38

LIST OF ABBREVIATIONS

AIF	apoptosis inducing factor
Apaf-1	apoptotic protease-activating factor 1
BH	Bcl-2 homology
BIR	baculovirus IAP repeat
bp	base pair
CARD	caspase recruitment domain
Caspase	cysteine aspartate-specific protease
DD	death domain
DED	death effector domain
DIABLO	direct IAP-binding protein with low isoelectric point (pI)
DISC	death-inducing signaling complex
DR	death receptor
FADD	Fas-associated death domain protein
FLIP	Flice inhibitory protein
IAP	inhibitor of apoptosis protein
IκB	inhibitor of NF-κB
IKK	I-κB kinase complex
JNK	c-Jun NH2-Terminal Kinase
MAPK	mitogen-activated protein kinase
NF-κB	nuclear factor κB
PCD	programmed cell death
PFM	position frequency matrix
PWM	position weight matrix
RAIDD	RIP-associated ICH-1 protein with death domain
RIP	receptor-interacting protein
Smac	second mitochondrial activator of caspases
TF	transcription factor
TFBS	transcription factor binding site
TNF	tumor necrosis factor
TNFSF	tumor necrosis factor superfamily
TNFR	tumor necrosis factor receptor
TNFRSF	tumor necrosis factor receptor superfamily
TRADD	tumor necrosis factor - associated death domain protein
TRAF	tumor necrosis factor receptor - associated factor
TRAIL	tumor necrosis factor α-related apoptosis-inducing ligand
XIAP	X-linked inhibitor of apoptosis protein

CHAPTER ONE: INTRODUCTION

1.1 Apoptosis and programmed cell death

The word "apoptosis" comes from an ancient Greek, meaning the "falling of leaves from a tree in autumn" or "falling of petals from a flower" (Lawen 2003). Apoptosis is now referred to as a type of cell death that orderly and efficiently removes damaged or unnecessary cells in metazoan organisms (for reviews, see Ashe and Berry 2003; Danial and Korsmeyer 2004; Lawen 2003). It is often used synonymously as the term programmed cell death (PCD), though some may argue that PCD refers to the temporal and spatial cell death during development and it occurs through apoptosis (Lawen 2003). Here we use them interchangeably.

Apoptosis plays a critical role in controlling cell populations during embryonic development in multi-cellular organisms, which is probably best illustrated by the tissue differentiation of the nematode *Caenorhabditis elegans*. The worm hermaphrodites have 1090 somatic cells, 131 of which commit suicide by apoptosis; the remaining 959 cells survive and develop into tissues (Danial and Korsmeyer 2004; Ellis and Horvitz 1986). In adults, apoptosis also operates to maintain normal tissue homeostasis, and serves as a defense mechanism against cells that might threaten the integrity of the organism itself, such as cells infected by viruses, cells with damaged DNA or endoplasmic reticulum (ER) stress, as well as autoimmune cells in the immune system (Ashe and Berry 2003; Danial and Korsmeyer 2004; Kaufman 1999).

Under normal circumstances, apoptosis is tightly controlled to ensure destroying of only unwanted cells. However, aberrant regulation of apoptosis has been implicated in the pathogenesis of a wide range of human diseases. Insufficient apoptosis can develop

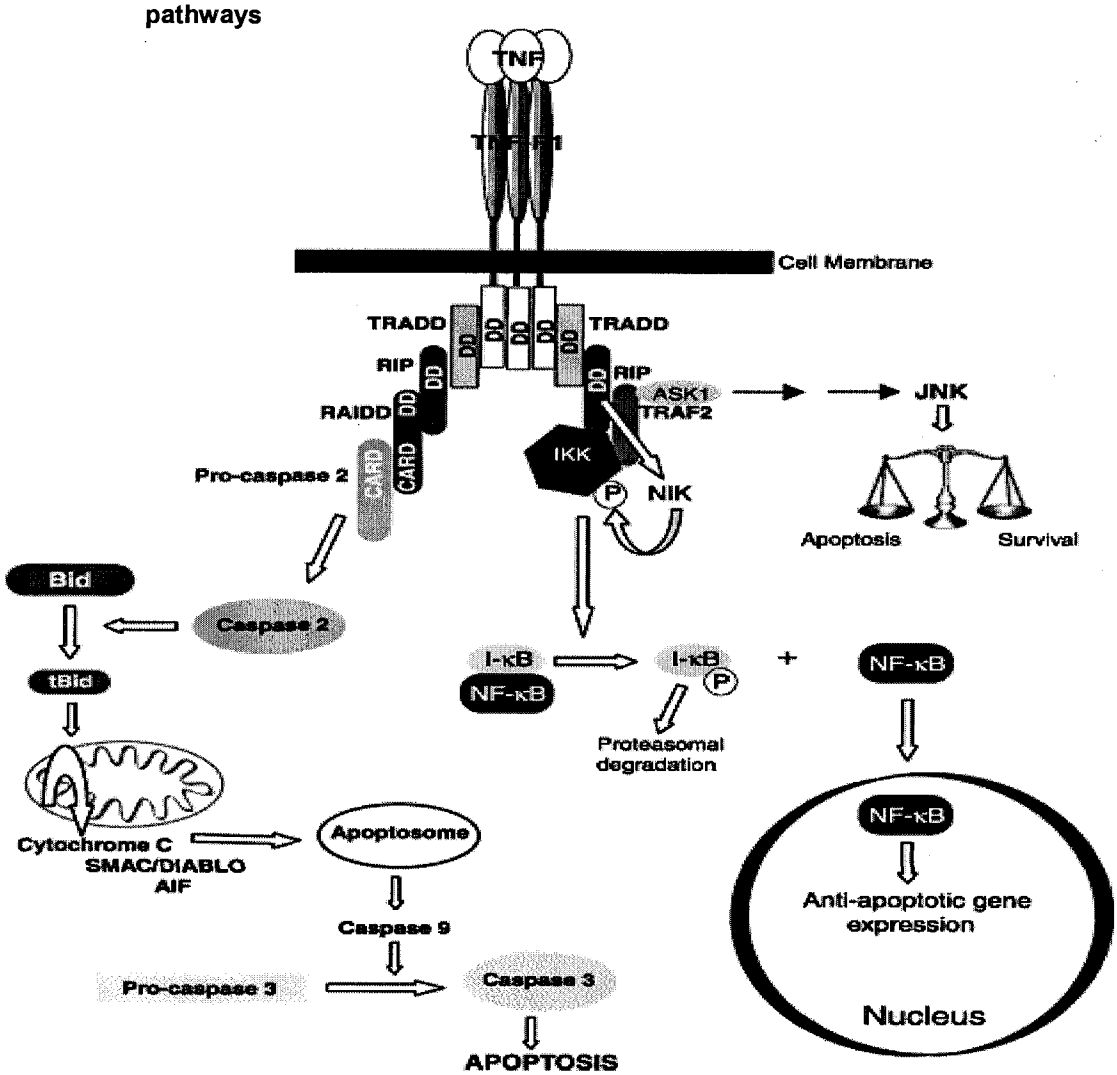
into cancers or autoimmunity, whereas excessive cell death is evident in acute and chronic degenerative disorders (e.g. Alzheimer's and Parkinson's diseases), immunodeficiency, and infertility (Danial and Korsmeyer 2004).

1.2 Apoptotic signaling pathways

How does an organism make the tough decision between cell death and survival? In other words, how are intracellular or extracellular apoptotic stimuli transmitted to invoke cellular responses for killing cells? Much research has been done in the past decades to identify these signal transduction cascades, and many apoptotic signaling pathways have been characterized (for reviews, see Ashe and Berry 2003; Danial and Korsmeyer 2004). Generally, two types of apoptotic signaling pathways were described: intrinsic pathway and extrinsic pathway. The intrinsic pathway is also known as the mitochondrial pathway since the mitochondrion plays a central role in it (Hockenbery *et al.* 1990). In the mitochondrial pathway, three apoptotic factors such as AIF, cytochrome C and Smac/DIABLO can be released from inside mitochondria to initiate the apoptosis program upon intracellular apoptotic stimuli, typically from intracellular stress. Two members of the Bcl-2 protein family, Bcl-2 or Bax, can block or promote the release of cytochrome C, respectively. In contrast, the extrinsic apoptotic pathways are mediated by death receptors located in the cell membranes that are activated by their extracellular ligands, and apoptotic stimuli come from extracellular sources such as UV radiation. Two major well-characterized extrinsic pathways are Fas (TNFRSF6, APO-1/CD95) - mediated death pathway and TNFR1 - mediated death pathway (Ashe and Berry 2003; Danial and Korsmeyer 2004). Fig.1 shows a schematic diagram of the TNFR1-mediated apoptotic signaling pathway. In this pathway, the ligand TNF binds to its receptor TNFR1 localized in the cell membrane. This process initiates the recruitment of proteins with DD domain and/or CARD domain, which are the major protein interaction domains in the DD

superfamily (Reed *et al.* 2003). These domain interactions lead to the sequential signaling cascade to activate Caspase-2, which then passes the signals to the intrinsic mitochondrial pathway. The final phase of these pathways is typically the activation of executioner caspases (i.e. effector caspases) of apoptosis, such as caspase-3, which degrades the cellular infrastructure by large-scale proteolysis. Additionally, some caspases can cleave other caspases and thus causing amplification of apoptotic signals during signaling cascades (Ashe and Berry 2003; Shi 2002).

Figure 1. The TNFR1-mediated extrinsic apoptotic signaling pathway and related pathways



This figure was used from Ashe and Berry (2003) with permission, mainly to illustrate the domain interaction leading to sequential transduction of extracellular apoptotic signals.

Caspases are a family of cysteine aspartate-specific proteases that are involved in apoptosis initiation and execution, or are required for proteolytic processing of certain pro-inflammatory cytokines (Reed *et al.* 2003; Shi 2002). To date, 13 mammalian caspases have been identified, though not all human or mouse homologs for each family member have been identified. Caspases are synthesized as pro-caspases, which are then proteolytically processed to their active forms at the conserved aspartate residues. All pro-caspases contain a highly homologous protease domain and an N-terminal prodomain. The protease domain contains two subunits of ~ 20 and 10 kDa, respectively, which associate to form a heterodimer following proteolytic processing. Two heterodimers then associate to form a tetramer, which is the active form of caspases (Fesik 2000). The N-terminal domain is of variable length depending on the functional category of the caspase. Initiator and inflammatory caspases have long prodomains (>100 amino acids), whereas effector caspases have short prodomains (<30 amino acids). Long prodomains contain specific motifs essential for caspase activity. These motifs may be either death effector domains (DEDs) as in caspase 8 and 10, or caspase recruitment domains (CARDs) as in caspases 1, 2, 4, 5, 9, 11, 12, 13, and 14 (Reed *et al.* 2003).

In addition, several other important apoptotic signaling pathways have been described, including NF-κB pathway and its major regulators of NF-κB, its inhibitor I-κB and I-κB kinase (IKK), as well as JNK/MAPK pathways (Shapira *et al.* 2004; Yamamoto and Gaynor 2004). These two pathways can interact with the TNFR1-mediated extrinsic pathway, as shown in Fig.1.

Despite the progress made towards revealing the various apoptotic-signaling pathways that can ultimately determine a cell's fate, the regulatory mechanisms of gene expression caused by apoptotic signals remain largely unknown. As is shown in the

Methods section below, there are only approximately 60 known binding sites for 26 distinct transcription factors in 18 apoptosis genes, most of which have been collected by the TRANSFAC database (Wingender *et al.* 2000; Wingender *et al.* 2001; Wingender 2004). To fully understand the regulation and deregulation (as in case of human diseases) of apoptotic signaling cascades, it is critical to identify and characterize transcription factors (TFs) and their *cis*-regulatory elements that play crucial roles in apoptosis.

1.3 Computational identification of transcription factor binding sites

In the pre-genomics days, experimental methods for regulatory element discovery such as nuclease protection assays and gel-shift analysis were used to confirm elements on a one-gene-one-element-a-time basis (Liu *et al.* 2004). In this post-genomics era, high-throughput computational approaches are increasingly needed to predict putative transcription factor binding sites (TFBS) for subsequent experimental validation. Developing computational methods for TFBS detection has now become an area of intense research in bioinformatics, and many algorithms have emerged in the past several years. Several TFBS prediction tools are briefly reviewed below.

The discovery of regulatory regions in intergenic sequences through cross-species comparison is often termed 'phylogenetic footprinting'. It is based on the observation that functionally important regions tend to be conserved over the course of evolution by selective pressure. Many putative TFBS are enriched in conserved non-coding genomic sequences (Fickett and Wasserman 2001; Levy *et al.* 2001; Wasserman *et al.* 2000). One strategy tries to find common motifs that are shared by multiple orthologous sequences; while the other begins with global alignment of orthologous sequences, followed by identification of conserved regions. Footprinter is a program designed specifically for phylogenetic footprinting (Blanchette and Tompa 2002;

Blanchette and Tompa 2003), and it detects highly conserved motifs in the homologous regions with regard to the phylogenetic relationship among the homologous sequences. In practice, the choice of species is critical in phylogenetic footprinting. Too great an evolutionary distance can result in regulatory alterations or difficulty in aligning short patches of identity between long sequences. Inadequate evolutionary distance may be insufficient for non-functional sequences to diverge while conserving the functional sequences (Lenhard *et al.* 2003). Thus, the arbitrary parameters are often difficult to choose and they heavily influence the prediction performance of such tools.

Although comparative genomics approaches have been shown to be very effective to significantly reduce the noise and search space in identifying putative *cis*-regulatory elements (Lenhard *et al.* 2003; Liu *et al.* 2004), these techniques usually only provide information about which region is conserved among two or more species. The challenge remains to assess whether these regions of homology are involved in regulation (Ureta-Vidal *et al.* 2003). This is why global alignment has often been used in conjunction with the binding profiles of known transcription factor binding sites, usually taken from the TRANSFAC database (Wingender *et al.* 2000; Wingender *et al.* 2001; Wingender 2004). Consite (Lenhard *et al.* 2003), and rVISTA (Loots *et al.* 2002) are two examples that make use of the integrated approaches. The DNA binding specificity of TFs is commonly modeled using position weight matrices (PWM) (Fickett and Wassermann 2000; Lenhard *et al.* 2003). From a set of binding site sequences determined experimentally for a specific transcription factor, a position frequency matrix (PFM) can be generated by simply counting the frequency of each nucleotide A, C, G, T on each position (Lenhard and Wassermann 2002; Lenhard *et al.* 2003). From this PFM, a standard computational procedure is applied to calculate a PWM, which consists of the log-odds score (or "weight") of each nucleotide on each position in relation to a background model (Fickett and Wassermann 2000). Using these PWM models and

pattern-finding algorithms, one can search the upstream promoter regions and yield a large list of putative TFBS. An advantage of this approach is that a TF is immediately associated with the predicted binding sites that the TF can possibly bind. However, the predictive power of this technique is often restricted by the quality of PWM models, which depends upon the number of experimentally determined binding sites as well as their sequence degeneracy (Lenhard *et al.* 2003).

To reduce the non-functional predicted sites, researchers have been trying to take advantage of available experimental data, such as gene expression data, as enhancing signals (Lenhard *et al.* 2003; Ureta-Vidal *et al.* 2003). Functional genomics data, primarily microarray expression data, have been used to improve the predictive performance. For genes clustered from the expression profiles, motif-finding algorithms are used to find over-represented motifs in their upstream regions, assuming that co-expressed genes are more likely to share a similar set of transcription factor binding sites. AlignACE (Roth *et al.* 1998) and DIALIGN (Morgenstern *et al.* 1998) are two programs of such techniques. There are also studies suggesting that genes encoding interacting proteins tend to be co-regulated (Hannenhalli and Levy 2003; Jansen *et al.* 2002; Manke *et al.* 2003; Simonis *et al.* 2004), and in this study we used this hypothesis as one of filtering procedures for the predicted binding sites in human apoptotic genes.

1.4 Aim of the thesis

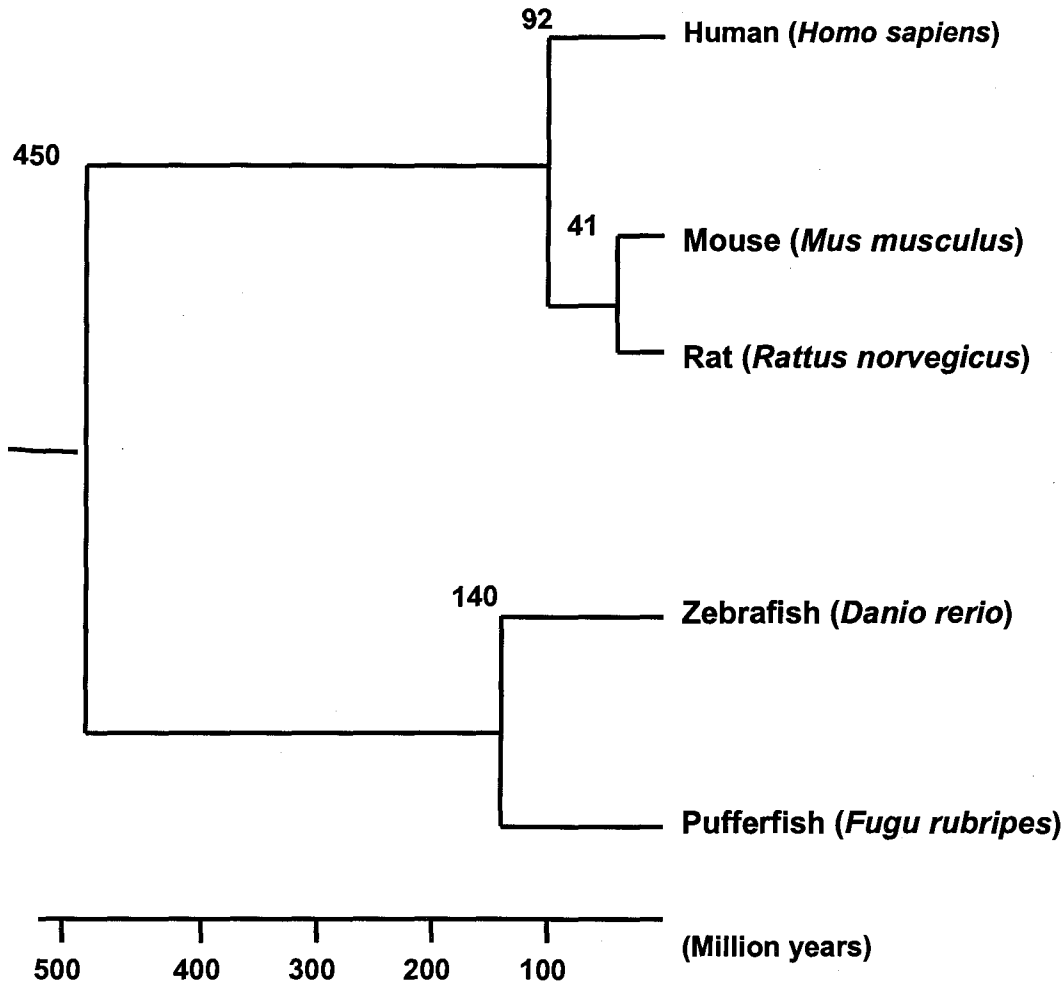
Apoptosis has become a major biomedical research area in recent years, which is explained by the fact that apoptosis is implicated in the pathogenesis of a wide variety of human diseases. Characterization of protein-protein interaction networks and regulatory networks during apoptotic signal transduction will help us to understand the mechanisms of the human diseases associated with apoptosis and develop rational strategy for their prevention and treatment. In this area, high-throughput computational

approaches may guide the design of downstream experiments to speed up discovery of apoptotic protein-protein interactions and regulatory elements. Regulatory networks that connecting transcription factors and their target genes in apoptosis will help us understand the transcriptional regulation during apoptotic signal transduction.

Proteins involved in apoptosis often contain evolutionarily conserved domains that can serve as signatures for identification, allowing one to apply bioinformatics techniques in the analysis of families of apoptosis-related proteins (Reed *et al.* 2003). The authors have classified apoptotic proteins into 16 protein families and signature domains and identified over 200 apoptosis genes in the genome of human or mouse. The major protein families known to be involved in apoptosis include caspases, Bcl-2 family, death domain superfamily, as well as tumor necrosis factor (TNF) superfamily and their receptors, and others.

In this study, we have examined the conservation of major apoptotic signaling pathways in human (*Homo sapiens*), and vertebrate model organisms mouse (*Mus musculus*), rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), and pufferfish (*Fugu rubripes*). The phylogenetic relationship of these 5 vertebrate species is represented in Fig. 2. Human and mouse (rat) diverged ~90 million years ago (MYA), human and pufferfish diverged ~450 MYA, and their complete genome sequences have been released (Ureta-Vidal *et al.* 2003). We have also extracted all experimentally determined protein-protein interactions of human apoptosis genes and constructed a protein interaction network, and investigate how protein-protein interaction data might be used in conjunction with sequence conservation for computational TFBS detection. Lastly, primarily based on our computational analyses we have constructed several *in silico* regulatory networks of several major apoptotic-signaling pathways to obtain insights into the transcriptional regulators that are critical in controlling apoptosis.

Figure 2. The phylogenetic relationship of the 5 vertebrate species



This figure was drawn based on Ureta-Vidal *et al.* (2003). The numbers at the corner of each branch represent the time in million years in which the species diverged.

CHAPTER TWO: METHODS

2.1 Identification of apoptosis genes in mammalian genomes

The protein accession numbers of apoptosis associated genes in human and mouse were primarily derived from the 227 genes compiled by Reed *et al.* (2003), and updated from NCBI LocusLink (Pruitt and Maglott 2001) for genes that were annotated or updated after their publication. The apoptosis genes in rat were identified at NCBI LocusLink using the human gene names or synonyms. Genes without available RefSeq (Pruitt and Maglott 2001) transcripts were excluded. The RefSeq accessions of all apoptosis genes in each mammalian species were batch retrieved from NCBI. For genes mapping to multiple alternatively spliced isoforms, only the longest transcript is used for obtaining upstream sequences below.

2.2 Identification of apoptosis genes in zebrafish and pufferfish genomes

The proteomes of both zebrafish and pufferfish have not been well annotated and no protein name could readily be found like in mammals. Thus, to identify the apoptosis relevant genes in these two lower vertebrate genomes, their whole gene sets were obtained from Ensembl via the EnsMart interface (Kasprzyk *et al.* 2004). The zebrafish gene set is derived from the whole genome shotgun assembly sequence version 3 (released on November 2003), whereas the pufferfish genome is based on v21.2c.1 (released on May 2004). Each gene set was formatted to be suitable for BLAST search. Apoptosis genes in these two genomes were identified by TBLASTN with default settings (Altschul *et al.* 1997). The protein sequence of each human gene was used to blast against the zebrafish or pufferfish gene set. If no human gene is available or no

significant hit was identified, the protein sequence of the mouse homolog was used. If again no mouse protein sequence is available or no significant hit was identified, the rat protein sequence was used instead. To further ensure data quality, all putative gene candidates were verified at the Ensembl web site (Stalker *et al.* 2004). If a hit is unambiguously annotated by Ensembl to be a homologue to a mammalian apoptosis gene (or occasionally to an apoptosis gene in other vertebrates) or contains a putative InterPro domain involved in apoptosis, this gene was annotated as an apoptosis gene in zebrafish or pufferfish genome.

2.3 Retrieval of upstream sequences

The region to be used for detecting *cis*-regulatory elements in eukaryotes is not well defined. In theory, the whole regulatory region for metazoan genes should include the 5'- and 3'-flanking regions, as well as the intronic sequences, and this is a large amount of sequence (Ureta-Vidal *et al.* 2003). Taking into account our computational power, we elected to choose only 3 kb upstream sequences for this analysis. The whole process could be easily scaled up for analyzing more sequences if required. The upstream region was measured from transcription start site (TSS) based on the RefSeq annotation. The transcript RefSeq accessions in mammals were used to obtain 3 kb upstream sequences from the UCSC Table Browser (Karolchik *et al.* 2004). The reference sequences in human, mouse, and rat are based on the latest available assemblies of July 2003, May 2004, and June 2003, respectively. For the zebrafish and pufferfish genes, their Ensembl transcript identifiers were used to obtain the 3 kb upstream sequences from Ensembl via the EnsMart interface (Kasprzyk *et al.* 2004).

2.4 Known transcription factors and their binding sites

The TRANSFAC database professional version 7.4 was licensed from the

BioBase (Wingender *et al.* 2000; Wingender *et al.* 2001; Wingender 2004). We only extracted the entries for vertebrate species, which include 4,754 binding sites, 786 transcription factors, and 490 PWM matrices representing binding profiles of known transcription factors.

2.5 The known apoptotic protein-protein interactions in humans

The experimentally determined protein-protein interactions of human apoptosis genes were extracted from the human protein reference database - HPRD (Peri *et al.* 2003), and only the interactions between the genes in our human apoptosis gene set were retained. The latest update of the protein-protein interaction data for our analysis was performed on June 20, 2004. The protein-protein interaction network was constructed using the open-source, Java-based Cytoscape (Shannon *et al.* 2003).

2.6 Data storage

To facilitate data storage and analysis, we designed a MySQL relational database for storing all genomic DNA and protein sequences, known and predicted transcription factors and their binding sites, as well as matrices from TRANSFAC (Wingender *et al.* 2000; Wingender *et al.* 2001; Wingender 2004). This local database is also the backend engine for the searchable web site described in 2.12.

2.7 Identification of conserved regions

A conserved region is determined by sequence identity percentage and length cutoffs. Conserved segments with percent identity X and length Y are defined to be regions in which every contiguous subsegment of length Y is at least X% identical to its paired sequence. The global alignment tool LAGAN (Brudno *et al.* 2003) was used for sequence alignment of the upstream or transcript sequences of each gene pair. The sequence identity percentage was calculated by using the BioPerl toolkit (Stajich *et al.*

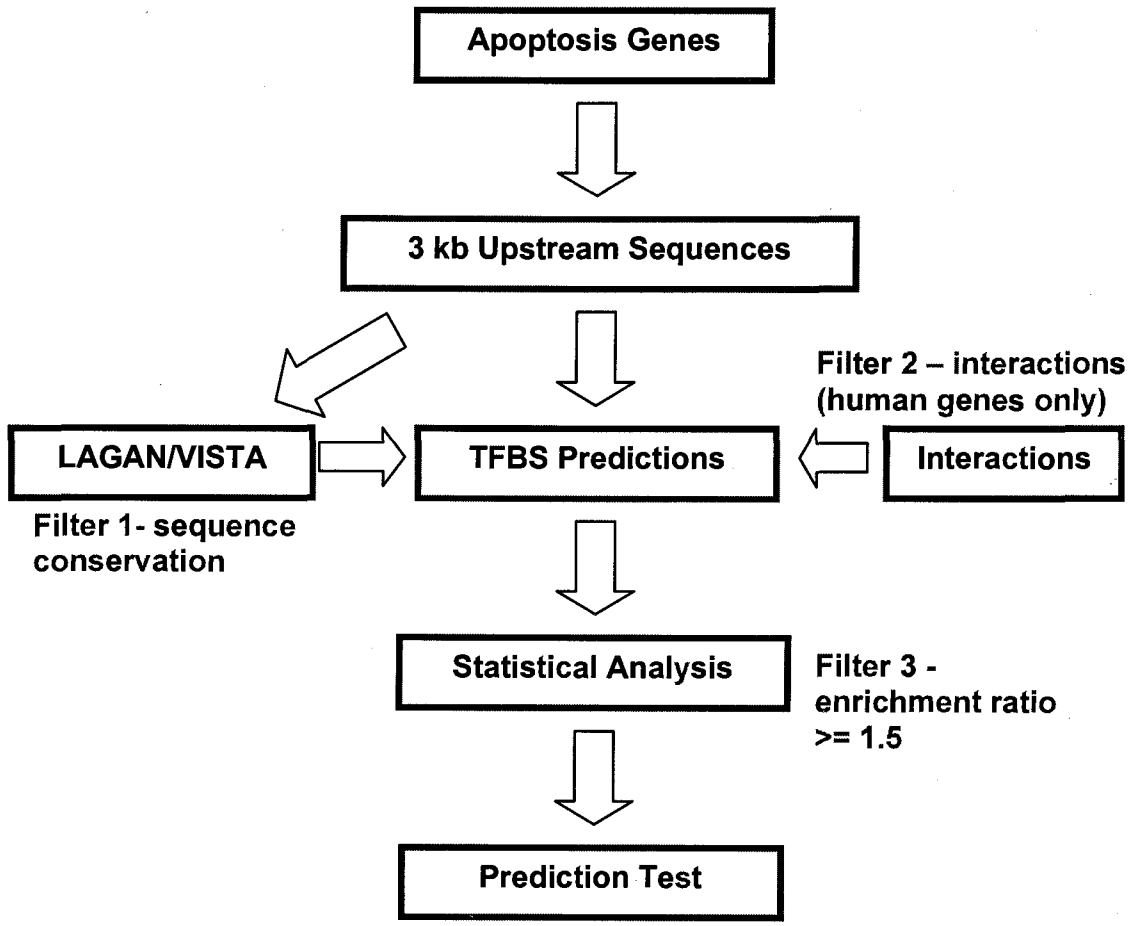
2002). The alignment visualization tool VISTA (Mayor *et al.* 2000) was used to identify conserved regions in the aligned upstream sequence pair. The VISTA window size is 21-bp, and the sequence identity threshold in this window is specified dynamically based upon the average identity of the sequence pair (i.e. 5% higher than the average identity). These segments are then merged to define the conserved regions between two upstream sequences of an apoptosis gene in two species (e.g. human and mouse).

2.8 TFBS identifications

Putative transcription factor binding sites are predicted and filtered by a three-step approach, which is schematically illustrated by the flow chart in Fig. 3. First, the TFBS software system (Lenhard and Wasserman 2002) was used for predicting transcription regulatory sites in the upstream regions of our genes in each species. The TF binding profiles were the 490 matrices derived from vertebrate genomes in the TRANSFAC database (Wingender *et al.* 2000; Wingender *et al.* 2001; Wingender 2004). Matrix thresholds of 70%, 75%, 80%, 85%, and 90% were compared. Hits that match each matrix above the predefined threshold were identified along the genomic sequences and stored in the database. Second, the algorithm identifies a subset of binding sites in the conserved regions between each species pair as determined above in 2.6. Third, for human genes with interacting partners, the algorithm identifies another subset of common binding sites matching the same PWM matrix in at least one pair of interacting genes. Union and intersection algorithm was used to integrate phylogenetic footprinting information and human protein-protein interaction data. The union subset of transcription factor binding sites consists of those sites that are either in the conserved regions between two species or that are shared by at least a pair of interacting genes, but the intersection subset was only considered once. The intersection subset of transcription factor binding sites consists of those sites that are both in the conserved

regions between two species and that are shared by at least a pair of interacting genes.

Figure 3. The flow chart of the TFBS prediction approach.



This figures illustrates schematically the logical flow of predicting and filtering transcription factor binding sites. All apoptosis genes were used to retrieve their 3 kb upstream sequences. These upstream sequences were used for detecting putative transcription factor binding sites and for sequence alignment (LAGAN) to identify conserved regions between species pairs (VISTA; Filter1). Another two filters are protein-protein interactions for human genes only and enrichment ratio (see Statistical analysis below). Prediction test (see Prediction performance testing below) was performed for the final sets of transcriptional factor binding sites.

2.9 Statistical analysis

The statistical significance of predictions was estimated by the over-representation of k-mer (a binding site of k nucleotides) in the upstream (non-coding) regions using exon sequences (transcript) as background model, an approach similar to

Hampson *et al.* (2002) and Xue *et al.* (2004). An enrichment ratio of a binding site in upstream against exon sequence was measured by S_{nc}, the ratio of C_{nc}, its occurrence in the upstream, with the C_{ex}, its occurrence in the corresponding exon sequence:

$$S_{nc} = C_{nc} / C_{ex} \quad (1)$$

$S_{nc} > 1$ correlates with over-representation of the binding site, and the larger the S_{nc} is, the more significantly the site is enriched in the upstream.

The nucleotide composition can be rather different between the upstream and coding regions, and thus the probability of obtaining a specific binding site is also different in the upstream or transcript sequence. In order to normalize this disparity, we calculated the average frequency of each nucleotide A, C, G, T in upstream and transcript. Given the P_a, P_c, P_g, and P_t for the average frequency of the 4 nucleotides A, C, G, T, respectively, the expected occurrence F of a k-mer with a hypothetical binding site sequence of $A_iG_jC_mT_n$ (where i + j+ m+ n = k) can be estimated as:

$$F = P_a^i P_g^j P_c^m P_t^n \quad (2)$$

Therefore, the normalized enrichment ratio S_{nc}' (normalized by the length of non-coding region L_{nc} and background exon L_{ex}, as well as the expected frequency F_{nc} and F_{ex}) for a binding site is equivalent to the measure of enrichment using the frequency in transcript region as background, which includes the theoretical ratio of k-mer occurrence in the non-coding region against the transcript region [Equation (3)].

$$S_{nc}' = (C_{nc}L_{ex}F_{ex}) / (C_{ex}L_{nc}F_{nc}) \quad (3)$$

We used this algorithm to normalize the nucleotide composition difference and sequence length of both upstream and coding regions, and calculate the enrichment of predicted sites for a transcription factor in the upstream region of a gene against its transcript sequences. For approximately estimating F, the TFBS degeneracy was not considered and the site sequence (e.g. $A_iG_jC_mT_n$ above) used is the binding site that has the highest score against the matrix of the transcription factor in the upstream or

transcript sequences. Binding sites that are at least 1.5 times enriched in the upstream regions compared to its coding regions were considered statistically significant.

2.10 Calculating the expected number of occurrences for each TF class

The method for calculating the expected number of binding site occurrences for each TF class is similar to that described by Zhang *et al.* (2002). Given a matrix representing the binding profile of a TF and its dominant binding site sequence, we can calculate the probability (p) of its occurrence in the upstream sequence as:

$$p = \prod_{i=1}^{N} \left(P_b \right)_{b \in S} \qquad (1)$$

where N is the length of the matrix, i is the position index, S represents a set of nucleotides (A, C, G, T), and P_b is the frequency of each nucleotide in the upstream sequence, i.e. P_a, P_c, P_g, or P_t. Assuming a uniform distribution of nucleotides in upstream sequences, P_a, P_c, P_g, or P_t is equal to 0.25 each. Then its expected occurrence (λ) in the upstream is calculated as $\lambda \cong Lp$, where L is the length of the upstream sequence, i.e. 3 kb. Since we used both strands for putative TFBS predictions, the total length of each upstream sequence is 6 kb. The average expected occurrence for each TF class is calculated based on all expected occurrence of all TFs in each class.

2.11 Predictive performance testing

A test data set with known transcription factor binding sites in the promoter regions of apoptosis genes was assembled from both the TRANSFAC database and literature. There are totally 61 known binding sites. After excluding genes without interacting partners in our human gene set (i.e. CIITA/C2TA, PUMA/BBC3, and TRIF), the test set consists of 54 binding sites in 15 distinct apoptosis genes for 26 distinct

transcription factors (Table 1). The binding sites of human Caspase-8 are not in the TRANSFAC database and were derived from literature (Liedtke *et al.* 2003). To assess the predictive performance, the sensitivity (SN) and specificity (SP) are defined based on Lenhard *et al.* (2003). SN is the percent correct predictions in the test set, that is, when a prediction and a known TFBS overlap by at least 50% given a corresponding transcription-factor binding profile. SP is defined as the number of predicted binding sites in the 3 kb upstream sequence but expressed as the average number of predicted binding sites along a 100 bp upstream sequence in both strands using 490 binding matrices from vertebrates. Control for using interaction data as filtering procedure was to test binding sites shared by gene pairs that have no known interacting relationships. To achieve this, gene pairs in the interaction tables are shuffled to ensure that the randomly generated genes pairs do not have known interaction data available from HPRD (Peri *et al.* 2003). For example, if gene A has interactions with gene B and gene C (A-B, A-C are now gene pairs in the interaction table), in the new pairs gene A was paired with a random human apoptosis gene except gene B and gene C. The predicted binding sites shared in the new gene pairs were used for testing similarly as the original gene pairs that have known interacting relationships.

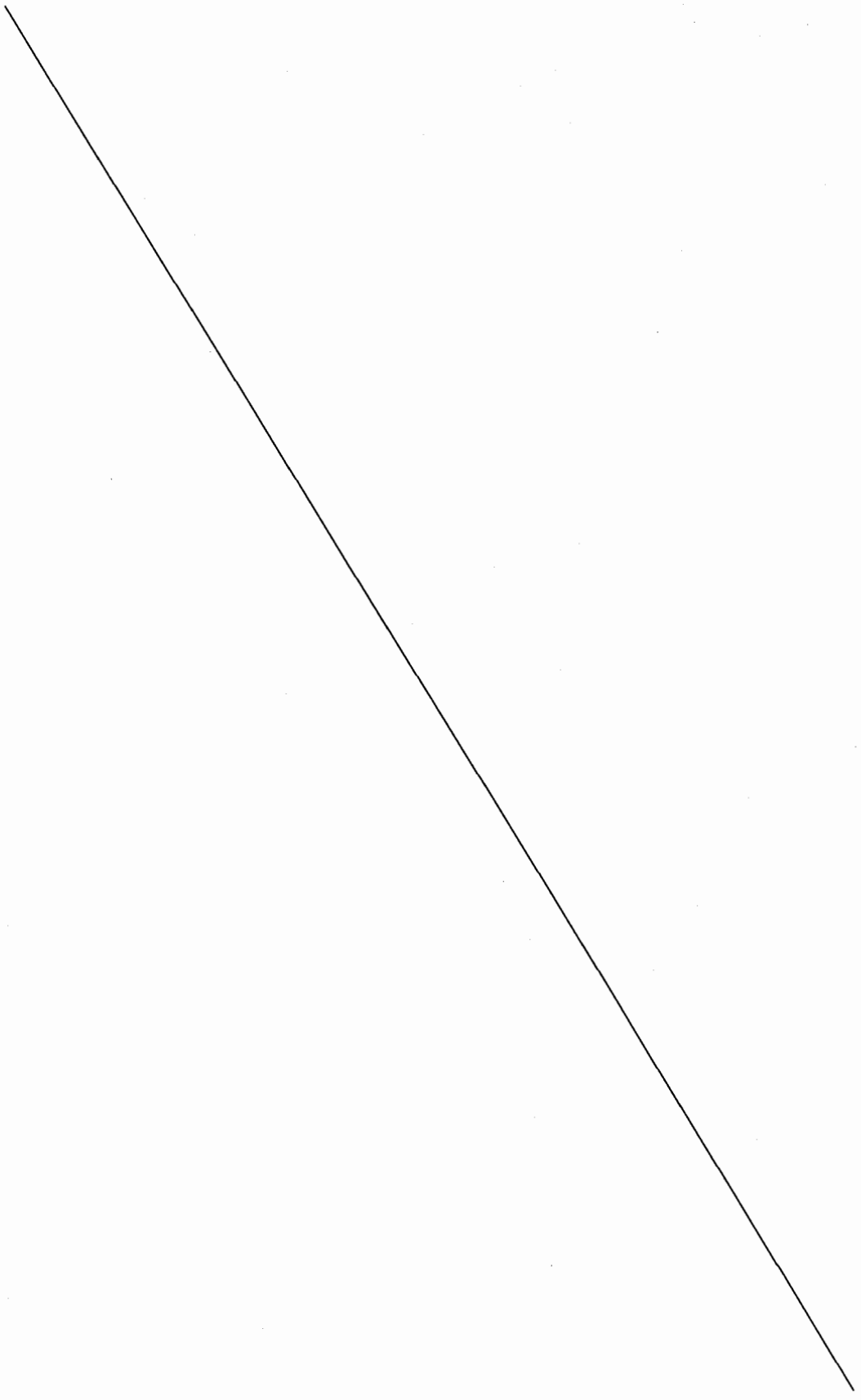
Table 1. The test data set consisting of 15 genes and 54 experimentally determined transcription-factor binding sites for 26 distinct transcription factors.

Target gene	Binding site sequence [1]	Site location [2]	Factor Name
A1a	ctcagatcccagccaGTGGACTTAGCCCctgtttgc	-134/-98	HNF-4alpha1
A1a	CAGCCAGTGGACTTAGCCCCTGTTTG	-128/-103	LF-A1
A1a	ATCCCAGCCAGTGGACTTAGCCCCTGTTTG	-125/-96	LF-A1
A1a	CTCCGATAACTG	-97/-86	LF-A2
A1a	TGGTTAATATTCACCAgc	-86/-69	HNF-1
A1a	GGGTGACCTTGGTTAATATT	-85/-66	LF-B2
A1a	ACCTTGGTTAATATTCACCAGCAG	-80/-57	HNF-1
A1a	GGGTGACCTTGGTTAATATTCACCAGCAG	-80/-52	HNF-1
A1a	GTTAATATTCACC	-78/-66	HNF-1
A1a	TGCCCCTCTGGATCCACTGCTTAA	-46/-23	LF-C
Apaf1	ctcAGACATGTCTggagaccctaggaCGACAAGCCCagg	-607/-569	p53
Bax	tcacaagttagAGACAAGCCTGGGCGTGGGCtatattg	-113/-83	p53
Bax	GGGCGTGGGC	-92/-83	Sp1
Bcl-x	GTTTCCCCCTCCCTGCGTCCCTCACTGAAACCTTGAACCCCATTGAGAAG	-906/-857	BCL-6
Bcl-x	CCAGGGAGTGACTTTCCGAGGAAGGCATTTCGGAGAAGACGGGGGTAGAA	-357/-308	BCL-6
Bcl-x	AGTCCACTGGTGCTTTCGATTTGACTTAAGTGAAGTATCTTGGAACCTAG	-287/-238	BCL-6
Caspase-1	ataaAGACATGCATATGCATGCAca	-85/-66	p53
Caspase-8	GGGCGTTCCC	-75/-66	NF-KB
Caspase-8	AGTGGGCGGGAGG	-97/-85	Sp1
Caspase-8	TTCCAAGAA	-198/-190	STAT1
DR4	-	-350/-344	AP1
IKBA	AATCGATCGTGGGAAACCCCAGGGAAAG	-63/-36	RBP-Jkappa
NFkB1	GAAACGTCATGGGAATTCCCCCCTCCGGG	-119/-90	RBP-Jkappa
TNFRSF5	gagggaatTTCCTTTGAAagagagcg	-529/-504	STAT6
TNFRSF5	gggaatgTTCTGGGGAAactcctgc	-136/-111	STAT6
TNFRSF6	aacccGGGCGTTCCCcagcg	-300/-281	NF-kappaB
TNFRSF6	aacccGGGCGTTCCCcagcg	-300/-281	Sp1
TNFSF2	tgtccaGGGCTAtggaAGTCGAgtatcg	-894/-866	LXR-alpha:RXR-alpha

Target gene	Binding site sequence [1]	Site location [2]	Factor Name
TNFSF2	CTCCC	-515/-511	LITAF
TNFSF2	ccaacTTTCCaaa	-187/-175	NF-AT1
TNFSF2	atCCCCGCCCCCgcg	-175/-161	Sp1
TNFSF2	GCCCCCGC	-169/-162	EGR1
TNFSF2	gagtGAGAAGAAAccgag	-161/-144	NF-AT1
TNFSF2	cactaccgcttcctccagaTGAGCTCAtgggtttctccaccaag	-125/-82	c-Jun
TNFSF2	tccagaTGAGCTCAtgggttt	-122/-102	c-Jun
TNFSF2	accgCTTCCTCCaga	-121/-107	NF-AT1
TNFSF2	ccgcTTCCtccag	-120/-108	c-Ets-1
TNFSF2	agctcatgggTTTCTCCA	-104/-87	NF-AT1
TNFSF2	ggcatGGGAATTTCCaactc	-102/-83	c-Rel
TNFSF2	aaccaaGGAAgtt	-87/-75	c-Ets-1
TNFSF2	aaggaagTTTTCCgctgg	-84/-67	NF-AT1
TNFSF2	gttTTCCgct	-78/-69	c-Ets-1
TNFSF2	gattctttCCCCGCCCtcctctcgccccagggaca	-61/-27	Sp1
TNFSF2	gattcTTTCCccgcct	-61/-46	NF-AT1
TNFSF1	TGGGGGCTTCCCC	-100/-88	NF-kappaB
TNFSF5	agcTAAATTTTATTTAATAATTATgcc	-564/-538	AKNA
TNFSF5	gataGGAAAatact	-269/-256	Unknown
TNFSF5	tcctGGAAAatgt	-71/-59	Unknown
TNFSF6	TAAATAAATAAGTAAATAAATA	-724/-703	FOXO3a
TNFSF6	gatctaatttctaaaGTGGGTGTagcaggtttttaac	-700/-664	EGR1
TNFSF6	attGTGGGCGGaaacttccagggg	-184/-161	EGR1
TNFSF6	attgtgggcGGAACTTccagggg	-184/-162	NF-AT2
TNFSF6	ctatGGAAActct	-144/-132	NFAT
TNFSF6	tcagctgcaaagtgaGTGGGTGTttctttgag	-129/-98	EGR1

(1) The site sequence in upper case is the required binding pattern; site sequence in lower case can be degenerate.

(2) The site location is relative to the transcription start site, which is set to be 0.



2.12 Construction of apoptotic regulatory networks

The putative transcription factors that have binding sites within 600 bp upstream of transcription start sites from major TF families for all gene components of the apoptotic signaling pathways are retrieved from the local MySQL database. The regulatory networks were constructed with Cytoscape (Shannon *et al.* 2003).

2.13 Development of the data-driven web site

The MySQL database described in 2.6 serves as the backend engine. The searchable web interface was developed with server-side scripting language PHP. User documentation is available at http://apoptosis.mbb.sfu.ca/help.php.

CHAPTER THREE: RESULTS

3.1 Apoptotic signaling pathways in vertebrate genomes

Based on the protein families classified by Reed *et al.* (2003), we identified 236 apoptosis genes with RefSeq transcripts in humans, and 223 in mice. However, only 147 RefSeq genes were identified in rats (see Appendix A). This discrepancy of gene numbers between rat and human/mouse is likely due to the rat genome status, though some apoptosis genes might have evolved significantly during mammalian evolution. With the recent release of the complete rat genome, more apoptosis genes could be found in rat. For the other two vertebrates, we identified 114 apoptosis related genes in the zebrafish genome and106 apoptosis genes in the pufferfish genome (see Appendix B) using TBLASTN with protein sequences from mammals as queries. These lower gene numbers might be attributed to the relative simplicity of apoptotic signaling pathways in these two lower vertebrate species. The apoptosis signaling pathways may have become more complex and many functionally redundant genes emerged over the course of vertebrate evolution (Le Bras *et al.* 2003).

Most caspases involved in apoptosis initiation and execution were identified in the zebrafish or pufferfish genomes (Table 2), except caspase-10. Caspase-10 was also not found in mouse or rat genome. Neither caspase-11 nor caspase-12 was identified in human; human caspase-4 and caspase-5 are orthologous to murine caspase-11 (Reed *et al.* 2003). In contrast, caspases involved in pro-inflammatory cytokine activation, most of which were identified in mammalian genomes, appeared to be absent in zebrafish and pufferfish genomes except Caspase-1.

Table 2. Functions of caspases and homologous genes identified in the 5 vertebrate genomes.

Caspase	Human	Mouse	Rat	Zebrafish	Pufferfish	Main function [1]
Caspase-1	+	+	+	+	+	Cytokine activation
Caspase-2	+	+	+	+	+	Apoptosis initiator
Caspase-3	+	+	+	+	+	Apoptosis effector
Caspase-4	+	+	-	-	-	Cytokine activation
Caspase-5	+	-	-	-	-	Cytokine activation
Caspase-6	+	+	+	+	+	Apoptosis effector
Caspase-7	+	+	+	+	+	Apoptosis effector
Caspase-8	+	+	+	+	+	Apoptosis initiator
Caspase-9	+	+	+	+	+	Apoptosis initiator
Caspase-10	+	-	-	-	-	Apoptosis initiator
Caspase-11	-[2]	+	+	-	-	Cytokine activation
Caspase-12	-[2]	+	+	-	-	Cytokine activation
Caspase-14	+	+	+	-	-	Cytokine activation

Plus sign (+) indicates that a homolog was identified in the species; minus sign (-) indicates that a homolog was not identified in the species.
(1) Not all these family members have been well characterized with respect to their physiological roles and targets, although it is known that distinct caspases play roles in apoptosis or inflammation (Ashe and Berry 2003).
(2) Human Caspase-4 and Caspase-5 are homologous to murine Caspase-11 (Reed *et al.* 2003), and the human ortholog of murine Caspase-12 is non-functional due to a termination codon prior to the region encoding the catalytic domain (Fischer *et al.* 2002).

For the key genes of the 4 major apoptotic signaling pathways described in Ashe and Berry (2003), all homologous genes for the mitochondrial intrinsic apoptotic pathway were identified in these 5 vertebrate genomes (Table 3). This is not unexpected, as most genes in this pathway are homologues of death genes in *C. elegans*. For example, the vertebrate homolog of Apaf-1 in *C elegans* is *ced-4* (Zou *et al.* 1997), the Bcl-2 homolog is *ced-9* (Hengartner and Horvitz 1994), and caspase-1 is homologous to *ced-3* (Yuan *et al.* 1993; *ced-3* also has sequence similarity to caspase-2 [Wormbase] and caspase-3 [Desnoyers and Hengartner 1997], but the *ced-3* substrate specificity is more similar to caspase-3 [Xue *et al.* 1996]). Hence, the intrinsic apoptotic pathway is well conserved during metazoan evolution.

For the death receptor - mediated extrinsic pathways, many components in the

TNFR-mediated pathway were found in the 5 vertebrate genomes, although the receptor TNFR1 (TNFRSF1A) has not been identified in the fish genomes. Some other members of the TNFR superfamily exist instead, which are presumably able to bind to the TNF ligands (e.g. TNF-α or TNF-β); or TNFR1 will ultimately be identified in zebrafish and pufferfish genomes. The major components of the NF-κB pathway also exist in all 5 genomes. On the other hand, three key genes in the Fas-mediated extrinsic pathway, i.e., Fas (TNFRSF6), FasL (TNFSF6) and adaptor protein FADD, are not identified in both zebrafish and pufferfish genomes. This is in agreement with an earlier report in the pufferfish genome (Le Bras *et al.* 2003). If these results were not caused by the annotation status of zebrafish and pufferfish genomes, they indicate that perhaps only with the exception of Fas-mediated extrinsic pathway, the core apoptotic pathways are evolutionarily conserved in vertebrates.

Table 3. Key genes involved in the four major apoptotic pathways and homologous genes identified in the 5 vertebrate genomes.

Gene	Species					Apoptotic pathway
	Human	Mouse	Rat	Zebrafish	Pufferfish	
AIF	+	+	+	+	-	Intrinsic
Bcl-2	+	+	+	+	+	Intrinsic
Bax	+	+	+	+	+	Intrinsic
Apaf1	+	+	+	+	+	Intrinsic
Caspase-9	+	+	+	+	+	Intrinsic
Smac	+	+	-	+	+	Intrinsic
XIAP	+	+	+	+	+	Intrinsic
Caspase-3	+	+	+	+	+	Intrinsic
TNFSF6	+	+	+	-	-	Fas extrinsic
Fas	+	+	+	-	-	Fas extrinsic
FADD	+	+	+	-	-	Fas extrinsic
Caspase-8	+	+	+	+	+	Fas extrinsic
Bid	+	+	+	-	-	Fas extrinsic
TNFSF1	+	+	+	+	-	TNFR1 extrinsic
TNFRSF1A	+	+	+	-	-	TNFR1 extrinsic
TRADD	+	+	+	-	+	TNFR1 extrinsic
RIP	+	+	+	+	+	TNFR1 extrinsic
TRAF2	+	+	+	+	+	TNFR1 extrinsic
Caspase-2	+	+	+	+	+	TNFR1 extrinsic
NF-κB	+	+	+	+	+	NF-κB
IκB	+	+	+	+	+	NF-κB
IKK	+	+	+	+	+	NF-κB

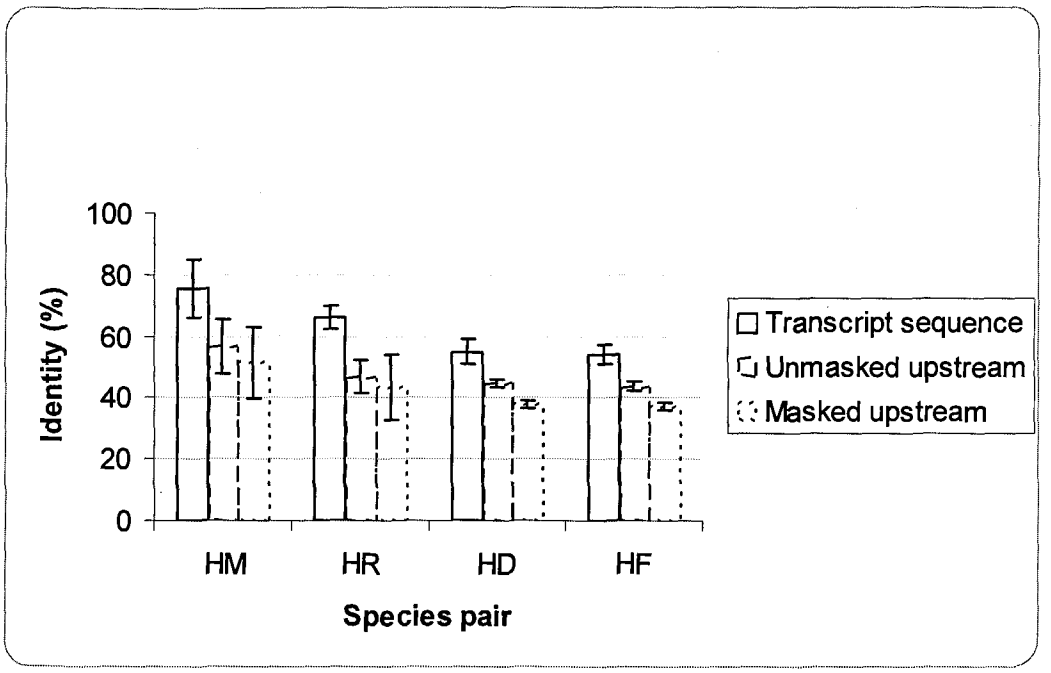
Plus sign (+) indicates that a homolog was identified in the species; minus sign (-) indicates that a homolog was not identified in the species. The key genes involved in each apoptotic pathway were based on Ashe and Berry (2003).

3.2 Sequence similarities of apoptosis genes and upstream regions

The determination of sequence conservation varies depending upon the chosen species, and the evolutionary rates can be considerably different between genes (Ureta-Vidal *et al.* 2003). Thus, it is crucial to estimate the sequence identity in our gene set for determining conserved regions across several species. Based on pairwise sequence

alignment by using the global alignment tool LAGAN, the average sequence identity of human and mouse apoptosis genes (transcripts) was ~75%, and ~57%, ~52% for unmasked and masked upstream sequences, respectively (Fig. 4). The average sequence identity is lower in human-rat, human-zebrafish, and human-pufferfish comparisons, but likewise, the gene sequences had highest sequence identity, followed by unmasked upstream sequences and masked upstream sequences. These results show that the upstream sequences are significantly less conserved than the transcript sequences. If the upstream is masked (as in our case for TFBS predictions), the sequence identity further declines, suggesting that repetitive elements contribute much to the sequence identity of the upstream sequences. Furthermore, the sequence identity was fairly diverse in our gene set (as indicated by the relatively high standard deviation in Fig. 4), making it difficult to set a fixed sequence conservation threshold for identifying conserved regions in phylogenetic footprinting studies. In the present study we chose to set the sequence identity threshold for each upstream sequence pair dynamically based on their average sequence identity. For example, if the masked upstream sequences of Apaf1 are on average 55% identical between human and mouse, we set 60% as the conservation threshold in the 21-bp sliding windows to identify their conserved regions.

Figure 4. Sequence identity of apoptosis transcripts and their upstream regions.



The figure shows the average sequence identity between each species pair for all apoptosis genes we have identified. The error bar represents standard deviation. HM = human and mouse comparison, HR = human and rat comparison, HD = human and Danio comparison, HF = human and Fugu comparison. The transcript sequence is the RefSeq sequence of each gene; Unmasked upstream is the 3 kb upstream sequence in which the repetitive elements are not masked; masked upstream is the 3 kb upstream sequence in which the repetitive elements are masked out by using RepeatMasker (Smit and Green. http://repeatmasker.org/).

3.3 Human protein-protein interaction networks in apoptosis

We extracted 366 distinct, experimentally determined protein-protein interactions covering 168 (~72%) human apoptosis genes, and constructed a human apoptotic protein interaction network (see Appendix C). This protein interaction network was defined as nodes representing human apoptotic proteins (genes) and edges representing all known interactions between them irrespective of their interaction type or condition. Thus, this network shows all currently known interactions of human apoptosis genes (proteins) that can take place under a certain biological context. In this entire network, each node (gene) has an average of ~2.2 edges (interactions). Table 4 lists the

number of direct interactions for highly interacting nodes, genes that have the number of interactions exceeding 3 times the average interaction number (i.e. 3×2.2=6.6). Genes with more than 20 direct interactions include TRAF2 (36), Caspase-8 (24) and Bcl-2 (23), and TRAF1 (22), all of which play critical roles in apoptosis. On the other hand, genes with only a single interacting partner include BAG3, Beclin, Bcl-3, Bcl-B, Bik, Bimp2, Bimp3, CARD9, COP, DR6, EDA-A1, EDAR, HIPPI, IFI16, IRAK-M, Mal, MALT-1, PYRIN, SIAH-1, SIAH-2, TEF2, TLR6, TNFRSF10C, TNFRSF10D, TNFSF7, TNFSF8, TNFSF12, and TNFSF18. TRAFs can be used as examples to infer functional importance form the number of interactions. TRAFs bind TNF receptors and their adapter proteins (e.g. TRADD), protein kinases involved in induction of NFκB and Jun amino-terminal kinase (JNKs), and serve as a bridge between TNFR1, NFκB and JNK pathways (Ashe and Berry 2003; Reed *et al.* 2003).

Table 4. Highly interacting genes in the protein-protein interaction network of human apoptotic pathways.

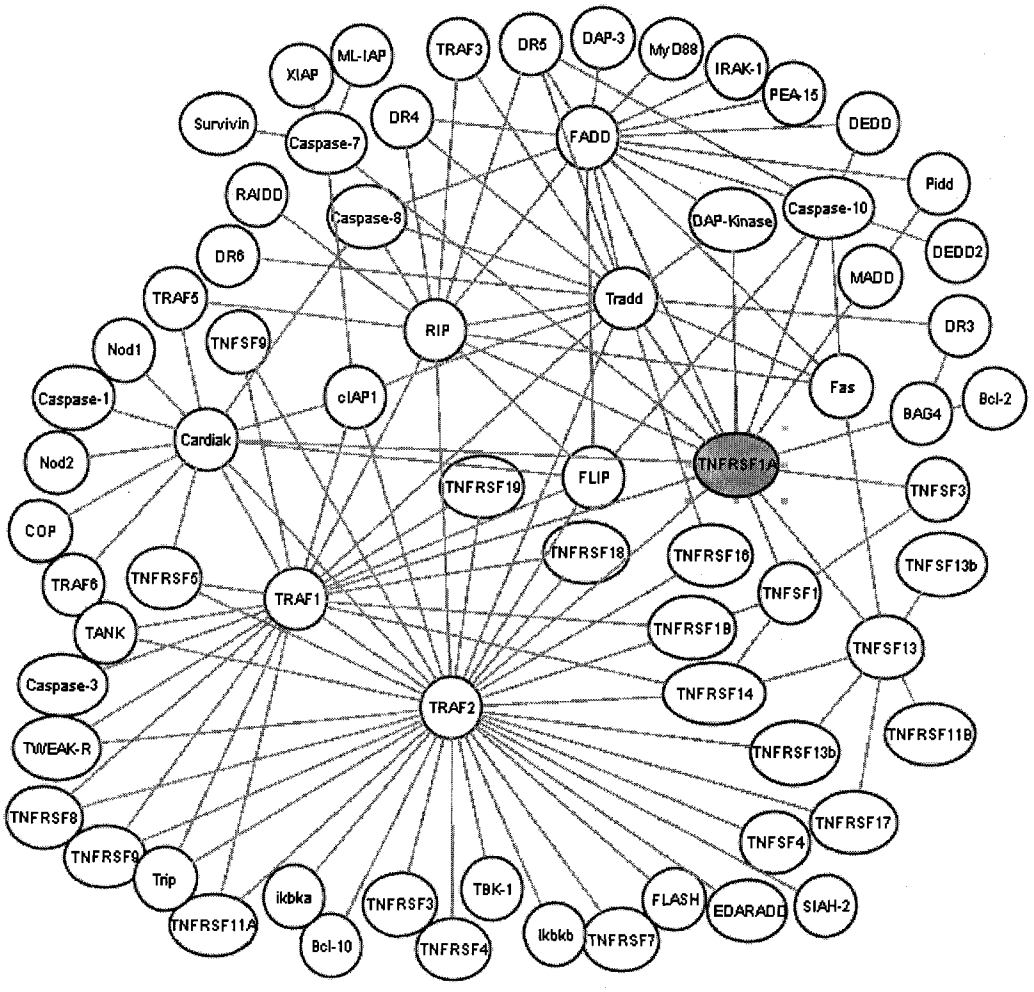
Gene	Number of Interactions	Gene	Number of Interactions
TRAF2	36	Caspase-9	9
Caspase-8	24	CIAP1	9
Bcl-2	23	Bcl-10	8
TRAF1	22	MyD88	8
TRAF3	19	Apaf1	7
Bcl-x	18	Caspase-10	7
TRAF6	18	Caspase-7	7
FADD	16	DR4	7
Tradd	15	DR5	7
RIP	14	Fas	7
TNFRSF1A	14	TNFRSF14	7
Cardiak	13	TNFRSF3	7
FLIP	13	TNFRSF5	7
TRAF5	13	TNFSF13	7
Caspase-3	11		

The genes are sorted by the number of interactions in descending order.

To further make these interaction data useful to molecular biologists, we built a two-layer interaction network for each human apoptosis gene that has protein-protein interaction data with other human apoptosis gene(s). A two-layer interaction network is

defined as a protein interaction network consisting of all the direct interactions of a particular gene and all direct interactions of its direct interacting partners. This interaction network is generally not too complex for better human visualization, yet contains much more information than just listing the direct interacting partners of a gene. It presents a broader context of interactions between related genes that are most likely involved in the same pathway or belong to a multi-protein complex. For example, in this two-layer interaction network centered on TNFR1 (Fig. 5), most genes involved in TNFR1-mediated extrinsic pathway (shown in Fig.1), including TNFSF1, TNFR1 (TNFRSF1A), adapter protein Tradd, RIP, RAIDD are all represented and the interactions between them and other closely related genes are clearly shown. Also, the TNFR1 complex involved in TNFR1-mediated apoptotic pathway includes TNFR1 (TNFRSF1A), TRADD, TRAF2, cIAP1, and kinase RIP1 (Micheau and Tschopp 2003). The two-layer interaction network of TNFR1 demonstrates that components of the TNFR1 complex are interacting with each other either directly or indirectly, which is not evident if we only examine the raw, pairwise interaction data.

Figure 5. A two-layer protein interaction network of human TNFR1 (TNFRSF1A).



This network is constructed using all interactions of the node of TNFR1 (TNFRSF1A), which is highlighted, and all interactions of interacting partners of TNFR1.

3.4 TFBS prediction using both phylogeny and interaction data

We investigated how protein-protein interaction data might be used to improve computational TFBS identification, mainly to reduce false positives and to improve prediction specificity. The prediction performance using TFBS Perl system and filtered with phylogenetic footprints and human protein-protein data is shown in Table 5. The 85% matrix threshold seems to be the best setting. At this matrix threshold, the sensitivity is the same as that of 70% matrix threshold; however, the number of

predictions is much lower, which suggests higher prediction specificity. Less stringent matrix threshold did not increase the prediction sensitivity but only dramatically increased false positives. Additionally, we applied a union/intersection algorithm to combine sequence conservation and protein-protein interaction data and identify binding sites in the conserved regions and/or binding sites shared by human interacting genes. The sensitivity of only keeping sites in the conserved regions drops slightly under all matrix thresholds; by selecting sites either in the conserved regions or shared by interacting genes, the sensitivity is the same as the prediction without any filtering procedures but the number of predictions drops, indicating that many false positives are eliminated and the specificity is improved. The sensitivity of only keeping binding sites shared by interacting human genes is lower than using phylogenetic footprinting information, partly because the interaction data are not comprehensive and more interactions might exist between human apoptosis genes. The sensitivity of control (method 4 in Table 5; predicted binding sites shared by gene pairs that currently have no known interaction data, see Methods for details) is much lower than that of keeping binding sites shared by interacting human genes (method 3). At 85% matrix threshold, the sensitivity of method 3 is 64.8% compared to 35.2% in its control (method 4). However, it is notable that the prediction sensitivity of the control is not extremely low, which could also be at least partly attributed to the fact that more interactions may be discovered between human apoptosis genes.

Table 5. TFBS prediction sensitivity (SN) and specificity (SP) using protein-protein interaction data and phylogenetic footprinting information.

Matrix threshold / Method [1]	70%		80%		85%		90%		95%	
	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP
1	79.4	628	79.4	107	79.4	41	64.7	14	40.1	4
2	77.6	169	75.7	28	75.7	10	50.2	4	34.6	1
3	72.8	465	68.4	69	64.8	26	60.6	8	32.5	2
4	42.6	-	40.7	-	35.2	-	29.6	-	18.5	-
5	79.4	490	79.4	79	79.4	29	64.7	10	40.1	2
6	63.7	165	61.1	18	61.1	7	42.3	2	22.3	0.5

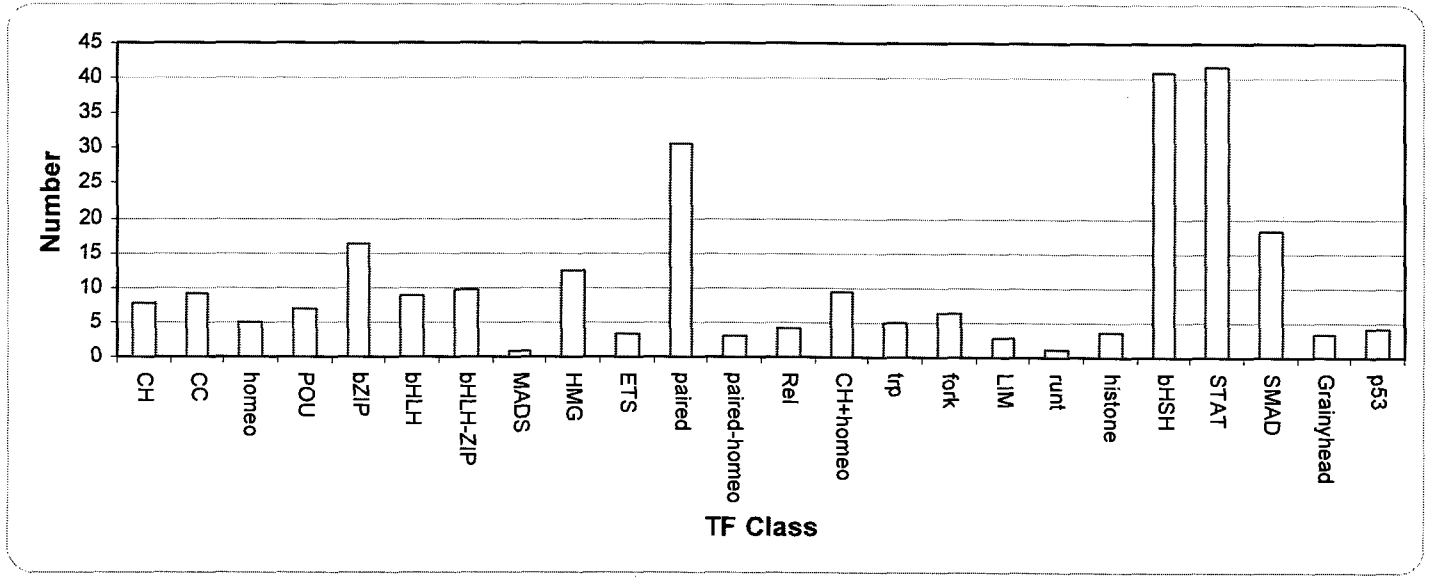
(1) Method: 1 = all predictions without filtering; 2 = only retain binding sites in the conserved regions between human and mouse; 3 = only retain binding sites shared by at least one pair of interacting human apoptosis genes; 4 = control of Method 3, average of 3 times shuffling of records in the interaction table; SP is not applicable for the control; 5 = union: (2) OR (3) (remove a redundant intersection portion); 6 = intersection: (2) AND (3).

3.5 Distribution of binding sites for each TF class

In order to estimate the relative importance of each TF class (family) in apoptosis, we surveyed the number of binding sites for TFs in each TF family in the 3 kb upstream of human apoptosis genes. After normalization by the number of matrices in each TF class, it is shown in Fig. 6 that STAT (e.g. STAT1), bHSH (e.g. AP-2) and paired (e.g. Pax1) classes have the highest numbers of binding sites in human apoptosis genes. STAT factors have been shown to play roles in development, cell growth, proliferation and apoptotic cell death, and there were many studies indicating that TFs in this family control expression of many apoptosis genes, including Bcl-2, Bcl-x, caspase-1, Fas receptor and FasL, and are involved in regulating p53 target genes (Stephanou and Latchman 2003; Stephanou *et al.* 2001; Vousden and Lu 2002). SMAD (e.g. Smad1), bZIP (e.g. c-Fos, c-Jun), CH (e.g. SP1), CC (zinc finger; e.g. GATA-1), forkhead (e.g. E2F), HMG (e.g. SRY) all have relatively high numbers of predicted sites.

However, the numbers of predicted binding site for ETS (e.g. c-ETS-1, PU.1), Rel (e.g. NF-κB1), and p53 TF families might be underestimated. One explanation for this is, the p53 response element is frequently found in the intronic regions of target genes (Mirza *et al.* 2003; Wang *et al.* 2001), and the intronic regions were not included in our predictions. As the average length of matrices is over 10 (see Appendix D) and the minimal length is 8 for the predicted binding sites of each TF class, it is almost unlikely that these binding sites could have occurred just by chance. This is non-trivial, as in an extreme case that a matrix merely has 5 nucleotides long; the number of expected occurrence for a binding site in a 6 kb upstream sequence (3 kb for both strands) would be ~6.

Figure 6. Average number of predicted transcription factor binding sites in the upstream regions.

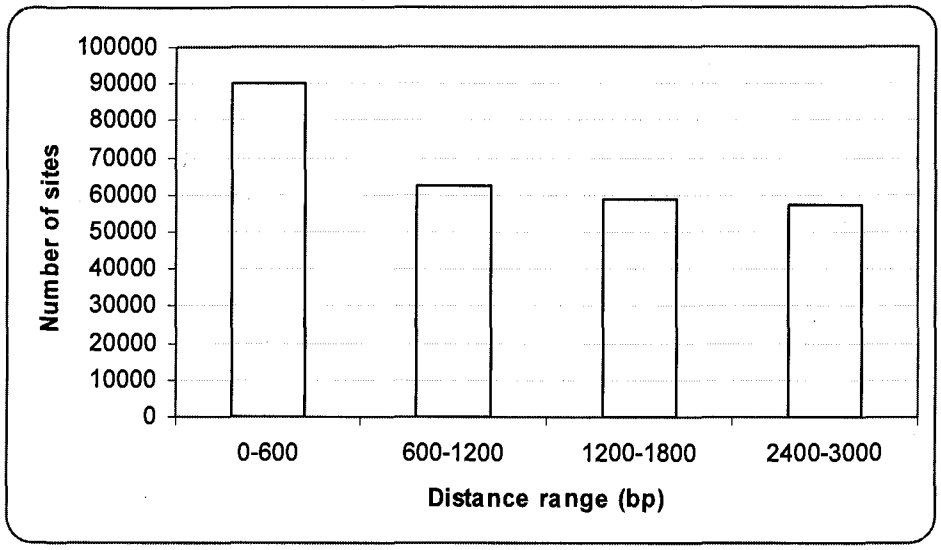


Only 24 TF classes are included, with P53 being considered a TF class. Some TF classes were excluded for at least two reasons: (1) the TF class is not from the vertebrate species; and (2) the PWM matrix describing the binding specificity of the TF class has unequal column sums (i.e. the sequences used to generate the PWM have unequal length) and cannot be used to predict TFBS in the upstream sequences. CH = zinc finger; CC = zinc twist; homeo = homeodomain or homeobox protein; POU = POU domain; bZIP = basic region + leucine zipper; bHLH = basic region + helix-loop-helix motif; bHLH-ZIP = basic region + helix-loop-helix motif + leucine zipper; MADS = MCM1-agamous-deficiens-SRF; HMG = high-mobility group protein-like factors; ETS = Factor family with homology to the ets-protooncogenes; paired = paired domain, paired box; paired-homeo = paired domain, paired box + homeo domain; Rel = Rel-related factor; CH+homeo = zinc finger + homeo domain; trp = tryptophan cluster; fork = fork head domain; LIM-homeo = LIM domain plus homeo domain; runt = runt homology domain; histone = histone fold; bHSH = basic region + helix-span-helix domain; STAT = signal transducers and activators of transcription; SMAD = SMA- and MAD (Mother against DPP) related proteins;

3.6 *In silico* construction of human apoptotic regulatory networks

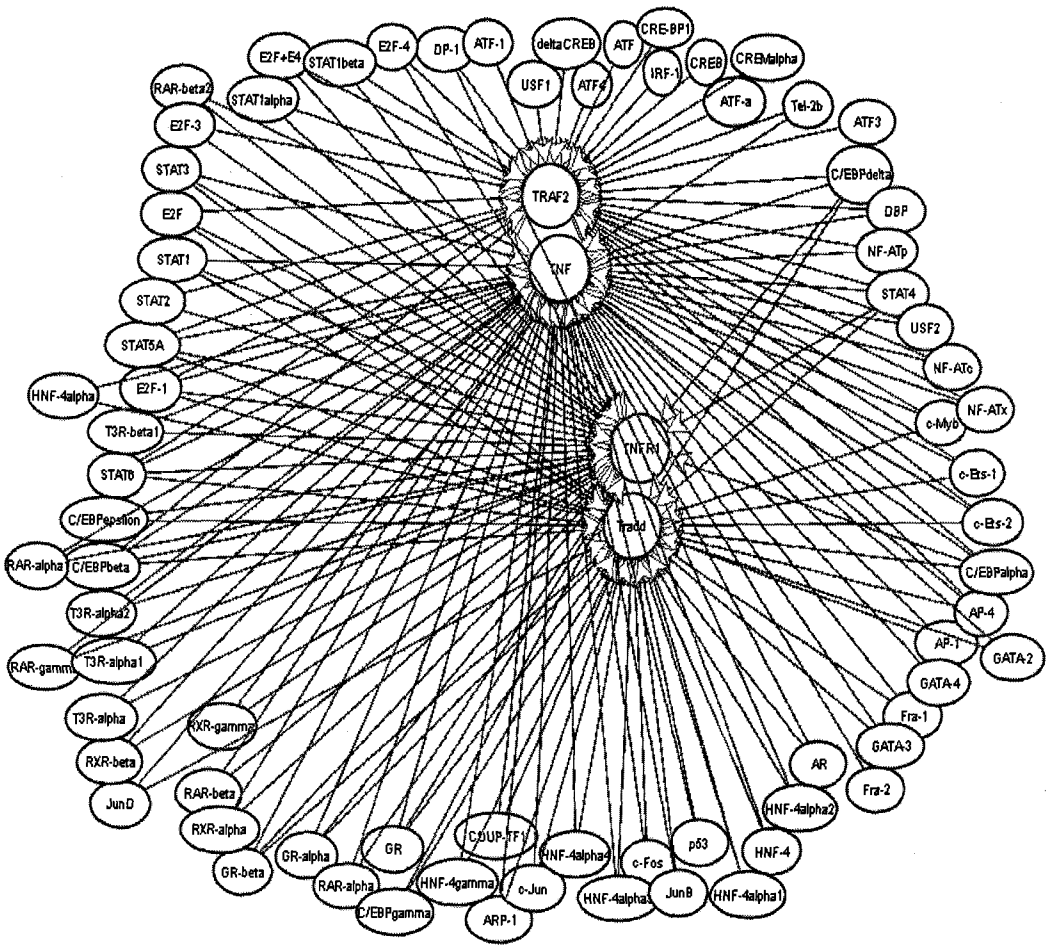
A regulatory network can be defined as a graph in which nodes represent either transcription factors or their regulated genes and edges indicate their regulatory relationship (Bar-Joseph *et al.* 2003; Ideker *et al.* 2002; Lee *et al.* 2002; Pilpel *et al.* 2001). For each of the major apoptosis signaling pathways, i.e. the mitochondrial intrinsic pathway, Fas-mediated extrinsic pathway, and TNFR1-mediated extrinsic pathway, we constructed an *in silico* using transcription factors that have detected binding sites within 600 bp upstream regions of human apoptosis genes. One reason for choosing 0-600 bp upstream regions is that there exist most predicted binding sites in these regions (Fig. 7). A regulatory subnetwork of the TNFR1 pathway is shown in Fig. 8. For human visualization purpose, this subnetwork only includes TNF (TNFSF1), TNFR1 (TNFRSF1A), TRADD, and TRAF2. The network is highly connected, suggesting that few TF families might regulate this pathway and many transcription factors control more than 2 genes in this pathway.

Figure 7. Distribution of predicted binding sites in different upstream regions.



This figure shows the number of all predicted binding sites (85% matrix threshold against all the 490 matrices) either located in conserved regions or shared by interacting gene pairs of human apoptosis genes in different upstream regions of 600 bp intervals. The distance is relative to a RefSeq annotated transcription start site (position 0). The distance range includes the lower boundary but not the upper boundary. For example, binding sites in 0-600 bp contain all sites that start from above 0 (>0) and end with 600 (<=600).

Figure 8. Computationally identified regulatory network for the TNFR1-mediated extrinsic apoptotic-signaling pathway



In this subnetwork, ovals represent transcription factors and circles in the middle represent target genes. An arrow between a TF and its target indicates a potential transcriptional regulatory relationship. For visualization purposes, only 4 key genes that are involved in this pathway and belong to the TNFR1 complex are shown. Table 6 summarizes the RN ratios of the regulatory networks covering all key genes in intrinsic pathway, Fas-mediated extrinsic pathway and TFNR1-mediated extrinsic pathway.

For estimating the potential significance of each transcription factor in each entire regulatory network (figures not shown as they are too complex for human visualization), we calculated its RN ratio of each transcription factor (Table 6). The RN ratio for a transcription factor is the number of target genes that this transcription factor is linked to in the regulatory network divided by the

total number of key genes in each pathway. Transcription factors such as AP-4, AR, C/EBPalpha, C/EBPbeta, C/EBPdelta, DBP, E2F, E2F-1, GATA-2, GR, GR-alpha, GR-beta, NF-κB, STAT1, STAT3, STAT4, STAT5A, and STAT6 appear to possess most connections to genes in the 3 regulatory networks. Other transcription factors connected to genes in two of the apoptotic pathways include AP-1, c-Ets-1, c-Fos, c-Jun, c-Myb, C/EBPgamma, E2F-3, E2F-4, and others. These data could be used to prioritize candidate transcription factors that might coordinately regulate several genes in the same pathway.

Table 6. Potential importance of transcription factors in the regulatory networks of three major apoptotic pathways.

Factor Name	Intrinsic pathway	Fas pathway	TNFR pathway
AP-1	0.875	0.571	0.375
AP-4	0.625	0.571	0.625
AR	0.625	0.714	0.625
ATF	0.750	0.714	0.250
ATF-1	0.875	0.714	0.250
ATF-a	0.750	0.714	0.250
ATF3	0.750	0.857	0.250
ATF4	0.750	0.714	0.250
c-Ets-1	0.500	0.429	0.500
c-Fos	0.875	0.571	0.375
c-Jun	0.875	0.714	0.375
c-Myb	0.250	0.714	0.375
c-Myc	0.625	0.286	0

Factor Name	Intrinsic pathway	Fas pathway	TNFR pathway
C/EBPalpha	1.000	0.857	0.875
C/EBPbeta	0.625	0.714	0.875
C/EBPgamma	0.250	0.714	0.625
C/EBPdelta	0.625	0.714	0.875
C/EBPepisilon	0	0.714	0.625
CRE-BP1	0.875	0.714	0.250
CREB	0.750	0.714	0.250
CREMalpha	0.750	0.714	0.250
DBP	0.625	0.571	0.750
deltaCREB	0.750	0.714	0.250
DP-1	0.250	0.714	0.625
E2F	0.625	0.714	0.750
E2F+E4	0.375	0.714	0.625
E2F-1	0.875	0.857	0.875
E2F-3	0.375	0.714	0.625
E2F-4	0.250	0.714	0.625
Fra-1	0.875	0.571	0.500
Fra-2	0.875	0.571	0.375
GATA-2	0.875	0.571	0.625
GATA-3	0.625	0.571	0.375
GR	1.000	0.714	0.875
GR-alpha	1.000	0.714	0.875
GR-beta	1.000	0.714	0.875
HNF-3alpha	0.625	0.429	0
HNF-3b	0.625	0.429	0
HNF-3beta	0.625	0.429	0
HNF-3gamma	0.625	0.429	0

Factor Name	Intrinsic pathway	Fas pathway	TNFR pathway
JunB	0.875	0.571	0.500
JunD	0.875	0.571	0.500
LCR-F1	0.625	0.571	0
NF-κB	0.750	0.500	0.500
P53	0.375	0.600	0.400
RAR-alpha	0.250	0.714	0.750
RAR-alpha1	0.250	0.714	0.750
RAR-beta	0.375	0.714	0.750
RAR-beta2	0.250	0.714	0.750
RAR-gamma	0	0.714	0.750
RXR-alpha	0.250	0.714	0.750
RXR-beta	0.375	0.714	0.750
RXR-gamma	0	0.714	0.750
STAT1	0.875	0.857	0.875
STAT3	0.750	0.857	0.875
STAT4	0.750	0.857	1.000
STAT5A	0.875	0.857	1.000
STAT6	0.875	0.857	1.000
T3R-alpha	0.375	0.714	0.750
T3R-alpha1	0.375	0.714	0.750
T3R-alpha2	0.250	0.714	0.750
T3R-beta1	0	0.714	0.750

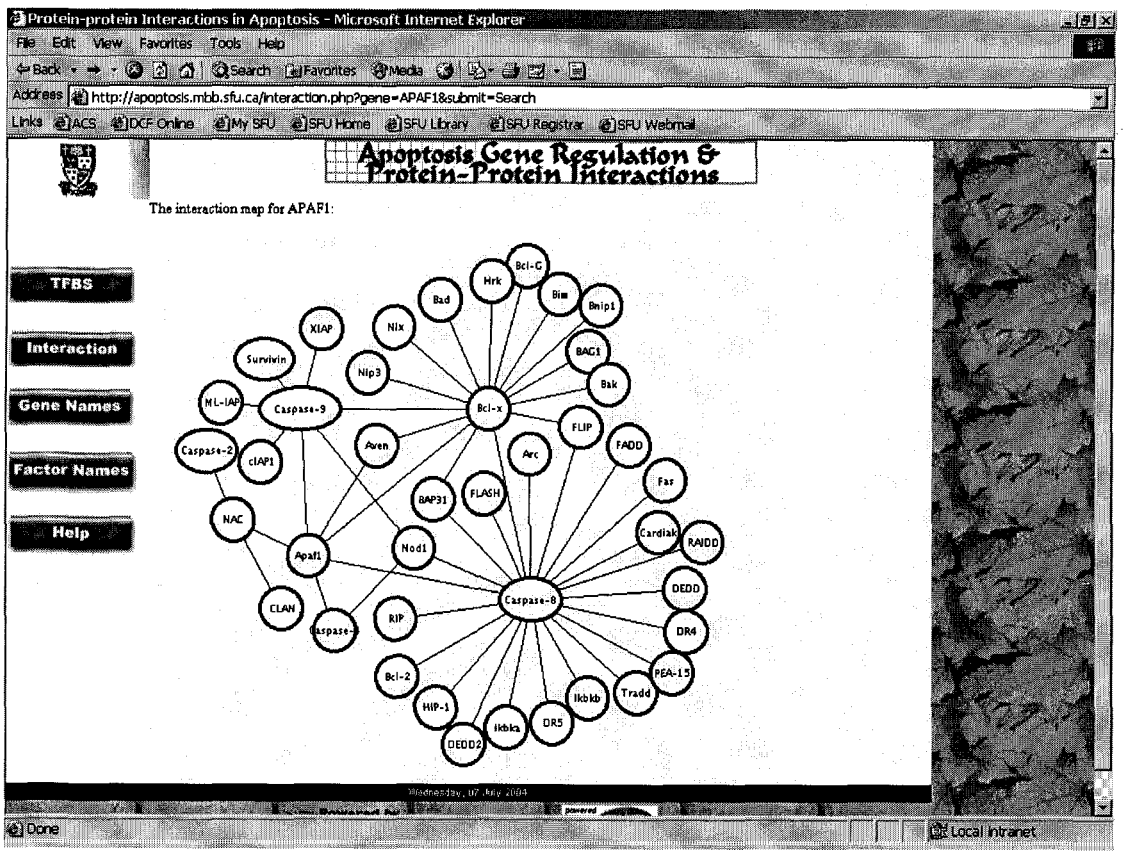
The NF-κB pathway is not included for this analysis as there are only three key components in the pathway. A ratio of 0 indicates that the transcription factor was not in the regulatory network.

3.7 A data-driven website for apoptosis research

One fundamental task of bioinformatics is to provide value-added resources that can directly benefit molecular biologists in their "wet-lab" experiments. Motivated by this philosophy, we developed a database-driven, dynamic web site for apoptosis

researchers. Suppose a biologist needs to know the interactions of most genes involved in the mitochondrial apoptotic signaling pathway, one can find the two-layer interaction network of Apaf1 (Fig. 9), which is ideal for human visualization. Many genes involved in this pathway, such as Apaf1, Caspase-9, NAC, and Caspase-3, are all displayed in this network. If such an interaction network is still complex due to many direct interactions of the gene and many interactions of its interacting partners, the interacting gene pairs can be shown to help identify the interaction gene pairs included in the network.

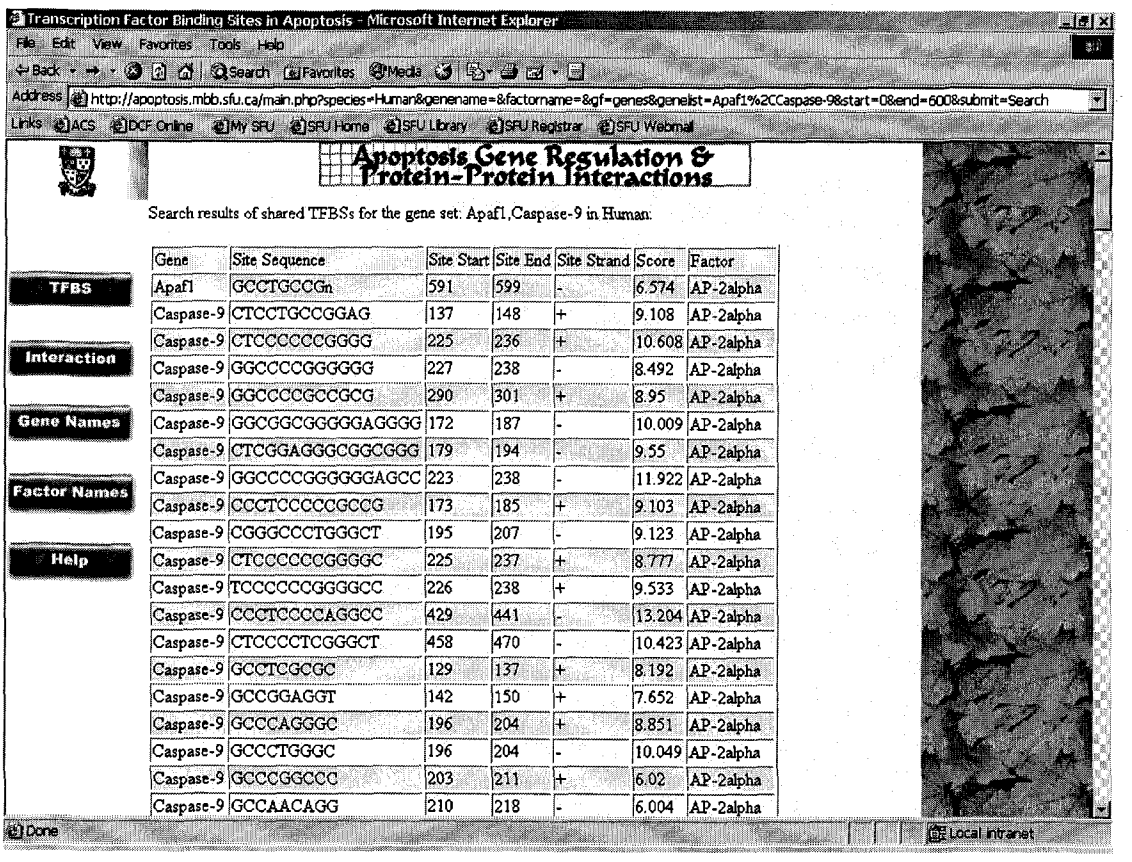
Figure 9. A web-based two-layer protein interaction network for the human Apaf1.



Furthermore, a user can search for all transcription factors common in a set of genes in the apoptotic pathways. This option was intended for searching potentially common transcription factors in components of a multi-protein complex or involved in the

same pathway, but is currently limited as the RefSeq of some genes might not be available and not included for our analyses. Part of the search results for putative common transcription factors in human Apaf-1 and Caspase-9 is shown in Fig. 10. Alternatively, for a transcription factor, one can find which genes in the apoptotic pathways might be its targets of regulation, which is recently termed "regulon" (Simmons *et al.* 2004). Another option is to find all potential transcription factors controlling an apoptosis gene.

Figure 10. The search results of putative transcription factors and binding sites shared by human Apaf1 and Caspase-9.



Gene	Site Sequence	Site Start	Site End	Site Strand	Score	Factor
Apaf1	GCCTGCCGn	591	599	-	6.574	AP-2alpha
Caspase-9	CTCCTGCCGGAG	137	148	+	9.108	AP-2alpha
Caspase-9	CTCCCCCCGGGG	225	236	+	10.608	AP-2alpha
Caspase-9	GGCCCCGGGGGG	227	238	-	8.492	AP-2alpha
Caspase-9	GGCCCCGCCGCG	290	301	+	8.95	AP-2alpha
Caspase-9	GGCGGCGGGGGAGGGG	172	187	-	10.009	AP-2alpha
Caspase-9	CTCGGAGGGCGGCGGG	179	194	-	9.55	AP-2alpha
Caspase-9	GGCCCCGGGGGGAGCC	223	238	-	11.922	AP-2alpha
Caspase-9	CCCTCCCCCGCCG	173	185	+	9.103	AP-2alpha
Caspase-9	CGGGCCCTGGGCT	195	207	-	9.123	AP-2alpha
Caspase-9	CTCCCCCCGGGGC	225	237	+	8.777	AP-2alpha
Caspase-9	TCCCCCCGGGGCC	226	238	+	9.533	AP-2alpha
Caspase-9	CCCTCCCCAGGCC	429	441	-	13.204	AP-2alpha
Caspase-9	CTCCCCTCGGGCT	458	470	-	10.423	AP-2alpha
Caspase-9	GCCTCGCGC	129	137	+	8.192	AP-2alpha
Caspase-9	GCCGGAGGT	142	150	+	7.652	AP-2alpha
Caspase-9	GCCCAGGGC	196	204	+	8.851	AP-2alpha
Caspase-9	GCCCTGGGC	196	204	-	10.049	AP-2alpha
Caspase-9	GCCCGGCCC	203	211	+	6.02	AP-2alpha
Caspase-9	GCCAACAGG	210	218	-	6.004	AP-2alpha

CHAPTER FOUR: DISCUSSION AND CONCLUSION

4.1 Usage of protein-protein interaction data for TFBS prediction

After various genomes have been sequenced and more are being sequenced in a high-throughput manner, several major bioinformatics challenges remain in the post-sequencing genomics era. One of them is the identification of regulatory regions regulating the expression of genes along the genomes. The main difficulty of TFBS prediction lies on the fact that transcription factor binding sites are very short (typically 6-12 bp in eukaryotes; longer in prokaryotes) and degenerate to tolerate considerable sequence variations, and thus many computationally predicted sites can occur randomly in the genome (Lenhard *et al.* 2003; Sharan *et al.* 2003).

If functional genomics data such as microarray gene expression data are available and can be integrated into TFBS predictions, the predictive performance of many computational approaches can be improved. Microarray data have already been used for identifying regulatory elements. For genes clustered by the expression profiles, motif-finding algorithms are used to find over-represented motifs in their upstream regions, since co-expressed genes are co-regulated and may share a set of similar transcription factor binding sites (Qian *et al.* 2003).

Here we explored how protein-protein interaction data might also be used for TFBS filtering. Our assumption is that interacting genes share a similar set of transcription factor binding sites. In prokaryotes a significant proportion of genes that are co-regulated at the transcriptional level code for proteins that interact physically (Teichmann and Babu 2002). In eukaryotes, mainly limited to yeast, gene expression profiles have been shown to correlate with protein-protein interactions (Ge *et al.* 2001;

Jansen *et al.* 2002; Teichmann and Babu 2002). Gene expression profiles are highly correlated for gene products that form multi-subunit protein complexes or involved in the same pathway (Staudt and Brown 2000), and genes with similar expression patterns are more likely to encode interacting proteins (Ge *et al.* 2001). Hannenhalli and Levy (2003) showed that the *cis*-identity, defined as the proportion of shared transcription factor binding sites (TFBS) between two *cis*-element profile (or *cis*-profile, refers to the collection of TFBS regulating the transcription of a gene), is higher for functionally linked interacting proteins as well as for members of a signal transduction pathway, which suggests similar transcriptional control of genes in a complex or pathway. Thus, these authors hypothesize that genes encoding for interacting proteins will be transcribed with a common set of regulatory signals. Simmons *et al.* (2004) studied transcriptional regulation of protein complexes in yeast and showed that the genes in multi-protein complexes are likely to be co-regulated either together or in smaller subgroups. We based our TFBS filtering on the assumption that interacting genes are more likely to share a similar set of transcription factor binding sites.

Our prediction results show that the prediction sensitivity of only keeping binding sites in conserved regions is lower than non-filtering prediction, because some binding sites may not be located in the conserved regions (Lenhard *et al.* 2003). There are studies showing examples of regulatory elements that are found in regions of low sequence identity. For example, there are two functional MARE motifs in pufferfish and chicken that is located in a poorly conserved region compared with human and mouse (Flint *et al.* 2001).

To improve TFBS predictions we used a simple union algorithm with protein-protein interaction data (i.e. retain binding sites either in the conserved regions or shared by at least a pair of interacting genes), the sensitivity is the same as the no-filtering approach, while the specificity is significantly improved (implied by fewer average predictions in a

100 bp upstream sequence).

This approach apparently has limitations. First, though there is some evidence that the interacting genes tend to share a similar *cis*-profile (Hannenhalli and Levy 2003; Simmons *et al.* 2004), it is not certain that this assumption can generally hold true. Interacting genes may have other mechanisms of regulation at translational level or post-translational level. The relationships between gene expression and genome-wide two-hybrid interaction data appear to be more tenuous (Ge *et al.* 2001; Gerstein and Jansen 2000; Jansen *et al.* 2002). Second, it depends on the known protein-protein interactions; some interactions may be yet to be discovered. Thus, with increasing amount of protein-protein interaction data, the TFBS predictive performance of using protein-protein interaction data would definitely improve. Third, the PWM compiled from experimentally determined TFBS available in TRANSFAC database poses a major limitation, because the computational approach described relies on the available DNA binding profiles of known transcription factors (Lenhard *et al.* 2003; Loots *et al.* 2002).

While the current computational tools have success in predicting TBFS within a special context, apparently there is still much scope for improvement. The prediction sensitivity is sometimes reasonably high, but the specificity is extremely low, yielding most false positives randomly distributed along the genomic sequences (Lenhard *et al.* 2003). Innovative algorithms, and perhaps even more importantly, high-quality functional genomics as well as proteomics data sets with respect to gene expression, are required to further advance this bioinformatics field. With the increasing amount of protein-protein interaction data generated by various ongoing proteomics efforts, we suggest that they could be used in a similar manner as microarray data for improving computational TFBS identification.

4.2 Human protein-protein interaction networks and regulatory networks in apoptosis

Many protein-protein interaction data disperse in the literature and recently, efforts have been made to gather these interaction data into databases and provide searchable user interfaces, such as BIND (Bader *et al.* 2003), the human protein reference database - HPRD (Peri *et al.* 2003), and others. However, these sites in general only present the interaction partners of a gene in a pairwise manner, greatly limiting the potential value of these experimentally determined protein interaction data.

For each of the 168 human apoptosis genes that have known protein-protein interaction data, we constructed a two-layer interaction network. These two-layer interaction networks are suitable for human visualization and intended to help biologists in apoptosis research. Each of them consists of protein interactions of related genes that are likely involved in the same pathway or a protein complex. Several multi-protein complexes have been characterized in the major apoptotic signaling pathways. The apoptosome complex, which is formed in the mitochondrial intrinsic pathway, contains Apaf-1, ATP, Caspase-9, cytochrome C, and NAC (Acehan *et al.* 2002; Chu *et al.* 2001; Li *et al.* 1997; Liu *et al.* 1996). The DISC complex in the Fas-mediated extrinsic pathway, and contains Fas (TNFRSF6), FasL (TNFSF6), FADD, and either Caspase-8 or Caspase-10 (Kischkel *et al.* 1995; Muzio *et al.* 1995; Wajant 2002). The inflammasome, which is mainly involved in cytokine activation, contains ASC, NALP1, Caspase-1, and Caspase-5 (Martinon *et al.* 2002). The TNFR1 complex includes TNFR1, TRADD, TRAF2, cIAP1, and the kinase RIP1 (known as complex I) assembles in plasma membrane rapidly to recruit IKK leading to NF-κB activation and survival; and the second complex includes TRADD, RIP1, FADD, Caspase-8 and -10, which forms a cytoplasmic complex (complex II) to initiate apoptosis (Micheau and Tschopp 2003). Components in these complexes must interact with each other, either directly or

indirectly. This information regarding protein complexes does normally not exist in the pairwise protein interaction data. However, if we add another layer of interactions, all these complexes can be represented in two-layer interaction networks (e.g. Fig. 5; all 168 two-layer interaction networks are available at http://apoptosis.mbb.sfu.ca/interaction.php). Hence, we hypothesize that the two-layer interaction network can help biologists to discover other interactions and/or multi-protein complexes in apoptotic pathways.

Transcriptional regulatory networks are key to our understanding of fundamental biological processes, and it can even offer insights into the defect of gene expression that is common in many human diseases (Qian *et al.* 2003). However, linking transcription factors and their target genes presents great challenges to genome biology. In eukaryotes, most studies concerning regulatory networks were performed in unicellular yeast *Saccharomyces cerevisiae* (Bar-Joseph *et al.* 2003; Ideker *et al.* 2002; Lee *et al.* 2002; Pilpel *et al.* 2001). However, there was no previous systematic analysis of transcriptional regulatory networks in human apoptotic signaling pathways.

We computationally constructed a regulatory network for each human apoptotic-signaling pathway and estimated the putative regulatory significance of each transcription factor. These *in silico* regulatory networks are generic (i.e. without a special biological context) and aimed to help understand the potential transcriptional regulation in apoptotic signal transduction. The network is highly connected, suggesting that few TF families might regulate this pathway and many transcription factors might regulate more than 2 genes in this network. Transcription factors such as AP-4, AR, C/EBPalpha, C/EBPbeta, C/EBPdelta, DBP, E2F, E2F-1, GATA-2, GR, GR-alpha, GR-beta, NF-κB, STAT1, STAT3, STAT4, STAT5A, and STAT6 appear to possess most connections to genes in the 3 regulatory networks. Other transcription factors connected to genes in two of the apoptotic pathways include AP-1, c-Ets-1, c-Fos, c-Jun, c-Myb, C/EBPgamma,

E2F-3, and E2F-4. These data could be used to prioritize candidate transcription factors that might coordinately regulate several genes in the same pathway.

To conclude this work, the following work has been done regarding genes that are involved in apoptotic signaling pathway:

1. Compiled genes known to be involved in apoptosis from 5 vertebrate genomes;
2. Extracted known protein-protein interactions between human apoptosis genes and constructed two-layer interaction networks suitable for human visualization;
3. Predicted many putative transcription factor binding sites for transcription factors that might regulate or co-regulate many genes in the apoptotic signaling pathways;
4. Constructed *in silico* regulatory networks to obtain insights into regulation of the major apoptotic signaling pathways;
5. Developed a database-driven web site to make the data available to the apoptosis research community.

REFERENCES

Acehan, D., Jiang, X., Morgan, D.G., Heuser, J.E., Wang, X., and Akey, C.W. 2002. Three-dimensional structure of the apoptosome: implications for assembly, procaspase-9 binding, and activation. *Mol. Cell* **9:** 423–432.

Altschul, S.F., Madden, T.L., Schaeffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acid Res.* **25:** 3389-3402.

Ashkenazi, A. and Dixit V.M. 1998. Death receptors: signaling and modulation. *Science* **281:** 1305–1308.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. 2002. Whole genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301-1310.

Ashe, P.C. and Berry, M.D. 2003. Apoptotic signaling cascades. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* **27:** 199-214.

Bader, G.D., Betel, D., and Hogue, C.W. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acid Res.* **31:** 248-250.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola T.S., Young, R.A., and Gifford, D.K. 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotech.* **21:** 1337-1342.

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

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **31:** 3840-3842.

Blanchette, M. and Tompa, M. 2003. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acid Res.* **31:** 3840-3842.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366:** 265-268.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: efficient tools for large -scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721-731.

Chu, Z.L., Pio, F., Xie, Z., Welsh, K., Krajewska, M., Krajewski, S., Godzik, A., and Reed, J.C. 2001. A novel enhancer of the Apaf1 apoptosome involved in cytochrome c-dependent caspase activation and apoptosis. *J. Biol. Chem.* **276:** 9239-9245.

Danial, N.N. and Korsmeyer, S.J. 2004. Cell death: critical control points. *Cell* **116:** 205-219.

Desnoyers, S. and Hengartner, M.O. 1997. Genetics of apoptosis. *Adv. Pharmacol.* **41:** 35-56.

Ellis, H.M. and Horvitz, H.R. 1986. Genetic control of programmed cell death in the nematode C. elegans. *Cell* **44:** 817-829.

Fesik, S.W. 2000. Insights into programmed cell death through structural biology. *Cell* **103:** 273-282.

Fischer, H., Koenig, U., Eckhart, L., and Tschachler, E. 2002. Human caspase-12 has acquired deleterious mutations. *Biochem. Biophys. Res. Comm.* **293:** 722-726.

Fickett, J.W. and Wassermann, W.W. 2000. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotech.* **11:** 19-24.

Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T., Frampton, J., Alter, B.P., Frischauf, A.M., and Higgs, D.R. 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Human Mol. Genet.***10:** 371-382.

Frith, M.C., Spouge, J.L., Hansen, U., and Weng, Z.P. 2002. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acid Res.* **30:** 3214-3224.

Ge, H., Liu, Z., Church, G.M., and Vidal, M. 2001. Correlation between transcriptome and interactome networkping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29:** 482-486.

Gerstein, M. and Jansen, R. 2000. The current excitement in bioinformatics – analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.* **10:** 574-584.

Gilligan, P., Brenner, S., and Venkatesh, B. 2002. Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene* **294:** 35-44.

Hampson, S., Kiber, D., and Baldi, P. 2002. Distribution patterns of over- represented k-mers in non-coding yeast DNA. *Bioinformatics* **18:** 513-528.

Hannenhalli, S. and Levy, S. 2003. Transcriptional regulation of protein complexes and biological pathways. *Mamm. Genome* **14:** 611-619.

Hengartner, M. O. and Horvitz, H. R. 1994. C. elegans cell survival gene ced-9 encodes a functional homolog of the mammalian proto-oncogene bcl-2. *Cell* **76:** 665–676.

Hockenbery, D., Nunez, G., Milliman, C., Schreiber, R.D., and Korsmeyer, S.J. 1990. Bcl-2 is an inner mitochondrial membrane protein that blocks programmed cell death. *Nature* **348:** 334-336.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. 2002. Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics* **18 (Suppl.1):** S233-S240.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533-538.

Jansen, R., Greenbaum, D., and Gerstein, M. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12:** 37-46.

Kaufman, R.J. 1999. Stress signaling from the lumen of the endoplasmic reticulum: coordination of gene transcriptional and translational controls. *Genes & Dev.* **13:** 1211–1233.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet C.W., Haussler, D., and Kent, W. J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acid Res.* **32 (Database issue):** D493–D496.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. 2004. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* **14:** 160-169.

Kischkel, F. C., Hellbardt, S., Behrmann, I., Germer, M., Pawlita, M., Krammer, P. H., and Peter, M. E. 1995. Cytotoxicity-dependent APO-1 (Fas/CD95)- associated proteins form a death-inducing signaling complex (DISC) with the receptor. *EMBO J.* **14:** 5579-5588.

Kutlu, B., Cardozo, A. K., Darville, M.I., Kruhoffer, M., Magnusson, N., Orntoft, T., and Eizirik, D. L. 2003. Disocvery of gene networks regulating cytokine - induced dysfunction and apoptosis in insulin-producing INS-1 cells. *Diabetes* **52:** 2701-2719.

Lawen, A. 2003. Apoptosis – an introduction. *BioEssays* **25:** 888-896.

Le Bras, M., Bensaad, K., and Soussi, T. 2003. Data mining the p53 pathway in the Fugu genome: evidence for strong conservation of apoptotic pathway. *Oncogene* **22:** 5082-5090.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298:** 799-804.

Lenhard, B. and Wasserman, W.W. 2002. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18:**1135-1136.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W. W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2:** 13.1-13.11.

Levy, S., Hannenhalli, S., and Workman, C. 2001. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17:** 871-877.

Li, P., Nijhawan, D., Budihardjo, I., Srinivasula, S.M., Ahmad, M., Alnemri, E.S., and Wang, X. 2003. Cytochrome C and dATP-dependent formation of Apaf-1/ caspase-9 complex initiates an apoptotic protease cascade. *Cell* **91:** 479-489.

Liedtke, C., Groger, N., Manns, M.P., and Trautwein, C. 2003. The human caspase-8 promoter sustains basal activity through SP1 and ETS-like transcription factors and can be up-regulated by a p53-dependent mechanism. *J. Biol. Chem.* **278:** 27593- 27604.

Liu, X., Kim C.N., Yang, J., Jemmerson, R., and Wang, X. 1996. Induction of apoptotic program in cell-free extracts: requirement for dATP and cytochrome c. *Cell* **86:** 147- 157.

Liu, Y., Wei, L., Batzoglou, S., Brutlag, D. L., Liu, J. S., and Liu, X. S. 2004. A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acid Res.* **32 (Web server issue):** W204-W207.

Liu, Y., Liu, X. S., Wei, L., Altman, R. B., and Batzoglou, S. 2004. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.* **14:** 451-458.

Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E.M. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12:** 832-839.

Manke, T., Bringas, R., and Vingron, M. 2003. Correlating protein-DNA and protein-protein interaction networks. *J. Mol. Biol.* **333:**75-85.

Martinon, F., Burns, K., and Tschopp, J. 2002. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of pro-IL-beta. *Mol. Cell* **10:** 417-426.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak I. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-1047.

Micheau, O. and Tschopp, J. 2004. Induction of TNF receptor I-mediated apoptosis via two sequential signaling complexes. *Cell* **114:** 181-190.

Mirza, A., Wu, Q., Wang, L.Q., McClanahan, T., Bishop, W.R., Gheyas, F., Ding, W., Hutchins, B., Hockenberry, T., Kirschmeier, P., Greene, J.R., and Liu, S. 2003. Global transcriptional program of p53 target genes during the process of apoptosis and cell cycle progression. *Oncogene* **22:** 3645-3654.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. 1998. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14:** 290-294.

Muzio, M., Chinnaiyan, A.M., Kischkel, F.C., O'Rourke, K., Shevchenko, A., Ni, J., Scaffidi, C., Bretz, J.D., Zhang, M., Gentz, R., Mann, M., Krammer, P. H., Peter, M.E., and Dixit, V.M. 1996. FLICE, a novel FADD-homologous ICE/ CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death- inducing signaling complex. *Cell* **85:** 817-827.

Nimwegen, E.V., Zavolan, M., Rajewsky, N., and Siggia, E.D. 2003. Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. *Proc. Natl. Acad. Sci.* **99:** 7323-7328.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K.B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A., and Pandey, A. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13:** 2363-2371.

Pilpel, Y., Sudarsanam, P., and Church, G.M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29:** 153-159.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene - centered resources. *Nucleic Acid Res.* **29:** 137-140.

Qian, J., Lin, J., Luscombe, N.M., Yu, H.Y., and Gerstein, M. 2003. Prediction of networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* **19:** 1917-1926.

Qin, Z. H., McCue, L.A., Thompson, W., Thompson, W., Mayerhofer, L., Lawrence, C.E., Liu, S.J. 2003. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotech.* **21:** 435-438.

Reed, J.C., Doctor, K., Rojas, A., Zapata, J.M., Stehlik, C., Fiorentino, L., Damiano, J., Roth, W., Matsuzawa, S., Newman, R., Takayama, S., Marusawa, H., Xu, F., Salvesen, G., Godzik, A., RIKEN GER Group, and GSL Members. 2003. Comparative analysis of apoptosis and inflammation genes of mice and humans. *Genome Res.* **13:** 1376-1388.

Reimertz, C., Kogel, D., Rami, A., Chittenden, T., and Prehn, J.H. 2003. Gene expression during ER stress-induced apoptosis in neurons: induction of the BH3-only protein Bbc3/PUMA and activation of the mitochondrial apoptosis pathway. *J. Cell Biol.* **162:** 587-597.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C. J., Bell, S.P., and Young, R.A. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306-2309.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole- genome mRNA quantitation. *Nat. Biotech.* **16:** 939-945.

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

Shapira, S., Harb, O.S., Caamano, J., and Hunter, C.A. 2004. The NF-kB signaling pathway: immune evasion and immunoregulation during toxoplasmosis. *Int. J. Parasitol.* **34:** 393-400.

Sharan, R., Ovcharenko, I., Ben-Hur, A., and Karp, R.M. 2003. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **19 Suppl. 1:** i283-i291.

Shi, Y. 2002. Mechanisms of caspase activation and inhibition during apoptosis. *Mol. Cell* **9:** 459-470.

Simonis, N., van Helden, J., Cohen, G.N., and Wodak, S.J. 2004. Transcriptional regulation of protein complexes in yeast. *Genome Biol.* **5:** R33.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12:** 1611-1618.

Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H.R., and Cox, A.V. 2004. The Ensembl web site: mechanics of a genome browser. *Genome Res.* **14:** 951-955.

Staudt, L.M. and Brown, P.O. 2000. Genomic views of the immune system. *Annu. Rev. Immunol.* **18:** 829-859.

Steenbergen, C., Afshari, C.A., Petranka, J. G., Collins, J., Martin, K., Bennett, L., Haugen, A., Bushel, P., and Murphy, E. 2003. Alterations in apoptotic signaling in human idiopathic cardiomyopathic hearts in failure. *Am. J. Physiol. Heart Circ. Physiol.* **284:** 268-274.

Stephanou, A. and Latchman, D.S. 2003. STAT-1: a novel regulator of apoptosis. *Int. J. Exp. Pathology* **84:** 239-244.

Stephanou, A., Scarabelli, T., Brar, B.K., Nakanishi, Y., Matsumura, M., Knight, R.A., Latchman, D.S. 2001. Induction of apoptosis and Fas receptor/Fas ligand expression by ischemia/reperfusion in cardiac myocytes requires serine 727 of the STAT-1 transcription factor but not tyrosine 701. *J. Biol. Chem.* **276:** 28340-28347.

Teichmann, S.A. and Babu, M.M. 2002. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotech.* **20:** 407-410.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4:** 251-262.

Vousden, K.H. and Lu, X. 2002. Live or let die: the cell's response to p53. *Nat. Rev. Cancer* **2:** 594-604.

Wajant, H. 2002. The Fas signaling pathway: more than a paradigm. *Science* **296:** 1635-1636.

Wang, L.Q., Wu, Q., Qiu, P., Mirza, A., McGuirk, M., Kirschmeier, P., Greene, J. R., Wang, Y., Pickett, C.B., and Liu, S. 2003. Analyses of p53 target genes in the human genome by bioinformatics and microarray approaches. *J. Biol. Chem.* **276:** 43604-43610.

Wassermann, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence C.E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225-228.

Wingender, E. 2004. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol.* **4:** 55-61.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. 2001. The TRANSFAC system on gene expression regulation. *Nucleic Acid Res.* **29:** 281-283.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acid Res.* **28:** 316-319.

Xue, D., Shaham, S., and Horvitz, H.R. 1996. The Caenorhabditis elegans cell- death protein ced-3 is a cysteine protease with substrate specificities similar to those of human CPP32 protease. *Genes & Dev.* **10:** 1073-1083.

Xue, W., Wang, J., Shen, Z.R., and Zhu, H.Q. 2004. Enrichment of transcriptional regulatory sites in non-coding genomic region. *Bioinformatics* **20:** 569-575.

Yamamoto, Y. and Gaynor, G.B. 2004. IκB kinases: key regulators of the NF-κB pathway. *Trends Biochem. Sci.* **29:** 72-79.

Yuan, J., Shaham, S., Ledoux, S., Ellis, H.M., and Horvitz, H.R. 1993. The C. elegans cell death gene ced-3 encodes a protein similar to mammalian interleukin-1 beta-converting enzyme. *Cell* **75:** 641-652.

Zhang, Z.L., Harrison, P.M., and Gerstein, M. 2002. Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes. *J. Mol. Biol.* **323:** 811-822.

Zou, H., Henzel, W.J., Liu, X., Lutschg, A., and Wang, X. 1997. Apaf-1, a human protein homologous to C. elegans CED-4, participates in cytochrome c- dependent activation of caspase-3. *Cell* **90:** 405–413.

WEB SITE REFERENCES

http://www.apache.org: Apache web server.

http://apoptosis.mbb.sfu.ca/main.php: A data-driven web site for searching apoptosis regulatory elements and two-layer human apoptotic protein-protein interaction networks, developed for this project.

http://www.biobase.com: TRANSFAC (professional version) web site.

http://www.bioperl.org: BioPerl toolkit.

http://www.cbil.upenn.edu/tess: TESS Transcription Element Search System, including free web access to the public version of the TRANSFAC database.

http://www.ensembl.org: Ensembl web site.

http://www.ensembl.org/Multi/martview: Ensembl EnsMart interface.

http://forkhead.cgb.ki.se/TFBS: TFBS Perl system documentation.

http://www.gene-regulation.com: TRANSFAC database documentation.

http://genome.ucsc.edu/cgi-bin/hgText: UCSC genome table browser.

http://lagan.stanford.edu/lagan_web/index.shtml: LAGAN alignment tool.

http://www.mysql.com: MySQL database server.

http://www.ncbi.nih.gov: National Centre for Biotechnology Information.

http://www.perl.org: Perl programming language web site.

http://www.php.net: PHP scripting language web site.

http://www-gsd.lbl.gov/vista/index.shtml: VISTA alignment and visualization tool.

APPENDICES

Appendix A: The RefSeq accession numbers of apoptosis genes identified in the mammalian genomes.

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
A1a	NM_000295	NM_009742	NM_022519
A1b	NM_130786	NM_007534	-
AIF	NM_004208	NM_012019	NM_031356
AIM2	NM_004833	XM_357160	-
Ankyrin-1	NM_020476	XM_144122	-
Ankyrin-2	NM_001148	NM_178655	XM_227735
Ankyrin-3	NM_020987	NM_170728	-
ANT1	NM_012469	XM_134169	-
ANT2	NM_001152	NM_007451	NM_057102
ANT3	NM_001636	-	-
Apaf1	NM_181861	NM_009684	NM_023979
Apollon	NM_016252	NM_007566	-
Arc	NM_015193	NM_030152	NM_053516
ASC	NM_013258	NM_023258	NM_172322
Aven	NM_020371	NM_028844	XM_230438
Bad	NM_004322	NM_007522	NM_022698
BAFF-R	NM_052945	NM_028075	-
BAG1	NM_004323	NM_009736	-
BAG2	NM_004282	NM_145392	-
BAG3	NM_004281	NM_013863	-
BAG4	NM_004874	NM_026121	-
BAG5	NM_004873	XM_127149	XM_345726

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
Bak	NM_001188	NM_007523	-
BAP31	NM_005745	NM_012060	-
BAR	NM_016561	NM_025976	-
Bax	NM_004324	NM_007527	NM_017059
Bcl-10	NM_003921	NM_009740	NM_031328
Bcl-2	NM_000633	NM_009741	NM_016993
Bcl-3	NM_005178	NM_033601	XM_223405
Bcl-B	NM_020396	NM_013479	NM_053733
Bcl-G	NM_138722	XM_132904	-
Bcl-L12	NM_138639	NM_029410	-
Bcl-w	NM_004050	NM_007537	NM_021850
Bcl-x	NM_001191	NM_009743	NM_031535
Beclin	NM_003766	NM_019584	NM_053739
Bid	NM_001196	NM_007544	NM_022684
Bik	NM_001197	NM_007546	-
Bim	NM_138621	NM_009754	NM_022612
Bimp1	NM_014550	NM_130859	XM_243622
Bimp2	NM_024110	NM_130886	-
Bimp3	NM_032415	NM_175362	-
Bmf	NM_033503	-	NM_139258
Bnip1	NM_001205	XM_355000	-
Bnip2	NM_004330	NM_016787	-
Bok	NM_032515	NM_016778	NM_017312
c-rel	NM_002908	NM_009044	XM_223688
CARD6	NM_032587	-	-
CARD9	NM_052814	-	NM_022303
Cardiak	NM_003821	NM_138952	-
CARP	NM_014391	NM_013468	NM_013220

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
Caspase-1	NM_033292	NM_009807	NM_012762
Caspase-10	NM_001230	-	-
Caspase-14	NM_012114	NM_009809	XM_234878
Caspase-2	NM_032982	NM_007610	NM_022522
Caspase-3	NM_004346	NM_009810	NM_012922
Caspase-4	NM_001225	NM_007609	-
Caspase-5	NM_004347	-	-
Caspase-6	NM_001226	NM_009811	NM_031775
Caspase-7	NM_033338	NM_007611	NM_022260
Caspase-8	NM_001228	NM_009812	NM_022277
Caspase-9	NM_001229	NM_015733	NM_031632
cIAP1	NM_003921	NM_007464	-
CIITA	NM_000246	NM_007575	NM_053529
CLAN	NM_021209	XM_140158	XM_216640
COP	NM_052889	-	-
CPAN	NM_004402	NM_007859	NM_053362
Cryopyrin	NM_004895	NM_145827	-
DAP-3	NM_033657	XM_144039	NM_173138
DAP-Kinase	NM_004938	-	NM_133392
DEDD	NM_004216	NM_011615	NM_031800
DEDD2	NM_133328	NM_026117	-
DFF45	NM_004401	NM_010044	NM_053679
DFFA-likeA	NM_198289	NM_007702	XM_214551
DFFA-likeB	NM_014430	NM_009894	-
DR3	NM_148965	NM_033042	XM_345611
DR4	NM_003844	-	-
DR5	NM_003842	NM_020275	XM_344431
DR6	NM_014452	NM_178589	-

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
EDA-A1	NM_021783	NM_010099	-
EDAR	NM_022336	NM_010100	-
EDARADD	NM_145861	NM_133643	-
EndoG	NM_004435	NM_007931	-
FADD	NM_003824	NM_010175	NM_152937
Fas	NM_000043	NM_007987	NM_139194
FLASH	NM_012115	-	XM_232860
FLIP	NM_003879	NM_009805	NM_057138
FSP27	XM_352506	NM_178373	-
HIP-1	NM_005338	NM_146001	-
HIPPI	NM_018010	NM_028680	-
Hrk	NM_003806	NM_007545	NM_057130
HtrA2	NM_013247	NM_019752	-
Iceberg	NM_021571	-	-
IFI16	NM_005531	NM_008329	-
ikba	NM_020529	-	-
ikbb	NM_002503	NM_010908	NM_030867
ikbe	NM_004556	NM_008690	-
ikbka	NM_001278	NM_007700	-
Ikbkb	XM_032491	NM_010546	NM_053355
Ikbke	NM_014002	NM_019777	XM_344139
ikbz	NM_031419	NM_030612	-
IL-1R	NM_004633	NM_008362	NM_133575
ILP2	NM_033341	-	-
IRAK-1	NM_001569	NM_008363	XM_343844
IRAK-2	NM_001570	NM_172161	-
IRAK-4	NM_016123	NM_029926	-
IRAK-M	NM_007199	NM_028679	-

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
KRC	NM_024503	NM_010657	-
MADD	NM_130470	NM_145527	NM_053585
Mal	NM_002371	NM_010762	NM_012798
MALT-1	NM_006785	NM_172833	-
NETWORK-1	NM_022151	NM_022323	XM_225513
Mcl-1	NM_021960	NM_008562	NM_021846
MEPRIN-1a	NM_005588	NM_008585	NM_013143
MEPRIN-1b	NM_005925	NM_008586	NM_013183
Mil-1	NM_015367	NM_153516	-
ML-IAP	NM_139317	XM_283820	-
MyD88	NM_002468	NM_010851	NM_198130
NAC	XM_293792	XM_193688	XM_340835
NAIP	NM_004536	NM_008670	XM_226742
NFkB1	NM_003998	NM_008689	XM_342346
NFkB2	NM_002502	NM_019408	-
Nfkbil1	NM_005007	NM_010909	-
NGFR	NM_002507	NM_033217	NM_012610
Nip3	NM_004052	NM_009760	NM_053420
Nix	NM_004331	NM_009761	NM_080888
NMP-84	NM_005131	NM_153552	-
Nod1	NM_006092	NM_172729	-
Nod2	NM_022162	NM_145857	-
NOP2	NM_182543	-	-
Noxa	NM_021127	NM_021451	-
p193	NM_006437	-	-
p50	NM_003998	NM_008689	XM_238811
p52	NM_001517	XM_133174	XM_227168
P53	NM_000546	NM_011640	NM_030989

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
P65	NM_031899	NM_028976	-
P73	NM_005427	NM_011642	-
PAN11	NM_153447	NM_011860	-
PAN2	NM_134444	NM_023697	NM_133524
PAN3	NM_175854	-	-
PEA-15	NM_003768	NM_011063	-
Pidd	NM_145886	NM_022654	XM_347291
POP1	NM_015029	NM_152894	-
POP2	NM_022135	NM_022318	NM_199113
Puma	NM_014417	NM_133234	-
PYRIN	NM_000243	NM_019453	NM_031634
Raf1	NM_002880	NM_029780	NM_012639
RAIDD	NM_003805	NM_009950	XM_235060
RelA	NM_021975	NM_009045	XM_238994
RelB	NM_006509	NM_009046	-
RELT	NM_032871	NM_177073	-
RIP	NM_003804	NM_009068	XM_225262
SIAH-1	NM_003031	NM_009172	NM_080905
SIAH-2	NM_005067	NM_009174	NM_134457
Smac	NM_019887	NM_023232	-
Smn	NM_022874	NM_011420	NM_022509
Survivin	NM_001168	NM_009689	-
TANK	NM_004180	NM_011529	NM_145788
TBK-1	NM_013254	NM_019786	-
TEF1	NM_021961	NM_009346	-
TEF2	NM_003563	NM_025287	-
TEF3	NM_003214	NM_197987	-
TEF4	NM_003598	NM_011565	XM_218630

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
TEF5	NM_003214	NM_011566	-
TIRP	NM_021649	NM_173394	-
TLR1	NM_003263	NM_030682	-
TLR10	NM_030956	-	-
TLR2	NM_003264	NM_011905	NM_198769
TLR3	NM_003265	NM_126166	NM_198791
TLR4	NM_138554	NM_021297	NM_019178
TLR5	NM_003268	NM_016928	XM_223016
TLR6	NM_006068	NM_011604	-
TLR7	NM_016562	NM_133211	-
TLR8	NM_016610	NM_133212	-
TLR9	NM_017442	NM_031178	-
TNFR1	NM_001065	NM_011609	NM_013091
TNFRSF10A	NM_003844	-	-
TNFRSF10B	NM_003842	NM_020275	-
TNFRSF10C	NM_003841	-	-
TNFRSF10D	NM_003840	-	-
TNFRSF11A	NM_003839	NM_009399	-
TNFRSF11B	NM_002546	NM_008764	NM_012870
TNFRSF12	NM_148965	NM_033042	-
TNFRSF13b	NM_012452	NM_021349	-
TNFRSF14	NM_003820	NM_178931	-
TNFRSF16	NM_002507	NM_033217	NM_053401
TNFRSF17	NM_001192	NM_011608	-
TNFRSF18	NM_004195	NM_009400	-
TNFRSF19	NM_018647	NM_013869	-
TNFRSF1A	NM_001065	NM_011609	NM_013091
TNFRSF1B	NM_001066	NM_011610	NM_130426

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
TNFRSF3	NM_002342	NM_010736	-
TNFRSF4	NM_003327	NM_011659	NM_013049
TNFRSF5	NM_001250	NM_170701	-
TNFRSF6	NM_000043	NM_007987	NM_139194
TNFRSF6B	NM_032945	-	-
TNFRSF7	NM_001242	XM_284241	-
TNFRSF8	NM_001243	NM_009401	NM_019135
TNFRSF9	NM_001561	NM_011612	-
TNFSF1	NM_000595	NM_010735	NM_080769
TNFSF10	NM_003810	NM_009425	NM_145681
TNFSF11	NM_003701	NM_011613	NM_057149
TNFSF12	NM_003809	NM_011614	-
TNFSF13	NM_003808	NM_023517	-
TNFSF13b	NM_006573	NM_033622	XM_213352
TNFSF14	NM_003807	NM_019418	-
TNFSF15	NM_005118	NM_177371	NM_145765
TNFSF18	NM_005092	NM_183391	-
TNFSF2	NM_000594	NM_013693	NM_012675
TNFSF3	NM_002341	NM_008518	-
TNFSF4	NM_003326	NM_009452	NM_053552
TNFSF5	NM_000074	NM_011616	NM_133542
TNFSF6	NM_000639	NM_010177	NM_012908
TNFSF7	NM_001252	NM_011617	-
TNFSF8	NM_001244	NM_009403	-
TNFSF9	NM_003811	NM_009404	-
Tradd	NM_003789	XM_134502	XM_341671
TRAF1	NM_005658	NM_009421	-
TRAF2	NM_021138	NM_009422	XM_217381

Gene Name	Human RefSeq	Mouse RefSeq	Rat RefSeq
TRAF3	NM_145725	NM_011632	-
TRAF4	NM_004295	NM_009423	-
TRAF5	NM_004619	NM_011633	-
TRAF6	NM_145803	NM_009424	XM_230377
TRIF	NM_014261	NM_174989	NM_053588
Trip	NM_005879	NM_011634	-
TTRAP	NM_016614	NM_019551	-
TUCAN	NM_014959	-	-
TWEAK-R	NM_016639	NM_013749	NM_181086
UNC5H1	XM_030300	NM_153131	NM_022206
UNC5H2	NM_170744	NM_029770	NM_022207
UNC5H3	NM_003728	NM_009472	-
UNC5H4	NM_080872	NM_153135	XM_224934
VDAC1	NM_003374	NM_011694	NM_031353
VDAC2	NM_003375	NM_011695	NM_031354
VDAC3	NM_005662	NM_011696	NM_031355
XEDAR	NM_021783	NM_175540	-
XIAP	NM_001167	NM_009688	NM_022231

Appendix B: The Ensembl IDs of apoptosis genes identified in the *Danio* and *Fugu* genomes.

Gene Name	*Danio* Ensembl Gene ID	*Fugu* Ensembl Gene ID
AIF	ENSDARG00000004596.2	-
Ankyrin-1	ENSDARG00000023273.1	-
Ankyrin-3	ENSDARG00000012091.2	SINFRUG00000132896.1
ANT2	ENSDARG00000023035.1	-
ANT3	ENSDARG00000017611.2	-
Apaf1	ENSDARG00000021239.2	SINFRUG00000151154.1
Apollon	ENSDARG00000016703.2	SINFRUG00000127619.1
ASC	ENSDARG00000025239.2	-
Bad	ENSDARG00000016986.2	-
BAG1	ENSDARG00000020895.2	SINFRUG00000127522.1
BAG2	ENSDARG00000002935.2	SINFRUG00000145698.1
BAG3	ENSDARG00000016349.2	SINFRUG00000124171.1
BAG4	ENSDARG00000003448.1	SINFRUG00000129587.1
BAG5	ENSDARG00000018864.2	SINFRUG00000148315.1
BAP31	ENSDARG00000022311.1	-
BAR	-	SINFRUG00000145185.1
Bax	ENSDARG00000020623.2	SINFRUG00000135145.1
Bcl-10	-	SINFRUG00000155270.1
Bcl-2	ENSDARG00000025613.1	SINFRUG00000155226.1
Bcl-G	ENSDARG00000024762.1	-
Bcl-x	ENSDARG00000008434.2	SINFRUG00000154885.1
Bimp1	-	SINFRUG00000151272.1
Bimp2	ENSDARG00000007176.2	-
Bimp3	ENSDARG00000015105.2	SINFRUG00000153721.1
Bnip1	ENSDARG00000011211.1	-

Gene Name	*Danio* Ensembl Gene ID	*Fugu* Ensembl Gene ID
Bnip2	ENSDARG00000018654.2	-
Bok	ENSDARG00000008082.2	SINFRUG00000125422.1
c-rel	ENSDARG00000003646.2	SINFRUG00000127895.1
CARD9	ENSDARG00000008151.2	-
Cardiak	-	SINFRUG00000125585.1
CARP	ENSDARG00000010568.2	SINFRUG00000140308.1
Caspase-1	ENSDARG00000008165.2	SINFRUG00000143650.1
Caspase-2	ENSDARG00000014202.2	SINFRUG00000123590.1
Caspase-3	ENSDARG00000017905.2	SINFRUG00000150854.1
Caspase-6	ENSDARG00000025608.1	SINFRUG00000153768.1
Caspase-7	ENSDARG00000016228.2	SINFRUG00000149757.1
Caspase-8	ENSDARG00000004166.2	SINFRUG00000137039.1
Caspase-9	ENSDARG00000004325.2	SINFRUG00000151828.1
cIAP1	-	SINFRUG00000121271.1
CPAN	ENSDARG00000009748.2	SINFRUG00000151815.1
Cryopyrin	ENSDARG00000002237.2	SINFRUG00000120866.1
DAP-3	-	SINFRUG00000139205.1
DAP-Kinase	ENSDARG00000010449.2	SINFRUG00000129387.1
DEDD	-	SINFRUG00000126787.1
DEDD2	ENSDARG00000002758.2	SINFRUG00000123753.1
DFF45	-	SINFRUG00000143240.1
DFFA-likeA	ENSDARG00000011058.2	-
DFFA-likeB	ENSDARG00000012640.1	SINFRUG00000128904.1
DR3	ENSDARG00000023511.1	-
DR6	ENSDARG00000028025.1	SINFRUG00000144989.1
EDA-A1	-	SINFRUG00000144147.1
EDAR	ENSDARG00000016846.2	SINFRUG00000137336.1
EndoG	ENSDARG00000013314.2	-

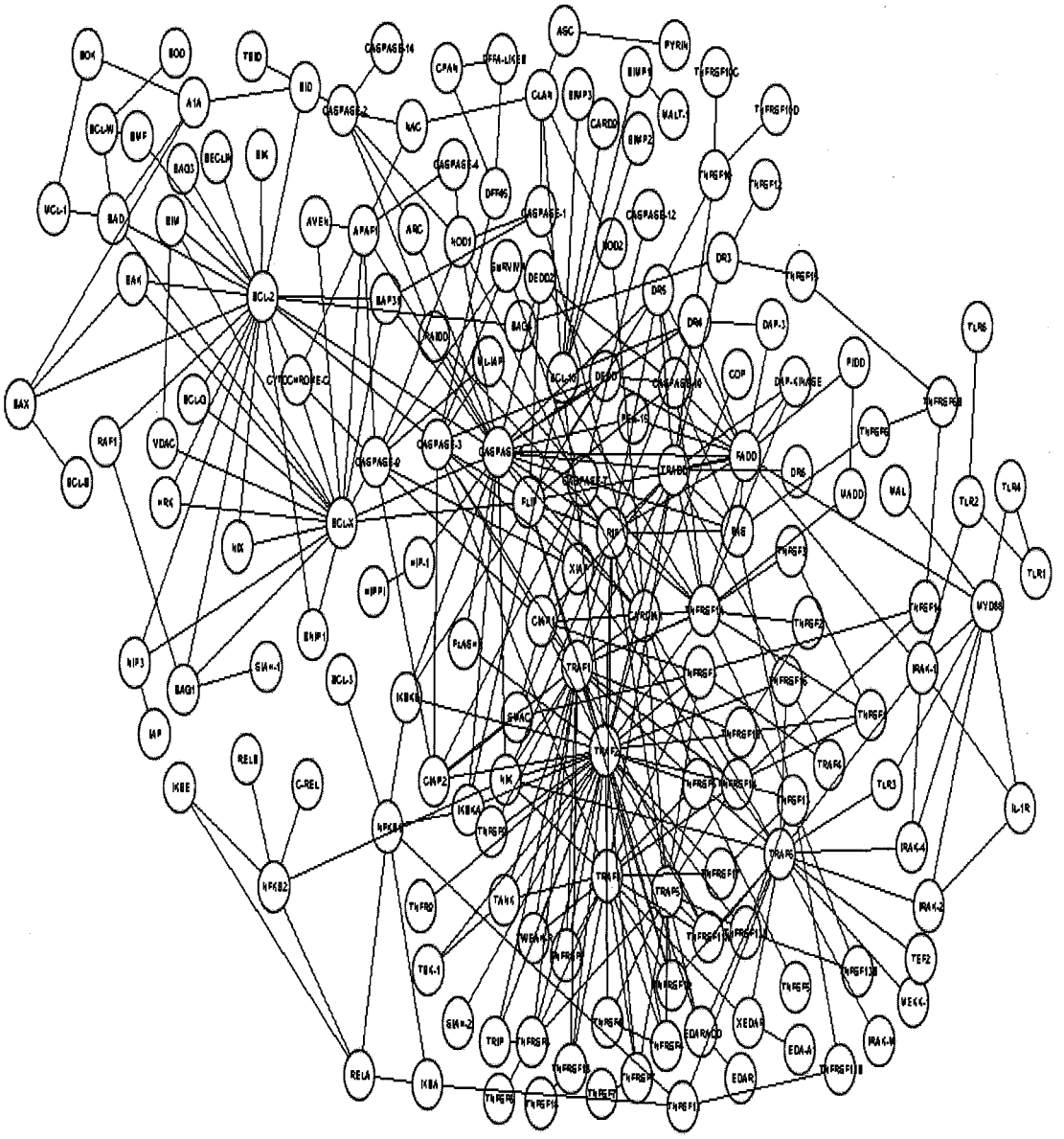
Gene Name	*Danio* Ensembl Gene ID	*Fugu* Ensembl Gene ID
FLASH	ENSDARG00000022718.1	-
FLIP	ENSDARG00000019149.2	-
FSP27	ENSDARG00000002891.1	-
HIP-1	ENSDARG00000012291.2	SINFRUG00000139764.1
HIPPI	ENSDARG00000021022.2	SINFRUG00000149841.1
HtrA2	ENSDARG00000003377.2	SINFRUG00000142884.1
ikba	ENSDARG00000005481.1	SINFRUG00000155476.1
ikbka	-	SINFRUG00000150960.1
Ikbkb	ENSDARG00000011732.2	SINFRUG00000145881.1
Ikbke	ENSDARG00000008987.2	SINFRUG00000154808.1
ikbz	-	SINFRUG00000153830.1
IRAK-1	-	SINFRUG00000137211.1
IRAK-4	ENSDARG00000010657.2	SINFRUG00000122284.1
IRAK-M	ENSDARG00000009541.2	-
KRC	ENSDARG00000002158.2	-
MADD	ENSDARG00000003495.2	SINFRUG00000129322.1
MALT-1	ENSDARG00000006052.2	SINFRUG00000153288.1
NETWORK-1	-	SINFRUG00000127514.1
Mcl-1	ENSDARG00000008363.2	-
MEPRIN-1a	ENSDARG00000008029.2	SINFRUG00000141433.1
MEPRIN-1b	-	SINFRUG00000142288.1
Mil-1	ENSDARG00000012343.2	-
MyD88	ENSDARG00000010169.2	SINFRUG00000143419.1
NAC	-	SINFRUG00000121087.1
NFkB1	-	SINFRUG00000139935.1
NFkB2	ENSDARG00000004043.2	-
NGFR	ENSDARG00000013019.2	SINFRUG00000140313.1
Nip3	ENSDARG00000019785.2	-

Gene Name	*Danio* Ensembl Gene ID	*Fugu* Ensembl Gene ID
Nix	ENSDARG00000025468.1	SINFRUG00000134338.1
NMP-84	ENSDARG00000011938.2	SINFRUG00000123317.1
Nod1	ENSDARG00000009801.2	SINFRUG00000144256.1
Nod2	ENSDARG00000010756.2	SINFRUG00000134206.1
P193	ENSDARG00000024235.1	SINFRUG00000139489.1
P52	ENSDARG00000004772.2	SINFRUG00000148742.1
P53	ENSDARG00000005535.2	-
P65	ENSDARG00000015126.2	-
P73	ENSDARG00000017953.2	-
PEA-15	ENSDARG00000014546.2	SINFRUG00000144776.1
Pidd	ENSDARG00000018596.2	SINFRUG00000133773.1
POP1	-	SINFRUG00000139383.1
POP2	ENSDARG00000012001.2	SINFRUG00000129594.1
PYRIN	-	SINFRUG00000153876.1
Raf1	ENSDARG00000008824.2	-
RAIDD	ENSDARG00000028192.1	SINFRUG00000128869.1
RelA	-	SINFRUG00000124092.1
RelB	-	SINFRUG00000123923.1
RIP	ENSDARG00000006677.2	SINFRUG00000155319.1
SIAH-1	ENSDARG00000003044.2	-
SIAH-2	ENSDARG00000026203.1	SINFRUG00000124017.1
Smac	ENSDARG00000003346.2	SINFRUG00000129198.1
Smn	ENSDARG00000018494.2	-
Survivin	ENSDARG00000015440.2	SINFRUG00000122152.1
TBK-1	ENSDARG00000011399.2	-
TEF1	ENSDARG00000028159.1	SINFRUG00000144598.1
TEF2	ENSDARG00000018779.2	-
TEF3	ENSDARG00000026508.1	SINFRUG00000142204.1

Gene Name	*Danio* Ensembl Gene ID	*Fugu* Ensembl Gene ID
TEF4	-	SINFRUG00000139713.1
TEF5	ENSDARG00000009569.2	SINFRUG00000153549.1
TLR1	ENSDARG00000010871.2	SINFRUG00000136489.1
TLR10	-	SINFRUG00000127281.1
TLR2	ENSDARG00000013167.2	SINFRUG00000148027.1
TLR3	ENSDARG00000016065.2	SINFRUG00000130570.1
TLR4	ENSDARG00000019742.2	SINFRUG00000135468.1
TLR5	ENSDARG00000003558.2	SINFRUG00000153794.1
TLR6	-	SINFRUG00000151268.1
TLR7	-	SINFRUG00000135870.1
TLR8	-	SINFRUG00000153451.1
TLR9	ENSDARG00000008467.2	-
TNFRSF11B	-	SINFRUG00000125406.1
TNFRSF14	ENSDARG00000012428.2	-
TNFRSF16	-	SINFRUG00000136786.1
TNFRSF19	ENSDARG00000025982.1	SINFRUG00000148227.1
TNFRSF1B	-	SINFRUG00000138929.1
TNFSF1	ENSDARG00000013598.2	-
TNFSF10	ENSDARG00000004196.2	SINFRUG00000130060.1
TNFSF13b	ENSDARG00000012945.1	-
TNFSF2	ENSDARG00000009511.2	-
Tradd	-	SINFRUG00000126483.1
TRAF1	ENSDARG00000011321.2	-
TRAF2	ENSDARG00000018205.2	SINFRUG00000123917.1
TRAF3	ENSDARG00000022000.1	-
TRAF4	ENSDARG00000003884.2	SINFRUG00000129233.1
TRAF6	ENSDARG00000007432.2	SINFRUG00000154715.1
Trip	-	SINFRUG00000154904.1

Gene Name	*Danio* Ensembl Gene ID	*Fugu* Ensembl Gene ID
TTRAP	ENSDARG00000016685.2	-
UNC5H1	-	SINFRUG00000129653.1
UNC5H2	ENSDARG00000007437.2	SINFRUG00000120852.1
UNC5H3	-	SINFRUG00000133834.1
UNC5H4	-	SINFRUG00000132851.1
VDAC1	ENSDARG00000021881.1	-
VDAC2	ENSDARG00000013623.2	-
VDAC3	ENSDARG00000003695.2	-
XIAP	ENSDARG00000016143.2	SINFRUG00000127720.1

Appendix C: The human apoptotic protein-protein interaction network



Appendix D: An example position frequency matrix (PFM) for each TF class in Fig. 6, its average matrix length and expected number of occurrences in upstream sequences

TF class	Matrix example	Average matrix length	Expected occurrences
CC	A: 5 11 11 0 50 1 18 6 6 7 C: 22 17 15 0 1 1 6 14 14 11 G: 20 20 10 53 0 1 18 24 18 19 T: 6 5 17 0 2 50 11 9 14 15	11.5	0.0008
CH	A: 32 24 14 17 0 0 19 2 0 21 17 3 9 C: 21 20 10 1 0 2 80 5 1 5 10 55 40 G: 35 56 65 89 108 106 0 99 99 76 72 21 32 T: 20 8 19 1 0 0 9 2 8 6 9 29 27	13.2	5.4×10^{-5}
homeo	A: 6 2 1 2 25 20 2 8 16 3 1 16 9 7 11 C: 0 0 0 0 0 2 0 3 0 1 4 1 11 14 7 G: 16 24 0 0 0 1 0 8 1 1 0 6 3 0 4 T: 3 0 25 24 1 3 24 7 9 21 21 3 3 5 3	15.9	1.6×10^{-6}
POU	A: 6 13 5 12 2 0 26 25 25 15 9 5 C: 11 4 5 1 0 0 0 0 1 5 6 6 G: 2 5 8 2 23 26 0 1 0 2 2 6 T: 7 4 8 11 1 0 0 0 0 4 9 9	13.8	2.9×10^{-5}
bZIP	A: 6 14 19 4 4 36 3 0 2 47 2 10 15 C: 15 12 3 2 2 4 13 0 44 0 8 24 12 G: 19 13 16 5 33 2 29 0 0 0 24 5 11 T: 7 8 9 36 8 5 2 47 1 0 13 8 9	13.0	8.9×10^{-5}
bHLH	A: 1 2 3 0 5 0 0 0 0 0 0 1 C: 2 1 0 5 0 0 1 0 0 1 2 0 G: 2 2 1 0 0 4 4 0 5 2 0 3 T: 0 0 1 0 0 1 0 5 0 2 3 1	12.0	0.0004
bHLH-ZIP	A: 2 2 0 7 0 0 1 0 1 1 C: 3 0 7 0 0 7 0 2 1 1 G: 2 2 0 0 7 0 1 5 3 2 T: 0 3 0 0 0 0 5 0 2 3	14.4	1.4×10^{-5}
MADS	A: 2 3 0 0 19 9 19 1 17 10 0 0 9 4 C: 4 4 21 21 0 1 0 2 0 2 0 0 7 9 G: 11 7 0 0 0 0 1 1 1 2 21 20 5 5 T: 4 7 0 0 2 11 1 17 3 7 0 1 0 3	14.5	1.1×10^{-5}
HMG	A: 6 10 15 21 25 0 28 27 10 16 7 11 C: 6 4 1 2 0 29 0 0 0 3 6 2 G: 10 3 2 1 0 0 0 0 1 5 13 10 T: 7 12 11 5 4 0 1 2 18 5 3 6	9.8	0.007
ETS	A: 8 0 9 2 0 14 14 1 C: 0 3 2 0 0 0 0 3 G: 1 11 3 12 14 0 0 10 T: 5 0 0 0 0 0 0 0	14.6	1.0×10^{-5}

TF class	Matrix example	Average matrix length	Expected occurrences
paired	A: 0 3 1 1 15 17 3 3 7 15 3 C: 18 1 0 1 1 0 14 0 10 2 16 G: 2 0 18 17 3 2 2 0 0 1 0 T: 0 16 1 1 1 1 1 17 3 2 1	16.5	6.6×10^{-7}
paired-homeo	A: 0 3 1 1 15 17 3 3 7 15 3 C: 18 1 0 1 1 0 14 0 10 2 16 G: 2 0 18 17 3 2 2 0 0 1 0 T: 0 16 1 1 1 1 1 17 3 2 1	13.6	3.9×10^{-5}
REL	A: 0 0 1 5 6 5 1 2 0 1 C: 5 0 0 1 5 1 0 0 15 16 G: 8 16 15 9 3 1 0 0 0 0 T: 4 1 1 2 3 10 16 15 2 0	10.0	0.006
CH+homeo	A: 3 5 0 2 1 12 0 0 0 0 4 3 1 C: 3 3 6 1 4 0 12 12 0 0 1 3 3 G: 2 1 1 3 0 0 0 0 0 12 1 6 1 T: 4 3 5 6 7 0 0 0 12 0 6 0 7	11.5	0.0007
Trp	A: 1 6 6 16 16 9 8 20 4 18 0 0 0 11 6 1 2 1 C: 12 15 10 5 11 16 4 1 49 18 0 0 0 1 6 7 2 0 G: 6 5 9 14 8 8 34 25 2 20 60 0 0 47 21 15 1211 T: 8 1 5 3 2 16 5 5 5 4 0 60 60 1 5 11 4 1	14.3	1.5×10^{-5}
forkhead	A: 1 2 1 0 0 0 0 0 5 5 5 5 1 0 2 C: 0 1 2 1 5 0 4 0 0 0 0 0 2 0 0 G: 0 0 2 4 0 5 1 5 0 0 0 0 0 2 3 T: 4 2 0 0 0 0 0 0 0 0 0 0 2 3 0	15.4	3.1×10^{-6}
LIM	A: 3 4 6 0 31 2 2 0 0 0 10 3 C: 16 10 12 31 0 0 8 0 0 14 5 8 G: 8 11 10 0 0 24 21 0 31 8 4 14 T: 3 6 3 0 0 5 0 31 0 9 10 6	12.0	0.0004
runt	A: 8 5 3 0 4 0 0 0 0 0 0 1 8 4 3 C: 4 4 3 4 6 1 0 3 0 0 1 4 0 5 9 G: 3 4 9 6 1 0 17 0 17 17 1 2 1 6 2 T: 2 4 2 7 6 16 0 14 0 0 15 10 8 2 3	15.7	2.1e-06
histone	A: 10 12 3 4 18 19 0 1 41 37 0 4 22 C: 11 17 10 12 5 1 41 39 0 2 1 22 3 G: 10 4 9 9 16 21 0 1 0 0 2 14 14 T: 9 8 19 16 2 0 0 0 0 2 38 1 2	12.4	0.0002
bHSH	A: 4 3 0 0 0 2 2 4 1 2 3 1 C: 6 0 11 13 12 5 8 5 1 2 9 3 G: 3 6 0 0 0 6 3 3 10 9 1 8 T: 0 4 2 0 1 0 0 1 1 0 0 1	12.6	0.0002
STAT	A: 5 9 7 9 4 0 0 0 0 4 9 1 21 C: 6 6 12 5 10 0 0 30 9 2 0 0 3 G: 11 8 3 6 5 0 1 0 1 14 21 23 5 T: 8 7 8 10 11 30 29 0 20 10 0 6 1	12.0	0.0004

TF class	Matrix example	Average matrix length	Expected occurrences
SMAD	A: 0 0 8 0 0 1 3 2 0 0 1 1 1 4 0 0 0 8 0 4 0 8 2 2 0 C: 10 10 0 1 1 5 2 1 10 10 2 1 4 0 0 2 7 1 0 5 10 0 3 7 4 G: 0 0 2 1 0 3 5 2 0 0 7 3 5 6 6 5 2 1 10 0 0 0 3 0 0 T: 0 0 0 8 9 1 0 5 0 0 0 5 0 0 4 3 1 0 0 1 0 2 2 1 6	12.5	0.0002
Grainyhead	A: 5 0 2 0 0 2 1 4 2 6 5 0 1 5 0 C: 0 17 0 0 1 3 2 6 3 6 1 1 7 6 1 5 G: 11 2 0 17 18 6 1 2 10 4 9 2 2 8 14 T: 3 0 17 2 0 8 15 7 4 3 4 0 10 5 0	11	0.001
P53	A: 4 5 15 0 17 0 0 0 0 0 0 0 2 0 17 0 0 0 0 0 C: 0 0 0 17 0 0 0 13 17 17 0 0 0 17 0 0 0 2 13 7 G: 13 12 2 0 0 0 17 0 0 0 17 17 15 0 0 0 17 0 0 2 T: 0 0 0 0 0 17 0 4 0 0 0 0 0 0 0 17 0 15 4 7	17	3.5×10^{-7}

Refer to Fig. 6 for the description of each TF class. The matrix example was the first one retrieved for each class. The expected number of occurrences was calculated based on the average length of matrices in a 6 kb upstream sequence.