Evolutionary Analysis of the Protein Domain Distribution in Eukaryotes

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
ingereichte

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik
vorgelegt
von Arli Aditya Parikesit.,Msc
geboren am 27. Juni 1979 in Jakarta

Die Annahme der Dissertation haben empfohlen:

1. Jun.Prof. Dr. Sonja J. Prohaska (Universität Leipzig, Deutschland)
2. Prof. Dr. Erich Bornberg-Bauer (Westfälische Wilhelms-Universität Münster, Deutschland)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 4. Dezember 2012 mit dem Gesamtprädikat "cum laude".
For Merry and Libby
# Contents

Abstract

Zusammenfassung

1 Introduction

2 Biological Background
   2.1 Gene Structure
   2.2 Protein Domain: Structure and Functions
      2.2.1 Protein Domain as Fundamental Unit
   2.3 Gene Regulation
      2.3.1 Transcription Factor (TF)
      2.3.2 Chromatin Regulation (CR)
      2.3.3 Translation
      2.3.4 Focusing on Transcription Factor and Chromatin Regulation

3 Technical Background
   3.1 Markov Chains, Hidden Markov Models, and the Viterbi Algorithm
      3.1.1 Markov Chain
      3.1.2 Hidden Markov Models
      3.1.3 The Viterbi Algorithm
   3.2 Gene Prediction
      3.2.1 Mechanism of Gene Prediction
      3.2.2 Gene Prediction Packages for this Research
      3.2.3 Problems of Gene Prediction
   3.3 Visualization and Scheme of Protein Domain
Investigations into the origin and evolution of regulatory mechanisms require quantitative estimates of the abundance and co-occurrence of functional protein domains among distantly related genomes. The metabolic and regulatory capabilities of an organism are implicit in its protein content. Currently available methods suffer for strong ascertainment biases, requiring methods for unbiased approaches to protein domain contents at genome-wide scales. The discussion will be highlighted on large scale patterns of similarities and differences of domain contents between phylum-level or even higher level taxonomic groups. This provides insights into large-scale evolutionary trends.

The complement of recognizable functional protein domains and their combinations convey essentially the same information and at the same time are much more readily accessible, although protein domain models trained for one phylogenetic group frequently fail on distantly related sequences. Transcription factors (TF) typically cooperate to activate or repress the expression of genes. They play a critical role in developmental processes. While Chromatin Regulation (CR) facilitates DNA organization and prevent DNA aggregation and tangling which is important for replication, segregation, and gene expression.

To compare the set of TFs and CRs between species, the genome annotation of equal quality was employed. However, the existing annotation suffers from bias in model organism. The similar count of transcripts are expected to be similar in mammals, but model organism such as human has more annotated transcripts than non model such as gorilla. Moreover, closely related species (e.g, dolphin and human) show a dramatically different distribution of TFs and CRs. Within vertebrates, this is unreasonable and contradicts phylogenetic knowledge. To overcome this problem, performing gene prediction followed by the detection of functional domains via HMM-based annotation of SCOP domains were proposed. This methods was demonstrated to lead toward consistent estimates for quantitative comparison. To emphasize the applicability, the protein domain distribution of putative TFs and CRs by quantitative and boolean means were analyzed. In particular, systematic studies of protein domain occurrences and co-occurrences to study avoidance or preferential co-occurrence of certain protein domains within TFs and CRs were utilized.
Pooling related domain models based on their GO-annotation in combination with *de novo* gene prediction methods provides estimates that seem to be less affected by phylogenetic biases. It was shown for 18 diverse representatives from all eukaryotic kingdoms that a pooled analysis of the tendencies for co-occurrence or avoidance of protein domains is indeed feasible. This type of analysis can reveal general large-scale patterns in the domain co-occurrence and helps to identify lineage-specific variations in the evolution of protein domains. Somewhat surprisingly, strong ubiquitous patterns governing the evolutionary behavior of specific functional classes were not found. Instead, there are strong variations between the major groups of Eukaryotes, pointing at systematic differences in their evolutionary constraints. Species-specific training is required, however, to account for the genomic peculiarities in many lineages. In contrast to earlier studies wide-spread statistically significant avoidance of protein domains associated with distinct functional high-level gene-ontology terms were found.


Um Vermeidung und Kookkurrenz von TFs und CRs zwischen verschiedenen Spezies unverzerrt untersuchen zu können, ist es wesentlich, dass deren Genomannotation von ähnlicher Qualität ist. Dazu wurde die Annotation der Domainen mit Hilfe einer Genvorhersage, gefolgt von einer Suche nach funktionellen Domainen mittels HMM-Methoden und bereits annotierten Domainen aus der SCOP Datenbank verbessert. Es konnte dadurch nachweislich die Verzerrung beim quantitativen Vergleich verringert werden. Um die gute Anwendbarkeit dieser Methode hervorzuheben, wurde zudem die Verteilung von
Proteindomainen vermeintlicher TFs und CRs durch quantitative Mittel analysiert. Die Vermeidung oder Kookkurrenz von bestimmten Proteindomainen in TFs und CRs konnte mit Hilfe von systematischen Studien über das Vorhandensein und die Abwesenheit von Proteindomainen untersucht werden.

Aknowledgement

Many thanks to my first advisor Sonja J. Prohaska, who taught me a lot about the latest technique in Bioinformatics and my second advisor Peter F. Stadler, who guided me for my research.

Thanks goes to my colleagues in the EvoDevo group: Axel, Christian Arnold, Irma, Lydia, and Tobias.

I also thanks Axel, Christian Arnold, Irma, Joe, and especially Lydia for proofreading the dissertation.

Thanks to Christian Otto for providing the German translation of the abstract.

Thanks to Wölfi for providing the LaTeX template for the dissertation.

Thanks to Petra for the helpful administration and Jens for excellent technical support.

Thanks also goes to all the other people in the Bioinformatics lab: Bernie, Konstantin, Veiko, Hakim, Sebastian, Gunnar, Gero, Fakhteh, Maribel, Steve, Frank, Stephanie, Helene, David, Jörg, Lydia, Christian, Abdullah, Jan, and Anne.

Many thanks to DAAD fellowship, and Max Planck Society for Mathematics in the Sciences fellowship who funded my research.

Thanks to Mom, Dad, Mitra, Dinda, Jurek, especially to Merry and Libby.

Thanks to my former master thesis advisor, Usman Sumo Friend Tambunan, who introduced me to the world of Bioinformatics.

Publication

This dissertation was written based upon these publications:
• Parikesit, Arli A.; Steiner, Lydia; Stadler, Peter F.; Prohaska, Sonja J. *Pittfalls of Ascertainment Biases in Genome Annotations-Computing Comparable Protein Domain Distributions in Eukarya.* 2012. *MDPI Computation* (on preparation).

• Parikesit, Arli A.; Stadler, Peter F.; Prohaska, Sonja J. *Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions.* 2011. *Genes* 2, no.4: 912-924.


PROTEINS are important building blocks of the cell. The range of an organism’s biochemical capabilities, both metabolic and regulatory, is thus largely encoded in its protein content. Protein was translated from the existing transcript in the cell. Transcription is the process of delivering genetic information to the cell. Mechanism of inheritance in life utilizes DNA, as the genetic material (Alberts et al., 2002). Massive efforts has been invested to annotate the genetic material in living organism, for example in human (Lander and et al, 2001). The whole genetic material from one organism is called genome.

The genome annotation projects have produced massive amount of transcription data. The plausible step for determining their functional feature is by annotating the domain quantity, which is the smallest functional unit of protein. In evolutionary point of view, individual proteins often have multiple ancestors that contributed with different domains to an extant protein (Koonin et al., 2000; Moore et al., 2008). Therefore, they form quite well-defined and stable units of selection (Taverna and Goldstein, 2000; Parikesit et al., 2011c).

The elucidation of the domain content of the existing transcript is possible due to the efforts of some research group. For example, the chromatin mechanism as a powerful computational device has been envisaged (Prohaska et al., 2010). Meanwhile, the domain losses and gain in the eukaryotic genome has been computed as well (Zmasek and Godzik, 2011). It was found also that the emergence of novel domain could be due to the environmental adaptation such as abiotic stress response (Moore and Bornberg-Bauer, 2012).

However, the utilization of existing transcript is not without problem. In Figure 1.1, The alternate splicing of transcript could hampering the annotation process due to its multiple transcripts. Figure 1.1 serves the purpose of explaining why to annotate domains directly in the genome rather than in transcriptome data is more plausible. Figure 1.1 shows that one
genomic region could have several transcripts. When many transcripts were detected, the same domain would be counted several times and creating bias in the calculations. This would pose overcounting domain annotations. It also highlights one of the problems: domains often span splice sites, hence one cannot annotate the genomic DNA without accounting for splicing. The phylogenetic distribution of the transcription factor domain from existing transcript is inconsistent, as seen in the discrepancy between Human and Dolphin (Parikesit et al., 2010a)

Moreover, our computational efforts have shown that existing genome annotation is not reliable for clade-wide comparison (Parikesit et al., 2010b). The Figure 1.2 shows the comparison of the domain counts for human and chimp from the existing annotations. There are clear differences between both human and chimp, as the human transcript annotated more domains than chimp. These differences are unacceptable due to the genomic similarity between both chimp and human. This discrepancy needs to be resolved by a new methodology. Extensive novel annotation effort of genomic DNA is necessary to generate an acceptable domain counts.

The domain annotation problem from the existing transcript has encouraged the development of a method to cope with it. The study is expected to give new insight on the domain distribution annotation among major clades. Most proteins contain more than single domain, thus investigating domain combinations are of particular interest to comprehend their functionality (Yang and Bourne, 2009). Specific domain-combinations for animal or vertebrate are found as well (Itoh et al., 2007). Investigating the domain co-occurrence would be interesting, because its network analysis demonstrates a growing core of combinations in multicellular organisms (Wuchty and Almaas, 2005). The abundance and co-occurrence of domains thus becomes the most natural and promising framework to understand patterns of protein evolution at kingdom-level time-scales, see e.g. (Kim and Caetano-Anollés, 2011; Yang and Bourne, 2009).

The subject of this thesis is to determine the tendency of domain avoidance or co-occurrences in genetic regulation of Eukaryotic genomes. A pipeline has been developed for domain annotation from the gene predictor (Parikesit et al., 2010b, 2011c). ADD (Arli’s Domain Distribution) pipeline is expected to give a fine-grained domain annotation, by using gene predictor package. The complete workflow of ADD pipeline could be seen in Figure 1.3. The ADD pipeline utilizes gene prediction for generating non-overlapping ORF, and annotate its domain content. Generated data from the gene prediction will be compared with the one from the existing transcripts, by using statistical analyses.
Figure 1.1: Alternate splicing. The color scales are as follows: yellow is UTR, domain 1 is blue, domain 2 is green, domain 3 is red, brown is exon. Several of the transcript have different domain and region content. Transcript 1 and 2 are having the same domain contents, but different UTR length. Transcript 3 and 4 are having two domain, which one of them are different with each other. Transcript 5 has no annotated domain.

Figure 1.2: Correlation of the number of protein domains. Domain prediction based on existing annotation yield systematic differences between human and chimp (Parikesit et al., 2010b)
Figure 1.3: Workflow for the estimation of domain abundance data. *De novo* gene annotation (l.h.s) was started, here using AUGUSTUS to obtain a collection of non-overlapping protein predictions that is as unbiased as possible. Moreover, the standard pipeline for protein domain annotation (r.h.s) is using the existing RefSeq database, by detecting its ORF. The figures in blue signify the detected genes, green signifies predicted Open Reading Frame (ORF), while red, orange, and purple signify annotated protein domain. Most studies instead start from protein databases that suffer from a variety of ascertainment biases. Protein domains from the Pfam or SUPERFAMILY database are mapped to the known or predicted proteins and form the basis for subsequent statistical analyses.
PROTEINS are very important in the cell. They are responsible for genetic regulation and various catalytic activities. The proteins are synthesized during the translation step by using the mRNA as a template (Sengbuch, 2003; Alberts et al., 2002). mRNA is copied by the transcription process from the genomic DNA. This system that governs the flow of the genetic information is called as 'the central dogma of molecular biology' (Figure 2.2) (Alberts et al., 2002; Wang and Shi, 2009).

As a template, genomic DNA was copied by the polymerase itself into RNA, for synthesizing protein as an end product (Alberts et al., 2002). Transcription is an information transmitting process from the DNA molecule to the RNA as the primary transcript (pre-mRNA) (Clancy, 2008a; Alberts et al., 2002; Lodish et al., 2000). This was followed by splicing process which is a modification of the nascent pre-mRNA taking place after or concurrently with its transcription. In this modifications, the introns are removed and exons are joined (Wu et al., 2006). An exon is a nucleic acid sequence that is represented in the mature form of an RNA molecule after portions of a pre-RNA (introns) have been removed (Mungall et al., 2011; Alberts et al., 2002).

The cell is the smallest living entity, as described by (Putnam, 2011). It comprises a set of different compartments, and for eukaryotes, it has a clearly defined nucleus. The chromosome, which is the organized structure of DNA and protein, is found most of the time in the nucleus. The tips of the chromosome are called telomeres, while its center is called centromere. When the chromosome was unwinded, it has the chromatin, which is the aggregate of histones and DNA. Last, the DNA is the molecule which inherited the genetic information (Figure 2.1) (Putnam, 2011).

The term 'central dogma' must be cited with caution, due to the old paradigm of
‘protein centric world’. In this paradigm, the role of the non-coding (nc)RNA is considered unimportant for the flow of genetic information. In Figure 2.2, the old paradigm is marked in bold. This point of view is no longer valid anymore because the role of ncRNA is important as well (Mattick, 2003). RNA splicing is a precondition for an exit from the nucleus to undergo the protein translation (Dahlberg et al., 2003; Alberts et al., 2002). The information content of the RNA could be altered by those preconditional steps, and this is important on how the cells, especially eukaryotes, transcribed the genome (de Roos, 2007; Alberts et al., 2002).

Finally, although the focus of this dissertation is the annotations of the proteins encoded by the genome, for some genes, RNA is the clear final product (Strachan and Read, 1999; Alberts et al., 2002). Like proteins, many of these RNAs fold into precise three-dimensional structures that have structural and catalytic roles in the cell (Alberts et al., 2002; Mattick, 2003). Moreover, it is shown in the Figure 2.2, that non-coding and SnoRNAs have also cardinal function which are important as coding RNAs (Mattick, 2003).
2.1 Gene Structure

Figure 2.2: Traditional and modern view of the genetic information flow, as proposed by (Mattick, 2003). The traditional view of genetic information flow was marked with bold arrows consisting only of protein coding genes. The modern view that primarily exist in eukaryotes, was included by thin arrows as well. In this new paradigm, the complicated networks of gene interaction was supported by various protein isoforms, as well as non-coding RNA (ncRNA). Figure modified from Mattick (2003), Fig.1.

2.1 Gene Structure

The Figure 2.3 shows the difference between prokaryotic and eukaryotic mRNA (Alberts et al., 2002; Cooper, 2000). Prokaryotic mRNA has multiple translation start sites, while eukaryotic mRNA has single translation start site.

**Promoter** It is a genomic region in front of a gene, which directed the binding of RNA polymerase to the DNA (Alberts et al., 2002). The purpose is to initialize the RNA synthesis when more proteins are needed (Alberts et al., 2002).

A type of promoter in eukaryotes is shown in the Figure 2.4. The Figure shows some regulatory elements, which are necessary for the transcription initialization. They are, i.e,
Biological Background

The Scheme of Prokaryotic and Eukaryotic mRNA

Figure 2.3: Prokaryotic and Eukaryotic mRNAs. UTRs at the 5’ and 3’ ends are the part of both types of mRNAs (Cooper, 2000). m7G caps, as well as 3’ poly-A tails, can only be found at eukaryotic mRNAs (Decker and Parker, 2002; Cooper, 2000). Figure taken from (Cooper, 2000), Chapter 7, Fig.6.

TATA box, DNA spacer, and regulatory sequence.

The TATA box is a DNA sequence (cis-regulatory element) found in the promoter region of genes in archea and eukaryotes. Approximately 24% of human genes contain a TATA box (Yang et al., 2007).

Spacer DNA are regions of non-transcribed DNA between tandemly repeated genes, such as ribosomal RNA genes in eukaryotes, and its function most likely involves ensuring the high rates of transcription associated with these genes (Alberts et al., 2002).

A regulatory sequence is a segment of DNA where regulatory proteins such as transcription factors bind preferentially. It is appropriately positioned in the genome; usually, a short distance ‘upstream’ of the gene being regulated. The binding would force these regulatory proteins to recruit another protein complex, e.g the RNA polymerase. In this way, they control gene expression and protein bio-synthesis as well (Stepanova et al., 2005).

**Coding and Non Coding sequence**  Coding sequences are the DNA and RNA sequences that are translated into proteins (Clancy and Brown, 2008). It is composed of exons encoding
2.1 Gene Structure

Figure 2.4: Promoter in eukaryotes. The assembling of polymerase and transcription factors is happening in the promoter DNA sequence. More explanation on the regulatory elements could be found in the main text (Gilbert, 2000a; Alberts et al., 2002). Figure taken from (Alberts et al., 2002), Fig.7-41

The boundaries of this region are the start codon close to 5' end and the stop codon close to 3' end.

The non-coding sequences are the DNA and RNA sequences that not translated into protein (Pray, 2008). However, the non-coding sequences could still contain function, e.g. ncRNA or snoRNA (Brown, 2002; Alberts et al., 2002). Non-coding sequences also have some recognizable specific features. They are non coding functional RNA, cis- and trans-regulatory elements, introns, pseudo-genes, repeat sequences, transposons, and telomeres. They are explained in the following:

- Non-coding functional RNAs are the RNA molecules that are not translated into proteins. Examples are rRNA, tRNA, and miRNA (Clancy, 2008c).
- Cis- and Trans-regulatory elements are the control sequence for the gene transcription, and may be located in 5' or 3' UTR or within introns (Morello and Breviario, 2008).
- Introns are sections of a gene that transcribed into the pre-mRNA sequence, but eventually omitted by RNA splicing (Brown, 2002).
- Pseudo-genes are related to the known genes that have been disabled to encode proteins (Torrents et al., 2003).
- Repeat sequences are the repetitive DNA sequences. These are tandem and interspersed repeat. Tandem repeat happens when repetitions of two or more nucleotides pattern are directly adjacent to each other. While interspersed repeats are the repetitive DNA sequences that dispersed into different locations in the genome (Strachan and Read, 1999).
- Transposons are DNA sequences that could change its relative position (self-transpose) in the genome (McClintock, 1950).
• Telomeres are a sequence region in the tip of the chromosomes that protected its structure from damage or collisions with other chromosomes (Alberts et al., 2002).

**Splice Site**  Splice site is a site for RNA splicing process that removes the intron and also the site of joining the exons sequences from the pre-mRNA (Clancy, 2008d; Alberts et al., 2002). Typically, eukaryotic pre-mRNA always include introns, and they are spliced by the snRNPs (small nuclear ribonucleoprotein particles). Then, the exons are joined (spliced) forming the mature mRNA. The scheme of the RNA splicing is shown in Figure 2.5.

![Figure 2.5: The Scheme of RNA Splicing](http://www.phschool.com/science/biology_place/biocoach/transcription/premrna.html)

A splice site represents the boundary between an exon and an intron, and its sequence is weakly conserved (Faustino and Cooper, 2003). These sequences are recognized by the splicing machinery, the spliceosome, that will splice them out (Clancy, 2008d). In each intron, they are found three sequence regions that are important for processing by the spliceosome. They are: the 5’ splice site, polypyrimidine tract with adenosine branch point, and the 3’ splice site (Faustino and Cooper, 2003). The splice sites in the different organism could have different consensus sequence, as shown in Figure 2.6 (Semlow and Staley, 2012).

Alternative splicing is a process by which the exons of the RNA produced by transcription of a gene (a primary gene transcript or pre-mRNA) are reconnected in multiple ways during RNA splicing. The resulting different mRNAs may be translated into different protein isoforms; thus, a single gene may code for multiple proteins (Black, 2003). Figure 1.1 shows that by
2.1 Gene Structure

Figure 2.6: The Scheme of Splice Sites in the pre-mRNA. The splice site consensus sequences for metazoa and budding yeast are shown in the figure. N: any bases. R:Purine. Y:Pyrimidin. Figure taken from (Semlow and Staley, 2012), Fig.1.

alternative splicing one region of DNA could encode for several different mRNA transcripts. Alternative spliced transcripts are generated when a variety of splice sites are chosen during pre-mRNA splicing (Dou et al., 2006).

Moreover, the Figure 2.7 shows the sequence logo of the splice sites. It was generated by (Rose, 2010). Here, it is shown that the frequency pattern of protein coding genes, non-coding genes, and UTRs are very similar.

Codon/ORF/start and stop codon Three consecutive nucleotides encoding one amino acid are called codon (Lodish et al., 2000). Most of the time, the three nucleotides only specify one amino acid. The set of codons is referred as 'the genetic code' (Clancy, 2008b). Figure 2.8 refers to the codon table that is called as the genetic codes.

An open reading frame (ORF) is the part of the mRNA that encodes for a protein. ORFs begin and end with specific codons, i.e start and stop codon. ORFs are always divisible by 3.

The start codon is the beginning of the mRNA, and it is needed for translation initialization (Ye, 2008; Alberts et al., 2002). It is almost always preceded by an 5' UTR. Alternate start codons (non ATG) are scarce in eukaryote’s nuclear genome. Mitochondrial genomes and prokaryotes use alternate start codons more extensively (mainly GUG and UUG). For example, E. coli employs 83% ATG (AUG), 14% GTG (GUG), 3% TTG (UUG) and one or two others (e.g., ATT and CTG). AUA codes for isoleucine in most organisms but for methionine in vertebrate mitochondrial mRNA (Xia et al., 2007). Bioinformatics programs usually permitting alternate start codons when searching for protein coding genes (Blattner et al., 1997).

The protein-coding was terminated by the existence of the stop codons, that is, one of the three of them (UUA, UAG, or UGA) (Sengbuch, 2003; Cooper, 2000). tRNA is not distinguishing them, and it did not point out to a specific amino acid, but act as a stop translation signal to the ribosome (Bertram et al., 2001; Alberts et al., 2002; Scheper et al., 2007). However, it is worthwhile to mention, that the alternative stop codons of AGA and AGG have been found as well (Sengupta and Higgs, 2005; Ivanov et al., 2001).
Figure 2.7: The Sequence Logo of the vertebrate Splice sites. The figure depicts nucleotide frequencies of real splice sites. It suggests that splice site of non-coding genes evolve similar to sites of protein-coding genes or UTRs (Hiller et al., 2009; Rose et al., 2011; Rose, 2010). Figure taken from (Rose, 2010), Fig. 3.3
2.1 Gene Structure

Figure 2.8: The genetic code: The dissimilarity of the nuclear and mitochondrial codon table, and the codons in the blue boxes are interpreted differently in both cases, while the mitochondrial version was proposed in blue. Figure taken from (Strachan and Read, 1999), Fig.1.22.

UTR  It is also known, that regulation of translation is due to signals in the UTR (Gray and Wickens, 1998). Untranslated region (UTR) refers to the two section around the coding sequence on the mRNA strand and these are two types, the 3'UTR and 5'UTR (Tuller et al., 2009). mRNA localization was directed by signals that located in the UTR (Kim-Ha et al., 1993; Alberts et al., 2002). The sequence logo of the splice sites is exposed in Figure 2.7 (Rose, 2010). Moreover, some notable examples of the UTR regulatory elements are the Interferon gamma 5' UTR regulatory element, the Bicoid 3'-UTR regulatory element, and the apolipoprotein B (apoB) 5' UTR cis regulatory element (Pontrelli et al., 2004; Macdonald and Struhl, 1988; Ben-Asouli et al., 2002)

The 50 base pairs of the intervening sequence were laid between transcription initiation points and translation, which is the leader sequence or 5' untranslated region (UTR) (Waters and Storz, 2009; Alberts et al., 2002). It could influence the translation initiation rate (Alberts et al., 2002). The ribosomal binding site (RBS) is a sequence on mRNA that is bound by the ribosome when protein translation is initiated (Komarova et al., 2002). It is located in 5' UTR (Laursen et al., 2005). The scheme of translation initiation is shown in Figure 2.9. In bacteria, 5'caps addition for translation initialization is not exist (McCarthy, 1998; Alberts et al., 2002; Cooper, 2000). The Shine-Dalgarno sequence marks the ribosomal-binding site of the bacteria and it is located some nucleotides before the AUG initiation codon (Alberts et al., 2002; Cooper, 2000).

3' UTR is a wide region, starting from the termination of the protein synthesis (stop codon), to the beginning of the poly-A tail (Alberts et al., 2002). The sequence AATAAA was included in this region, and it serves as polyadenylation signal (Dhaese et al., 1983; Alberts et al., 2002). At this sequence starts the ‘tail’ of the RNA transcript, which is 200 to 300 adenosine
Biological Background

Figure 2.9: Translation initiation. Shine-Delgarno sequence are the initiation sites in prokaryotic mRNA, and it precedes the AUG initiation codon (Cooper, 2000). The 40S ribosomal subunit was hitched to eukaryotic mRNA by the assistance of m7G caps (Jackson et al., 2010; Cooper, 2000). Figure taken from (Cooper, 2000), Chapter 7, Fig.7.

residues long(Gilbert, 2000b; Alberts et al., 2002).

**Polycistronic genes** It is defined as a single mRNA that encode for two or more peptides or proteins (Pi et al., 2009). Polycistronic genes are the feature of prokaryotic organism (Kozak, 1999). The Figure 2.3 shows a prokaryotic mRNA which is polycistronic. For example, the transcriptomes of Kinetoplastida (Leishmania and Trypanosoma) which have unusual structures, consists of polycistronic mRNAs (Martínez-Calvillo et al., 2010). Moreover, some polycistronic mRNAs are also found in Caenorhabditis elegans and Drosophila melanogaster (Pi et al., 2009). In Figure 2.10, the mechanism of polycistronic transcription in Trypanosoma is explained. The transsplicing process of spliced leader RNA to the 5’ end of the 5’ UTR of protein coding genes is the primary factor of the mRNA maturation, as shown in Figure 2.10. The polycistronically transcribed protein-coding gene of Trypanosoma brucei is transformed by coupled trans splicing and polyadenylation into monocistronic mRNAs (Siegel et al., 2011).

**Transsplicing** It is the union of two different primary RNA transcript into one by ligating them end to end. It results in one RNA transcript (Clancy, 2008d). For example, the
Figure 2.10: Polycistronic transcription and processing of mRNAs in trypanosomes. The cleavage of polycistronic transcription units was lead by the transsplicing of spliced leader RNA to the 5' end of the 5' UTR of protein coding genes. The 5' conserved spliced leader sequence and 3' poly(A) tail are the part of mature mRNA. The spliced leader sequence was 'bridged' with sequence tags. The mapping of splice acceptor sites (SAS) and polyadenylation sites (PAS) could be done by the 5' end of a gene, the poly(A) tail and the 3’ end of a gene. Figure taken from (Siegel et al., 2011), Fig.1.

Polycistronic mRNAs of Kinetoplastida are produced by transsplicing (Martínez-Calvillo et al., 2010). The Figure 2.11 shows the mechanism of transsplicings. Spliced leader RNA transsplicing is a mechanism of gene expression and involves spliced leader sequences (SL) in RNA processing (Davis, 2012). The conserved sequence of SL, ranging in size from 15-51 nt, is added to the 5’ end of mRNAs (Davis, 2012; Bonen, 1993). The mRNA is having a new cap (see Figure 2.11).

This kind of RNA maturation was also found in other kinetoplastida, euglena, nematodes, flatworms, ctenophores, hexactinellida sponges, chaetognaths, crustaceans, cnidarians, and rotifers, as well as tunicates, and primitive chordates (Davis, 2012; Bonen, 1993). It is suggested that the presence and apparent independent evolution of trans-splicing in many invertebrate phyla and early chordates represents an evolutionarily important form of gene expression (Davis, 2012). The mRNA metabolism that comprises mRNA processing, export, translation, and stability is fully dependent on the mRNA cap and 5’ UTR of mRNAs (Yeats, 2009; Davis, 2012).
Biological Background

**Figure 2.11: Spliced Leader (SL) Trans-Splicing.** In this Figure, the SL RNA and pre-mRNA is ligated into mature mRNA by the trans-splicing mechanism. Figure taken from (Davis, 2012), Fig.1.

**Intron Length and Abundance** The abundance of introns in a genome could be reflected by the average numbers of introns per gene. A genome is intron-poor when the average number of introns per gene is \( \leq 0.1 \) (Irimia et al., 2007). It could be seen in the Table 2.1, that several organisms are having the shortage of introns, i.e. *Giardia lamblia*, *Trichomonas vaginalis*, *Trypanosoma brucei*, and *Leishmania major*. Usually, the lack of intron is a feature of prokaryotic organism (Adam, 2001).

In *Giardia lamblia* there is only 4 introns detected that’s incorporated in 2 genes. One gene is responsible for generating mature mRNAs that encodes heat shock protein 90 which controlling the cellular proteins. The other one is encoding a dynein molecular motor protein, involved in the motility of the flagella (Kamikawa et al., 2011). The mature mRNAs of these two genes were formed from the transsplicing. *Trypanosoma* has rare occurrences of intron while most of its mRNAs are transspliced by addition to the 5’ end of the spliced leader (SL) sequence (Mair et al., 2000). Introns in only two genes of *Trypanosoma* were identified. These genes encode for poly(A) polymerase and DNA/RNA helicase enzymes. It was found no additional genes with introns (Siegel et al., 2010). There are no normal introns at all in *Leishmania*, and it relies solely on transsplicing for mRNAs processing (Kazemi, 2011). The scarcity of intron make the median length count not feasible in the statistical point of view.
Table 2.1: The Intron statistics. Each species with its different genome assembly are having different number of intron per gene and median intron length value. Data for the intron per gene is taken from [http://www.ncbi.nlm.nih.gov/mapview/](http://www.ncbi.nlm.nih.gov/mapview/). Based on information from (Fritz-Laylin and et al., 2010), the *Giardia lamblia* and *Trypanosoma brucei* median intron are still undetermined. The median intron length data by species and citation: *Naegleria gruberi* (Fritz-Laylin and et al., 2010), *Thalassiosira pseudonana* (Norden-Krichmar et al., 2011), *Tetrahymena* (Salim et al., 2011), *Phytophthora* (Haas and et al., 2009), *Clamydomonas* (Labadorf et al., 2010), *Oryza sativa* (Ner-Gaon et al., 2007), *Arabidopsis thaliana* (Hong et al., 2006), *Schizosaccaromyces pombe* (Zhang and Marr, 1994), *Caenoharbditis elegans* (Stein et al., 2003), *Drosophila* (Sela et al., 2010), human (Lander and et al, 2001). The species with '*' marks are having coding-introns. 'ND' refers to 'not determined'. 'NDA' refers to 'no data available'. $\delta$ refers to the newly computed median intron length from RefSeq annotation.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome Assembly</th>
<th>Intron per gene</th>
<th>median intron length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Giardia lamblia</em></td>
<td>WBC6</td>
<td>0.00137</td>
<td>ND</td>
</tr>
<tr>
<td><em>Trichomonas vaginalis</em></td>
<td>TrichDB-1.2</td>
<td>0.00059</td>
<td>NDA</td>
</tr>
<tr>
<td><em>Trypanosoma brucei</em></td>
<td>Tb927_May08_v4</td>
<td>0.00098</td>
<td>ND</td>
</tr>
<tr>
<td><em>Leishmania major</em></td>
<td>Lmj_20070731_V5.2</td>
<td>0.00044</td>
<td>NDA</td>
</tr>
<tr>
<td><em>Naegleria gruberi</em></td>
<td>Naegr1</td>
<td>0.70000</td>
<td>60</td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>PlasmoDB-7.0</td>
<td>1.34053</td>
<td>143 $\delta$</td>
</tr>
<tr>
<td><em>Tetrahymena</em></td>
<td>tta1_oct2008</td>
<td>3.72720</td>
<td>86</td>
</tr>
<tr>
<td><em>Thalassiosira pseudonana</em></td>
<td>Thaps3</td>
<td>1.42301</td>
<td>90</td>
</tr>
<tr>
<td><em>Phytophthora ramorum</em></td>
<td>Phyra1_1</td>
<td>1.58140</td>
<td>74</td>
</tr>
<tr>
<td><em>Clamydomonas</em></td>
<td>Chlre4</td>
<td>7.34670</td>
<td>232</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>OSV6.2</td>
<td>1.70526</td>
<td>139</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>TAIR9.55</td>
<td>2.04926</td>
<td>98</td>
</tr>
<tr>
<td><em>Dictyostelium</em></td>
<td>DDB</td>
<td>1.35402</td>
<td>105 $\delta$</td>
</tr>
<tr>
<td><em>Aspergillus niger</em></td>
<td>CADRE</td>
<td>3.36514</td>
<td>62</td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>EF1</td>
<td>1.01762</td>
<td>63</td>
</tr>
<tr>
<td><em>Caenoharbditis elegans</em></td>
<td>WS200</td>
<td>5.11277</td>
<td>66</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>BDGP5.13</td>
<td>3.84178</td>
<td>70</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>hg19</td>
<td>5.49877</td>
<td>1023</td>
</tr>
</tbody>
</table>
2.2 Protein Domain: Structure and Functions

The range of an organism’s biochemical capabilities, both metabolic and regulatory, is largely encoded in its protein content. Proteins embody a wide variety of functions in a cell, ranging from enzymatic activity to structural scaffolding (Cortese et al., 2008). This is true even though RNA-based mechanisms can play a fundamental role as in the case of post-transcriptional regulation by microRNAs (Mattick and Makumin, 2006). In fact, the presence or absence of RNAi pathways, for instance, can be inferred from the presence or absence of its protein components (Drinnenberg et al., 2009). Large-scale trends in evolution such as an increased complexity of transcriptional regulation (Melzer, 2011; Shelest, 2008), the diversification of chromatin modification (Prohaska et al., 2010), or novel modes of post-transcriptional processing are visible in comparisons of the predicted protein complements and thus are focal features of most genome papers (Zanivan et al., 2007; Parikesit et al., 2011c).

Proteins have a diverse range of functional groups, e.g. alcohols, thiols, and thioethers. However, the broad spectrum of protein function arises from the combination of these groups in various sequences (Berg et al., 2002).

Protein structure is very important for its functional feature, and specific structure has to be obtained for it (Yang et al., 2005a; Ochiai, 2008). Several levels of the protein structure can be assigned to uphold its functional features (Colafranceschi et al., 2008; Ochiai, 2008). In the following, the different level are explained using spinach ferredoxin as an example (Ochiai, 2008):

- **Primary structure**: In the Figure 2.12A, the *primary* structure, as the basic level, is the single chain of the polypeptide (Helmenstine, 2012; Ochiai, 2008). It is the bars connecting all amino acid sequence.

- **Secondary structure**: It is the hydrogen bonds patterns between amide and carboxyl groups backbone in the polypeptide. The coiled ribbons is the shape of α-helices, while planar tapes was a representation of β-strands (Figure 2.12B).

- **Tertiary structure**: It is the three-dimensional structure of a single polypeptide chain, and it comprises of different combinations of the secondary structures (α helices, β strands, and loops) (Figure 2.12C).

- **Quaternary structure** is the multiple folded or coiled protein arrangement in a multi-subunit complex. (Ochiai, 2008) (Figure 2.12D).
2.2 Protein Domain: Structure and Functions

Most proteins are composed of smaller building blocks. A protein domain typically forms a compact three-dimensional structure that is frequently stable and foldable on its own and conveys a specific molecular function such as a particular catalytic activity or binding specificity. Protein domains are characterized by local amino-acid patterns and hence can be annotated computationally in protein sequences (Capozzi et al., 2007; Parikesit et al., 2011c).

Protein could be consisted of more than one structural domains, and one domain could emerged in a diverse kind of proteins (Alberts et al., 2002). The length of domains could be varied from between 25 amino acids until up to 500 amino acids, and it could formed a functional units (Brocchieri and Karlin, 2005).

The distribution of domains within proteins is not completely uniform. For instance, about fifty specific domains are preferentially found in alternatively spliced exons and hence...
Biological Background

systematically lacking in some protein variants (Resch et al., 2004). As most proteins contain more than a single domain, domain combinations are of particular interest when aiming at a more detailed understanding of the novel functions (Yang and Bourne, 2009).

Protein domains can be recombined in a combinatorial fashion to produce new functionalities over large time-scales. Individual proteins often have multiple ancestors that contributed with different domains to an extant protein (Moore et al., 2008; Koonin et al., 2000). From an evolutionary perspective, furthermore, they form quite well-defined and stable units of selection (Parikesit et al., 2011c; Pereira-Leal et al., 2006). As an alternative to reconstructing protein evolution, one thus may focus on tracing the distribution of individual domains (Ponting and Russell, 2002; Caetano-Anolles and Caetano-Anolles, 2003; Jin et al., 2009; Yang and Bourne, 2009). In a recent study of chromatin evolution, we demonstrated that it is indeed feasible to determine large-scale trends in regulatory capabilities based on domain content (Prohaska et al., 2010). Interestingly, domains differ in their intrinsic propensity to co-occur with many different other domains. This versatility, however, is primarily dependent upon the position of the domain at the end of proteins and their occurrence in single domain proteins. This can be explained by fusions and fissions as the most frequent genomic operations creating novel domain combinations (Weiner et al., 2008). This is an ongoing evolutionary process. On the other hand, some "promiscuous" domains, in particular those involved in protein-protein interactions, have a propensity to appear in particularly wide variety of different domain architectures (Basu et al., 2008). For instance, there are many animal-specific or even vertebrate-specific domain-combinations (Itoh et al., 2007).

More global trends can be uncovered by considering aggregate statistics of domains and domain combinations. The average number of domains in a protein, for instance, increases systematically along the human lineage (Yang and Bourne, 2009). Network analysis of domain co-occurrences, furthermore, demonstrates a growing core of combinations in multicellular organisms (Wuchty and Almaas, 2005).

2.3 Gene Regulation

The expression of genomically encoded information is subject to tight regulation and control in all organisms that have been studied in detail (Panis et al., 2010). These regulatory rules are implemented in a highly complex network of several biochemically distinct mechanism that act at multiple levels of the gene expression cascade (Hao et al., 2012). They include specific chromatin states, the action of transcription factors, regulated mRNA export, alternative splicing, translational control, post-transcriptional and post-translational modifications, and controlled degradation of both RNA and polypeptides (Lackner, 2007).

Surprisingly, it appears that different phylogenetic clades emphasize certain types
of mechanisms while reducing or even abolishing others. Regulation in eubacteria, for example, appears to be dominated by transcription factors networks, trypanosomes use the post-transcriptional processing of large polycistronic transcripts, ciliates utilize extensive amplification of DNA in creating their macro-nuclei, and crown group eukaryotes have evolved an elaborates system of histone modifications (Cavalier-Smith, 2010). An understanding of the diversity of life thus requires the investigation of the origin(s) and evolution of these different regulatory mechanisms and their interplay (Cremer and Cremer, 2001; Ptashne, 1986; Thieffry et al., 1998).

2.3.1 Transcription Factor (TF)

Transcription factors (TF) typically cooperate to activate or repress the expression of target genes (Ouyang et al., 2009). They play critical roles in essentially every developmental process, from the proliferation and differentiation of stem cells to the maintenance of differentiated cells in adult organisms (Tam and Lim, 2008; Parikesit et al., 2011a). TF have many important function, that preserve the integrity of the cell, and ensure its growth (Phillips and Hoopes, 2008). Those are, namely, cell signaling, RNA splicing, siRNA control mechanisms, and chromatin modifications (Phillips and Hoopes, 2008). Figure 2.13A shows the specific regulatory properties of TFs.

**Zinc Finger**  A zinc finger is a large superfamily of protein domains that can bind to DNA, e.g the three tandem 3-bp subsites (Bulyk et al., 2001). They are interaction modules that bind DNA, RNA, proteins, or small molecules. A zinc finger consists of two antiparallel $\beta$ strands, and an $\alpha$ helix. The zinc ion is crucial for the stability of this domain type. Zinc finger is a part of transcription factor regulation domains (Bateman and et al, 2008; Parikesit et al., 2010b). Figure 2.13B is the zinc finger protein domain.

**Wing Helix**  Winged-helix is a DNA-binding domain which binds to specific DNA sequences, e.g the major groove of a duplex oligonucleotide derived from the transthyretin gene promoter (Gajiwala and Burley, 2000). Consisting of about 110 amino acids, the winged-helix domain has four helices and a two-strand beta-sheet. Wing-helix is a part of transcription factor regulation domains (EBI, 2009; Parikesit et al., 2010b). Figure 2.13C is the Wing Helix domain.

2.3.2 Chromatin Regulation (CR)

DNA in a cell was surrounded and associated with the the chromosomal architectural proteins (ChAPs) (Prohaska et al., 2010). These proteins plays important role to maintain the integrity of the DNA structure (Wang, 2006; Prohaska et al., 2010). Chromatin is the union of the ChAPs and DNA, and it has various contents in different organism (Mansperger, 2007;
Biological Background

A. Transcription Factor Mechanism

Figure 2.13: (A) The binding of transcription-factor binding sites (TFBS) to the TF. Figure taken from (Wasserman and Sandelin, 2004), Fig.1. (B) Zinc fingers are a large superfamily of protein domains that can bind to DNA (and occasionally single- or double-stranded RNA and proteins). Figure taken from (Bateman and et al, 2008), http://pfam.sanger.ac.uk/family/PF00096. (C) Wing Helix is another DNA-binding domain which binds to specific DNA sequences. Figure taken from (EBI, 2009), Fig.321c

Prohaska et al., 2010). The function of Chromatin regulation (CR) is to ensures efficient DNA replication, segregation and gene expression (Prohaska et al., 2010).

Post-translational histone modifications are the regulators of the chromatin structure including, for example, acetylation, methylation, and ubiquitinylation (Ram et al., 2011).

Histone Modification Domains  It was argued, that the mechanism of histone modification is following a set of rules (Wang et al., 2007). The ‘writer’ domains responsible to write specific histone modification, while their removal operation was conducted by the ‘eraser’ domains (Wang et al., 2007). The modifications were recognized by the ‘readers’ domain (Wang et al., 2007). In most of the time, the writers, readers, and erasers domain are associated with DNA binding domains (Prohaska et al., 2010).
2.3 Gene Regulation

**Writer domain: SET** SET is an abbreviation for *Drosophila’s* Su(var)3-9, ‘Enhancer of Zeste’ proteins, and Trithorax protein. SET is a conserved region that was found first in *Drosophila* Trithorax protein, and it is also identified in *Drosophila* Su(var)3-9 and ‘Enhancer of Zeste’ proteins (EMBL, 2011). Beside those three proteins, SET domain also available in others as well. The SET domains has catalytic activity as lysine methyltransferase (PFAM, 2012b). The protein-protein interaction was the functionality feature of SET domains (PFAM, 2012b).

**Eraser domain: Sir2** Histone deacetylation or mono-ribosyltransfer are the catalytic features of Sir2 domain (PFAM, 2012a). It also involved in the important metabolic pathways of the three domains of life (PFAM, 2012a). Moreover, the dissociation of the *Plasmodium falciparum’s* telomere cluster is also one of Sir2’s features (Mancio-Silva *et al.*, 2008).

**Reader domain: SANT** SANT domain has 50 amino acid motif, and it present in the nuclear receptor co-repressors. Recent studies suggested that SANT domains might be a histone-tail-binding module and also take part in the chromatin-remodelling complexes. SANT domains play part in histone acetylation (Boyer *et al.*, 2004).

**Correlation of Transcription Factor and Chromatin Regulation** Transcriptional activity and histone acetylation are associated in eukaryotic cells (Allfrey *et al.*, 1964; Pogo *et al.*, 1966). Transcriptionally active chromatin is associated with histones acetylated core and histones tails lysine residues undergo acetylation. That will cause the tails charge would be neutralized and their DNA affinity would be decreased (Sealy and Chalkley, 1978; Vidali *et al.*, 1978; Hebbes *et al.*, 1988; Hong *et al.*, 1993). This would cause the alteration of nucleosomal conformation that could augmented the transcription regulatory proteins toward the chromatin templates (Norton *et al.*, 1989; Lee *et al.*, 1993; Vettese-Dadey *et al.*, 1996). These phenomenon exposes that increased transcriptional activity in vivo was resulted from the histone acetylation. However, the relationship between transcriptional activity and histone acetylation or about their molecular mechanism is still unclear (Struhl, 1998).

### 2.3.3 Translation

The important step for preventing or stopping the translation is the utilization of mRNA repressor, which is involves specific proteins. The regulator genes encoded these repressor proteins for attaching themselves into the certain DNA segment known as the operator. The binding is preventing the RNA polymerase for generating mRNA (Gilbert, 2000a; Ibba and Soll, 2000). The mRNA degradation is the important step for the regulation and levels of gene expression, and it is determined by the regulation its half life (Beelman C and Parker, 2020).
Biological Background

1995; Wagner and Lykke-Andersen, 2002; Valencia-Sanchez et al., 2006; Garneau et al., 2007). This process is controlled by endo and exonucleases enzymes (Beelman C and Parker, 1995).

2.3.4 Focusing on Transcription Factor and Chromatin Regulation

The reason of focusing towards transcription factor and chromatin regulation, is that it is interested to examine the reciprocal effects of the unclear interactions between both gene regulators. The determination of the domain contents of both transcription factor and chromatin regulation protein domain will be envisaged. It is hoped that by elucidating the domain content, stronger the biological conclusion on the mechanistic insight of transcription factor and chromatin regulation could be obtained. It is also expected that the inter-relationship of chromatin regulation and transcription factor could be comprehended.
This dissertation uses gene prediction and domain annotation package extensively, and the related terms about them will be explained in this chapter. For example, the Markov chain, hidden markov models, and the Viterbi algorithm that underlies the methodology of gene predictors and domain annotation are explained in details. Additionally, it will be exposed how the protein domain works as a evolutionary unit will be exposed. Lastly, There will be more details about the domain annotations and the functional annotation.

3.1 Markov Chains, Hidden Markov Models, and the Viterbi Algorithm

This section is explaining about the interconnected terms of Markov chains, the basic principle of the Markov models, the basic and advanced application of Markov models, and about Viterbi Algorithm, which is critically important for the whole method. Speech recognition is the most widely discussed HMM research topic since its introduction in the early 1970s (Durbin et al., 1998). The CpG island is taken as an example, because it is an important indication for the start of the gene, and this is why the identification of CpG islands could pinpoint the gene location (Mneimneh, 2010).

The given example for explaining this section is the case of CpG Island (Durbin et al., 1998). It is a stretch of DNA sequence that is densely populated with C and G bases (Glass et al., 2007). There is a tendency of the C bases to be methylated, and then converted into T (Durbin et al., 1998).
Technical Background

3.1.1 Markov Chain

The Markov chain is the simplest probabilistic model to determine existence of the CpG Island region in a certain short sequence (Huson, 2008). Figure 3.1 is a Markov chain for the DNA sequences that exist in a CpG Island (Durbin et al., 1998). The state of the DNA sequences are in the each A,C, G, and T alphabets (Chang et al., 2005). Moreover, there is a probability parameter, which is assigned to each arrow that determine the probability of one state following the others (Durbin et al., 1998). In this respect, the probabilistic equation is written in Equation 3.1. The state sequence of the path $x$, the $i$th state of the path is called $x_{i-1}$.

$$a_{st}=P(x_i = t | x_{i-1} = s)$$ (3.1)

Figure 3.1: The Markov chains as a collection of 'states', and the arrows between the states connecting each particular residue. Figure taken from (Durbin et al., 1998), Fig.3.1.

3.1.2 Hidden Markov Models

However, simulating an island of CpGs in the middle of a 'non-island' could not be feasibly done in that simple model (Durbin et al., 1998). It is because the simple Markov Chains did not covers both available states (Konstantopoulos, 2009). The construction of a model that cover the whole states should be conducted (Durbin et al., 1998). Figure 3.2 shows the complete scheme of Hidden Markov Models (HMM), which are utilized for larger sequence.
3.1 Markov Chains, Hidden Markov Models, and the Viterbi Algorithm

The explanation of the model is as follows. First, each consecutive sequences is assigned with ‘-’ in subscript, when referring to the emission of CpG Island region, and in the other hand, each consecutive sequences would be assigned with ‘+’ in subscript when referring to the non-island region. The formal definition of a HMM in Equation 3.3 follow the Markov model in the previous section. The state sequence of the path \( \pi \), the \( i \)th state of the path is called \( \pi_{i-1} \).

\[
a_{kl} = P(\pi_i = l | \pi_{i-1} = k) \tag{3.2}
\]

Figure 3.2: The HMM model of CpG islands. The transitions complete set is shown in each set, just like the simpler markov chain. Figure taken from (Durbin et al., 1998), Fig.3.3.

Moreover, due to the decoupling of the symbol \( b \) from state \( k \), the new parameter of \( e_k(b) \) was introduced as the emission probability in Equation 3.3

\[
e_k(b) = P(x_i = b | \pi_i = k) \tag{3.3}
\]

3.1.3 The Viterbi Algorithm

The Viterbi Algorithm is a dynamic programming algorithm that is used to decode the observing sequence (Durbin et al., 1998). The existence of a CpG Island would be told by the HMM predicted path, based upon assumption that the CpG Island and other regions are assigned to the each state (Churchill, 2005). The assumption that either CpG islands or other regions were assigned to each state of the predicted path in HMM will pave the way for the predicted sequence (Panchin et al., 2011). The highest probability in one path should be chosen for the prediction, following Equation 3.4 (Durbin et al., 1998).
Technical Background

\[ \pi^* = \arg\max_\pi P(x, \pi) \] (3.4)

The \( \pi^* \) can be found as the most probable path recursively. Suppose the probability \( v_k(i) \) of the most probable path ending in state \( k \) with observation \( i \) is known for all the states \( k \). The observation \( x_{i+1} \) could be calculated for its probability as follows:

\[ v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl}) \] (3.5)

The initial condition \( v_0 = 1 \) would start all sequences in state 0 (the begin state). Backtracing could be utilized for finding the actual sequence state, when the pointers is kept backwards.

3.2 Gene Prediction

In what follows, the difference of the gene prediction methods will be exposed, as well as their problems. The explanation of the different kind of gene prediction mechanism will be explained. The problem of gene prediction will be described. Gene prediction is the early step of the re-annotation process.

3.2.1 Mechanism of Gene Prediction

**Homology-Based** The homology of sequence identity between the genomic sequence query with the database entry of the existing sequence may indicate that the sequence contain certain gene (Yao et al., 2005). The homology-based method is based on simple idea: genes and regulatory sites as evolutionary functional elements are inclined to be more conserved than its non-functional counterparts. Thus, biological function could be present due to local sequence similarity (Taher et al., 2003). There are several way to employ this method:

- Aligning the known sequence to the genome as the gene prediction.
- Completing the evidence from the guide of the known sequence for completing the gene structure.

The number of known gene sequence is determining the efficacy of those methods, hence its limitation is the completion of biological databases (Blanco, 2011). The homology based predictors that use query alignments toward the genome are also known as *de novo* gene predictor (Bernal, 2008).

**Ab initio** The *ab initio* gene prediction is using computational and statistical methods to detect certain properties of the gene, such as protein coding sequence, spice sites, start and stop codon (Yao et al., 2005). This method is independent of sequence similarity with an existing database (Yao et al., 2005). This dissertation is using the *Ab initio* method.
3.2 Gene Prediction

*Ab initio* gene predictors are defining the parameters of experimental-based real genes. Those are the models of splice donor sequence, splice acceptor sequence, intron and exon length distribution, open reading frame requirement in coding exons, requirement that introns maintain reading frame, and the start and stop of transcription (Borenstein, 2011). Those parameters are utilized to obtain the best prediction of genes from existing genomes.

*Ab initio* method depends on two important statistical parameters, namely sensitivity and specificity. Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function. Sensitivity measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of the predicted gene compared with actual gene). Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of the unpredicted gene compared with the actual gene).

\[
\text{Sensitivity} = \frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{Number of False Negative}} = \text{Probability of positive test} \tag{3.6}
\]

\[
\text{Specificity} = \frac{\text{Number of True Negative}}{\text{Number of True Negative} + \text{Number of True Positive}} = \text{Probability of negative test} \tag{3.7}
\]

Care must be taken when adjusting the both parameter. In one hand, when the sensitivity is high, then the specificity would be low, and it means that more false positive on the prediction engine. On the other hand, when the specificity is high, then the sensitivity would be low, and it means that more false negative results would be produced.

### 3.2.2 Gene Prediction Packages for this Research

**GENSCAN.** Definition of essential terms in gene finding is necessary for its utilization (Burge and Karlin, 1998). The successful gene prediction is measured by it 'sensitivity', and 'specificity' (Burge and Karlin, 1998). Moreover, 'sensitivity' is defined as the correctly predicted 'true' sites, for example the sites of exons and donor splicing (Burge and Karlin, 1998). While 'specificity' is defined as correct 'predicted' site proportion (Burge and Karlin, 1998). Adjusting the sensitivity and specificity into a simultant high level could led to an efficient and effective gene prediction (Burge and Karlin, 1998). Four type of exons are important for gene finding (Burge and Karlin, 1998): initial exons, internal exons, terminal exons, and single-exon genes (Burge and Karlin, 1998). The existance of those four exon types pose a challenge for each gene-finding package (Burge and Karlin, 1998). **GENSCAN** used HMMs for predicting genes (Burge and Karlin, 1998).
**Technical Background**

**GENEMARK.** The protein-coding potential of a DNA sequence (within a sliding window) is determined by GENEMARK by utilizing species-specific parameters of the Markov models of coding and non-coding regions (Besemer and Borodovsky, 2005). Local variations of coding potential was delineated by this approach (Lukashin and Borodovsky, 1998). The protein-coding potential distribution along a sequence is shown by the details of the GENEMARK annotation (Besemer and Borodovsky, 2005; Lukashin and Borodovsky, 1998). Genes and intergenic regions in a sequence are predicted by GENEMARK as a whole (Lukashin and Borodovsky, 1998). The 'grammar' of gene organization was reflected by the utilized HMM models (Besemer and Borodovsky, 2005). The maximum likely parse of the whole DNA sequence into protein coding genes (with possible introns) and intergenic regions were identified with GENEMARK program (Besemer and Borodovsky, 2005; Lukashin and Borodovsky, 1998).

**AUGUSTUS.** Important biological signals are mathematically modeled by gene prediction programs (Stormo, 2000; Stanke and Waack, 2003). Those modeled signals are, for examples, splice sites or the translation start and end points (Stormo, 2000; Stanke and Waack, 2003). It is the AUGUSTUS program, that using trained biological signals based upon the set of known gene structure (Koonin and Galperin, 2003; Stanke and Waack, 2003).

The clear advantage of AUGUSTUS package are they could be employed toward the sequence with unknown homologies, or even non-existent homologies (Stanke and Waack, 2003). However, the evaluation of the AUGUSTUS package efficiency and effectivity is most of the time based upon one gene with short sequences, and almost nothing of flanking DNA (Stanke and Waack, 2003). AUGUSTUS is using an HMM for predicting genes (Stanke and Waack, 2003). New method on the accuracy of intron lengths modeling was a feature of AUGUSTUS that could be utilized in other HMM based gene prediction package (Stanke and Waack, 2003). Length distribution clustering of short introns typically have a certain length (Stanke and Waack, 2003; Stanke et al., 2004). The length distribution of short introns is modeled accurately by AUGUSTUS and a geometric distribution was applied for the lengths of long introns (Stanke and Waack, 2003; Stanke et al., 2006). The models of the splice site have been simplified as a probabilistic model of empirical distribution was used. A method to train the model parameter dependent on GC-content of the input sequence was applied as well (Stanke and Waack, 2003; Stanke et al., 2008).

**Differences between GENSCAN, GENEMARK, and AUGUSTUS.** GENSCAN is an untrained gene predictor that use identical gene models for predicting genes. GENEMARK and AUGUSTUS are trained gene predictors that could be trained for specific species. Trained gene predictors could be useful for predicting genes in a newly sequenced genome.

However, among three of them, AUGUSTUS is definitely the gene predictor with the most complete functionality. AUGUSTUS can train cDNA, EST, Protein, and DNA/Genes data;
while GENEMARK is limited to DNA/genes only.

### 3.2.3 Problems of Gene Prediction

The homology-based method suffers from the dependence on the existing sequence database (Sharma et al., 2012). Only limited amount of genes could be found by comparing with other genomes (Yao et al., 2005). The existence of artifacts in the existing database also disrupt the accuracy of this method (Reese et al., 2000).

Ab initio methods are less accurate than the similarity method because of the limitation of the training sets (Yao et al., 2005). However, hopes are rising up, because more training sets are now available (Bonneau and Baker, 2001).

**Gene Structure in Specific Organism.** Species-specific gene prediction problems are present as well. Several major lineages of the Eukarya have gene structures and a genomic organization that is very different from the situation in animals, fungi, or plants. For example, both Giardia lamblia and Trichomonas vaginalis are extremely intron-poor; Trichomonas vaginalis in addition features very large numbers of paralogs. Kinetoplastids (Trypanosoma and Leishmania) produce large polycistronic transcripts from which individual mature mRNAs are produced by trans-splicing, cis-splicing, and polyadenylation, and these phenomenons are explained in detail in Section 2.1 (Michaeli, 2011; Thomas et al., 2009). Trans-splicing is also prevalent in the nematodes, but absent from most other animal genomes. Intron-sizes differ dramatically between invertebrates and vertebrates, where intron-sizes of more than 10 kb are not at all uncommon. Another problem is posed by the extreme sequence composition as in the AT-rich genome of Plasmodium falciparum, and this could be problem a for gene predictors due to the utilization of GC-content specific signal models (Lu et al., 2007; Zhang, 1998).

### 3.3 Visualization and Scheme of Protein Domain

A protein domain is an independent functional and evolutionary unit. Figure 3.5 shows that a protein domain can conduct chemical binding and interaction with other chemical substances.

Figure 3.3 shows HMG-box and WD40 transcription factor domains. HMG-box domain is involved in the regulation of DNA-dependent processes, such as transcription, replication, and DNA repair. WD40 is involved in variety of functions ranging from signal transduction and transcription regulation to cell cycle control, autophagy and apoptosis.

Proteins are composed of recognizable protein domains that can be re-combined in a combinatorial fashion to produce new functionalities over large time-scales. Figure 3.4 shows the PcrA Helicase proteins, which consist of four distinct domains: domain 1A, 1B, and two domains 2B.
Technical Background

Figure 3.3: (a) HMG-box and (b) WD40 Domains. Both HMG-box and WD40 are transcription factor domains. Figure taken from http://pfam.sanger.ac.uk/family/PF00505 and http://pfam.sanger.ac.uk/family/WD40.

Figure 3.4: Scheme of PcrA Helicase Protein and its domain. This protein is constituted from more than one domain. Each domain is signified by different colors. Figure taken from (snf2.net, 2005), Fig.2.
3.4 Protein Domain Annotations

The interesting point about protein domain is its recombination feature with the other domains to obtain new functionality. For example, the classical Rossman domains are combined from seven different domains (Bashont and Chothia, 2002). Figure 3.6 shows how the domain combination can be done.

3.4 Protein Domain Annotations

The frequency of insertion/deletions and the conserved residues can indicate the multiple sequence alignments of protein (Zhang et al., 2010; Eddy, 1996, 1998). In this respect, a profile, which defined as "a consensus primary structure model consisting of position-specific residue scores and insertion or deletion penalties", could be feasibly utilized for protein domain annotation (Lambert et al., 2003; Eddy, 1996). Number of research groups have been successfully developed Profile based method that derived on 3D structures and multiple sequence alignment, and these method is already utilized widely (Eddy, 1996). However, the complicated models with many free parameters pose a problem with the profile method (Eddy, 1996, 1998). The position-specific residue scoring, scoring gaps and indels, and the combination of structural and sequence information are the main problems that could not solved until recently (de Blasio, 2007; Eddy, 1996). Thus, profile methods using HMMs may
Technical Background

Figure 3.6: Scheme of domain combination. The 'blue' box is the DNA-binding domain which localizes to DNA site. The 'red' box is the recruiting RNA polymerase domain. When the 'blue' and 'red' domains are combining each other, they obtain a new function for selective gene transcription.

Protein domains and their combinations contain information about the functions in a cell.

be satisfactorily answers to those problems (Eddy, 1996).

SUPERFAMILY The protein domains with known structure are annotated in the SUPERFAMILY database (Wilson et al., 2007, 2009). The expert-curated HMM periodically updates the database, and it exemplifies the known protein domains with their structural features (de Lima Morais et al., 2011). The SCOP protein database was the classification source for these domain annotation (de Lima Morais et al., 2011; Andreeva et al., 2004). The hierarchy of SCOP database are classified into four levels, namely Class, fold, superfamily and family (de Lima Morais et al., 2011). It was classified based upon its similarity nature, in the level of sequence, evolutionary, and structural (de Lima Morais et al., 2011; Andreeva et al., 2008). The superfamily level is definitely the main focus of the SUPERFAMILY database (de Lima Morais et al., 2011). The evidence of a common ancestor from its structure and sequence proof are the basis to put two domains into the same superfamily classification (de Lima Morais et al., 2011; Gough, 2002). The complete domain annotation could be found in the SUPERFAMILY website, that present a bulk of proteins and superfAMILY analyzing methods, and definitely with its search feature (Gough and et al., 2002) (de Lima Morais et al., 2011). Potential investigation that could be done by the users are the representation of the superfamily, phylogenetic examination, protein domain networks and architectures, and complete phylogenetic distribution of the superfamilies in the whole kingdom of life.
(de Lima Morais et al., 2011; Vogel, 2010). The procedure 'T99' of SAM package was employed for building up the SUPERFAMILY database (Karplus and Hu, 2001). The model building was generated by multiple alignments of homologous sequences (Gough, 2002). A seed, which could be a different kind of sequences, whether they are aligned or not at all, are given to the pipeline (Gough, 2002; Gough and Chothia, 2002). Massive amount of sequences for iterating the extracted homologues were utilized to sum up the expanding alignment (Gough, 2002).

Method

These steps are the default T99 procedure, following (Gough, 2002; Karplus and Hu, 2001) (see also Figure 3.7):

1. **WU-BLASTP** ([http://blast.wustl.edu/blast](http://blast.wustl.edu/blast)) was utilized by the initial sequence(s), and two sequences sets were find based upon a search into the massive non-redundant protein database. The query sequence close homologues are the first set, while the second one are their distant homologues.

2. The seed sequence was aligned with its close homologues in the non-redundant database (NRDB).

3. The step 2 produced sequence aligments, which could be utilized for building initial an HMM

4. Out of the step 1, this model should find for the second homologues step, in order to be re-aligned with it, and generating a much better new alignment.

5. The iterations of steps 2,3, and 4 produces final alignment, that are sufficient to build a model. The SAM package has tools for this necessity (Karplus and Hu, 2001).

**NRDB90** is a massive non-redundant sequence database that is employed by the procedure of 'T99' (Holm and Sander, 1998). It consists of with almost 400000 sequences, although in the beginning it had only around 200000, and they are having less than 90 % similarity in this database (Gough, 2002).

**PFAM**  

PFAM enjoys its status as the primary reference of domain annotation that is derived from the homology of the sequence alignment (Punta et al., 2012). The significant sequence similarity are the classification based for protein families, and its homology are annotated in PFAM database (Punta et al., 2012; Bateman et al., 2004). The **HMMER3** programs package was used for detecting the similarity (Punta et al., 2012). Two types of families are the cornerstone of PFAM: Pfam-A, that resulted from manual curation, and Pfam-B, which derived from the automatic method (Punta et al., 2012; Finn et al., 2006).

These following steps are the pipeline to build Pfam-A families, following (Punta et al., 2012) (see figure 3.8):
1. Seed alignment was generated from the excellent quality of multiple sequence alignment.

2. A seed alignment was employed for generating a HMM profile.

3. The UniProtKB sequence database was searched against the HMM profile.

4. The family inclusion of the sequence regions are based upon the accepted gathering thresholds (GAs).

Matches towards the NCBI non-redundant database was also given by PFAM (Punta et al., 2012). They are, among others, HMM based family conservation, domain architectures description, the co-occurrences with others, and the taxonomic range of the domain (Punta et al., 2012).

The search conducted by the HMM profile, whether internal or external one, are primarily influenced by the quality of the seed alignment (Wistrand and Sonnhammer, 2005; Punta et al., 2012). In this end, the seed alignment is the center of the curatorial effort (Punta et al., 2012). Some functional aspect similarity are expected from the PFAM family members that have a common history of evolution (Enright, 2002). The representation of the functional units should be available in the PFAM families, and its combination could add up protein with a distinct functional feature (Punta et al., 2012). PFAM database ultimate goal is annotating the families of the protein sequences as much as possible, for the purpose of aiding genomic research (Shachar and Linial, 2004; Punta et al., 2012). However, homology is not the same with functional similarity, and vigilance must be taken upon using functional annotation on family membership (Götz et al., 2008; Kim and Caetano-Anollés, 2012; Punta et al., 2012).
3.4.1 Differences between SUPERFAMILY and PFAM

The SUPERFAMILY families tend to be less specific than PFAM due of picking up more distant evolutionary relationship of HMMs (Williams, 2006). It implies that often one SUPERFAMILY family corresponds to more than one PFAM family. The alignment of PFAM is independent of its structural feature, while SUPERFAMILY must be based on the known structure of the protein. Therefore, SUPERFAMILY can identify a domain based on a known structure (Williams, 2006).

3.5 GO Annotations

The Gene Ontology (GO), is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species (Consortium, 2008). The information integration of gene products and standardization of sequences classifications, are the main features of The GO database (Consortium, 2004). The three model organism databases, namely FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI) project, are the pioneer of the GO annotation efforts (Consortium, 2004). The GO consortium has included many databases repository from different kind of genomes project (Consortium, 2004).
3.5.1 Ontologies

The provided ontology in GO database has unique definition and identifiers, that arranged in ‘is-a’ and ‘part-of’ classifiers (Consortium, 2004; Blair, 2010). The needs of sharing reliable ontologies to the scientific community is high, due to the extensive genomic research and its data integration (Consortium, 2004). The annotation features of Enzyme Commission (EC) database is deemed as insufficient by bioinformaticians, due of its limitation for describing the gene and/or protein functional features (The Gene Ontology Consortium, 2000; Consortium, 2004). The specific EC system for enzymatic chemical reaction doesn’t apply to the non-enzymatic proteins, and its dynamic interaction within (The Gene Ontology Consortium, 2000). The confusion that derived from the term ‘function’ was exposed, when it is applied for specific biochemical goals (The Gene Ontology Consortium, 2000; Consortium, 2004). These are the main reasons why three independent ontologies should be built (The Gene Ontology Consortium, 2000).

The molecular level activities, such as catalytic and binding function, was denoted in molecular function (MF) ontology (The Gene Ontology Consortium, 2000; Consortium, 2004). It should be noted, that those activities, instead of the molecular entities, that are represented in GO MF (The Gene Ontology Consortium, 2000; Consortium, 2004). The assemblies of the molecular functions are described with biological goals are annotated in biological process (BP) ontology (The Gene Ontology Consortium, 2000). The location of the GO ontologies are described in cellular component (CC), at the levels of subcellular structures and macromolecular complexes (The Gene Ontology Consortium, 2000).

3.5.2 Mapping between GO and the SUPERFAMILY

The SUPERFAMILY domains and GO terms mapping is available at:
http://supfam.cs.bris.ac.uk/SUPERFAMILY/Domain2GO/Domain2GO_supported_only_by_all.txt

This section describes functional annotation of domain superfamilies. As already explained in section 2.2, domains are the functional building blocks of proteins, and its common ancestry is classified into superfamilies. SCOP is the database that maintain the superfamilies of protein domains (Murzin et al., 1995; Andreeva et al., 2004). In this respect, domain superfamilies have already annotated in functional manners (Vogel et al., 2004, 2005). This section will explain more about those annotation efforts, within the frame of its scheme of functionality, better eukaryotic superfamilies annotation, and improved SCOP classification (Vogel, 2010).

Function Scheme. The debate upon the definition of the protein or domain ‘function’ is not resolved yet until today, because of its contextual manner (Vogel, 2010). However, the role of the protein is the main information for the annotation process of domain superfamilies and it
3.5 GO Annotations

is mainly applied for certain pathway in the cell/organism (Nasir et al., 2011; Vogel, 2010). The combined definition of ‘molecular function’ and ‘biological process’ of GO annotation is the accepted formal definition for ‘function’ in GO annotation (Consortium, 2004; Vogel, 2010). COGs database scheme are similar with the functional terms of GO (Tatusov et al., 2003). In this end, a preparation of seven general function, with much more detailed 50 categories, was conducted. The mapping details of the functional categories is in the Table D.1 in the appendix (Vogel, 2010). The general categories of function are as follows: information, regulation, metabolism, intra-cellular processes, extra-cellular processes, general, and other/unknown (Nasir et al., 2011; Vogel, 2010). The diversification of functional features are occurring in some members of the superfamilies, especially the large groups (Vogel, 2010). The superfamilies have a clear description on their functional features, based upon its improved functional annotation (Todd et al., 2001; Vogel, 2010).

Annotation Scheme. The described functional scheme was utilized for the domain superfamily annotation into the SCOP classes. The information from SCOP (Andreeva et al., 2004), InterPro (Mulder et al., 2003, 2005), PFAM (Finn et al., 2006), SwissProt (Boeckmann et al., 2005) and literature (Vogel, 2010) are the basis of this annotation process. The position and GO functional annotation of PFAM domains into InterPro was used as a control (Mulder et al., 2003). The sequence similarity is the basis for the mapping process of PFAM into SCOP domain superfamilies. The consistency of the manual domain annotation was manifested in the GO annotation for PFAM, and with SUPERFAMILY as well (Consortium, 2004; Bateman et al., 2004; Madera et al., 2004). The scientific community already verified the validity of the large superfamilies annotation that are present in the eukaryotes genome (Vogel and Chothia, 2006a). This mapping process are already involving a broader community of scientist (Bashton, 2004; Vogel, 2010). Moreover, the error rate of this annotations is < 10% for large superfamilies, and <20% for all superfamilies (Vogel, 2010). SCOP domain superfamily distribution of functional annotation is shown in Figure 3.9. The most abundant category is the metabolism superfamily, for example, the enzymes (Vogel and Chothia, 2006b; Vogel, 2010). The metabolism functional annotation was mapped into almost of half of the superfamilies, and others categories have less than 15% of the total annotated domain superfamilies (Vogel, 2010). The metabolic superfamilies are found their one-third of its annotation in human genome, and it also covering one sixth of the all domains (Madera et al., 2004). Moreover, the unknown functions found its way into approximately 10% of the superfamilies (Schnoes et al., 2009; Vogel, 2010).

Functional Categories. In this end, 50 detailed function categories which map to seven more general function categories are listed in Table D.1 in the appendix (Vogel, 2010).
**Technical Background**

Figure 3.9: Domain function distributions. The *SUPERFAMILY* classification of *SCOP* (Andreeva *et al.*, 2004; Vogel, 2010). Figure taken from (Vogel, 2010), Fig.1.

### 3.5.3 Mapping between GO and the PFAM

The *InterPro2GO* database was utilized for the purpose of *PFAM* to *GO* mapping (Burge *et al.*, 2012; Hunter *et al.*, 2009). Combination of the various proteomic based databases are deposited in the Interpro database, and this includes *PFAM* as well (Koonin and Galperin, 2003; Hunter *et al.*, 2009; Burge *et al.*, 2012). The distinct InterPro entries were prepared for grouping the fingerprint of the identical protein family or domain (EBI, 2012). Moreover, UniProtKB database was utilized for cross-referencing the InterPro entry (EBI, 2012). The association of *GO* terms with InterPro entries can be reached when certain hits of InterPro entry are referring to the similarly functional proteins (Burge *et al.*, 2012).

The InterPro domains and GO terms mapping is available at:

http://www.geneontology.org/external2go/interpro2go

The InterPro team annotated the InterPro2GO file manually (EBI, 2011; Burge *et al.*, 2012). Matching comparison of InterPro and Protein entries are conducted by the curators, following (Burge *et al.*, 2012), by these steps:

1. Statistical verification of keywords, comments, and DE lines.
2. Examination of the conservation of the common annotation
3. Specify the exact GO terms that could be applied to the family

The accuracy of this method is exceptionally high (91-100%) (Camon and et.al., 2005).
THE quantitative view of protein domain distribution will be elaborated in this section. Why is it interesting to compare protein domain distribution between species?

The emergence of higher organisms was facilitated by a dramatic increases in the complexity of gene regulatory mechanism (Chen and Rajewsky, 2007; Parikesit et al., 2010b). This is achieved not only by the addition of novel regulatory mechanism but also by expansion of existing mechanisms. Such an expansion is usually characterized by an increase in functionally paralogous proteins or an increase in combinations of functional domains. Protein domains, thus, constitute units in evolutionary terms. They can be readily recombined in different arrangements leading to proteins that utilize different combinations of the same (types of) molecular interactions to fulfill different higher-level functions (Apic et al., 2001; Orengo and Thornton, 2005; Buljan and Bateman, 2009).

Over very large evolutionary time scales, such as those of interest in a comparative analysis of the eukaryotic kingdom, it becomes impossible in many cases to identify orthologous proteins since larger proteins are often manifested in a composite of domains deriving from several ancestral sources (Moore et al., 2008; Koonin et al., 2000). Fusions, fissions, and terminal loss have turned out to be much more frequent than the innovation of novel protein domains (Bornberg-Bauer et al., 2010; Zmasek and Godzik, 2011). Thus, the abundance and co-occurrence of domains becomes the most natural and promising framework to understand patterns of protein evolution at kingdom-level time-scales (see e.g. Yang and Bourne (2009)). In Zmasek and Godzik (2011), for instance, it was shown that frequent gains and losses of domains have lead to significant differences in functional profiles of major eukaryotic clades. Based on their results, the authors argue for the existence of a last eukaryotic common ancestor and suggesting that animals are gaining increased regulatory complexity at the
expense of their metabolic capabilities (Koonin, 2009; Parikesit et al., 2010b). Similarly, the rise of chromatin-based regulation mechanisms in crown-group eukaryotes can be traced by considering abundances and co-occurrences of the relevant protein domains (Prohaska et al., 2010).

4.1 Domain Annotation Data

Large scale phylogenetic analysis can shed light on the relative amounts of functional domains and their combinations and interactions involved in certain regulatory networks (Zhu et al., 2007; Parikesit et al., 2010a). Typically, studies of this type are based on existing annotation, for instance, the protein annotations compiled in databases like KEGG, ENSEMBL and PFAM (Finn et al., 2006). Moreover, based on the method in Prohaska et al. (2010), we used annotations and models of protein domains from the SUPERFAMILY database.

Several databases, most notably PFAM (Punta et al., 2012) and SUPERFAMILY (de Lima Morais et al., 2011), provide large collections of domain descriptions in the form of HMMs for annotating domains in the genome. Since protein domains are also regarded as functional units, the same databases provide maps to link domains with GO terms. These maps are obtained at least in part computationally as GO terms are primarily associated with entire proteins (Schug et al., 2002; de Lima Morais et al., 2011). Conserved protein function can be computed from the domain content as well (Forslund and Sonnhammer, 2008).

4.2 Pipeline and Results from the existing SUPERFAMILY annotation

The annotation pipeline of this section was prepared following Prohaska et al. (2010), by utilizing the SUPER script and the existing domain set. The function of the SUPER script could be partitioned into three parts. First, the script is compiling and annotating the existing gene and species-IDs. Second, it counts the occurrences of the designated domains in the five kingdoms of life. Third, the statistical result was shown, with an optional graphic plot. The latest version of the existing gene annotation, and species list was downloaded from the SUPERFAMILY database. Then, the domain distribution was calculated.

4.2.1 Chromatin Regulation

The chromatin computation discovery by Prohaska et al. (2010) has shown that domain combination of chromatin-associated enzymes in eukaryotic cells could be used to detect evolutionary patterns. Diverse modification patterns and better specificity is a result of reader domains combination. The crown-group Eukarya has an increased tendency of the utilization of reader domain combinations and it should be noted that Chromalveolata and...
4.3 Discrepancies in Existing Gene Annotations

Metazoa are having the most frequent combinations of methylation and acetylation readers (Prohaska et al., 2010).

According to Prohaska et al. (2010), methylation readers and RNA-binders has a unique and strong link in animals. RNA-binders was relied heavily by chromatin modification in animals more than any other eukaryotes group. However, the association of chromatin modification and RNA binding domain is not well annotated. It is assumed that nucleic acid binding are always interpreted by default as DNA binding, although some of them actually bind RNA. The zinc finger domain is an example of a domain, annotated as DNA binding that acts as RNA binder and DNA/RNA duplexes binder. Prohaska et al. (2010) states further that the chromatin modification machinery, such as (de)methylation, is anchored by information of the underlying nucleic acid sequence.

The Figure 4.1 shows the distribution matrix of CR domains across the kingdoms. It is shown, that most of the domains are annotated in all clades. It is also exposed, that functional domains and their combinations show increasing complexity with the complexity of the organism. It is shown as well, that annotation of Bacteria and Arachea is not as complete as their eukaryotic counterparts. Sudden expansions of the regulatory networks fall together with major innovations and potential changes to the regulatory concepts. In chromatin regulation, the binding of modified histones seem to be a eukaryotic innovation, enabling propagation, inheritance, and context dependent interpretation of histone modifications (Parikesit et al., 2010a; Prohaska et al., 2010).

4.2.2 Transcriptional Regulation

The Figure 4.2 clearly shows, that the massive expansion of transcription factor families and the pervasive combinatorial control of genes by multiple transcription factors is tightly correlated.

Moreover, results in Figure 4.3 show massive problems with data quality: closely related species (e.g, dolphin and human) show a dramatically different distribution of transcription factors and chromatin domains. Within vertebrates, this is not reasonable and contradicts biological knowledge (Parikesit et al., 2010a).

4.3 Discrepancies in Existing Gene Annotations

This section is dedicated for verifying the integrity and consistency of the existing gene annotation. The number of transcripts annotated as transcription factor transcript of the mammalian genomes is counted and compared to each other. This preliminary test is necessary to determine, whether the existing gene annotation is sufficient to use in the pipeline or not. The discrepancies of existing gene annotations are shown in Figure 4.4. There are a number of discrepancies in gene annotation, even within the same family. The best examples
Figure 4.1: Chromatin regulation distribution matrix. The x-axis contains clades from the 5 kingdoms of life, while the domain classes are shown in the y-axis. The Domain class are annotated in Eukaryotes, except in the ‘Rub’ class. No domain occurrences data on ‘Rub’. The black color signifies the most frequent occurrences, while the white signifies no occurrences. List of Abbreviation: ChAp (Chromosomal Architectural Protein), Dac (De-acetylation), Dme (De-methylation), Dph (De-phosphorylation), Dub (De-ubiquitination), Mac (acetylation modifiyer), Mme (methylation modifiyer), Mph (phosphorylation modifiyer), Mub (ubiquitination modifiyer), R– (reader of an unmodified side chain), Rac (acetylation reader), Rme (methylation reader), Rph (phosphorylation reader), Rub (ubiquitination reader), A (Arachea), B (Bacteria), E[BK] (basal Eukaryotes and Kinetoplastids), E[AO] (Chromalveolata), EF (Fungi), EM (Metazoa), and EV (Viridiplantae).

Figure 4.2: Transcription factor domain versus total protein plot. The occurrences of transcription factor domains are shown in the x-axis, while the occurrences of total proteins are shown in y-axis. The number of transcription factors scales with total number of proteins. The linear regression plot shows, that more total protein correlates with more transcription factor. The data points are the genomes of the various organism, that are counted for their TF domains and protein contents by using the existing SUPERFAMILY domain annotation. (Parikesit et al., 2010a)
are human and gorilla. The annotation quality of the Gorilla’s genome is worse than the human one. It is shown in Figure 4.4 that human has much more annotated transcript than Gorilla. It is argued, that the annotated transcript in the same order (in this case, the mammals) should be similar (ENSEMBL, 2012; Parikesit et al., 2010b). In order to correct this discrepancy, a gene prediction step extending the pipeline should be utilized (Parikesit et al., 2010b).

The GENSCAN annotation in Table 4.1 has shown a reasonable amount of predicted genes and domains for the tested species. However, GENSCAN doesn’t seem to work equally well on all species. It is shown that in kinetoplastids (Trypanosoma and Leishmania) that their genes are undercounted and this phenomenon happens as well in Plasmodium and Tetrahymena. The problem is caused by the unusual structure of their transcriptomes, which consists of long, polycistronic mRNAs that are processed by transsplicing (Martínez-Calvillo et al., 2010). The hits fall into a highly conserved polycistron of more than 10kb length, for which genscan predicts a “polyprotein” which is a protein that, after synthesis, is cleaved to produce several functionally distinct polypeptides. Interestingly, no spurious co-occurrences are found in the nematode C. elegans, whose polycistronic messages contain much fewer proteins. The second artifact are two hits in Phytophthora: one is again a putative artifact of genscan, which here predicts a chimera of RNA polymerase III subunit C34 and a hypothetical zinc-finger protein. The second hit covers a protein annotated as homolog of the EAP30 subunit of the ELL complex containing two winged-helix domains. In the latter case, the zinc-finger domain is most likely located in an additional downstream exon that is conserved between Phytophthora sojae and Phytophthora ramorum.

GeneMark (Lomsadze et al., 2005) was also tested as an alternative gene predictor and obtained comparable results. The focus on GENSCAN was implemented, because: (1) it has been reported to perform well across distantly related species (teleost fishes, nematodes, amphioxus, and fungi) without retraining its internal model (Korf, 2004), (2) because it is much faster than the alternatives, and (3) because it is the most widely used gene predictor (Miller et al., 2004).

4.4 The Necessity of Domain Re-annotation

Protein domains are re-annotated using the following three different collections of (putative) polypeptides for each genome: (1) computational translations of annotated transcripts available in sequence databases, (2) conceptual translations of the entire genomic DNA in all six reading frames (3 in each direction), and (3) protein predictions generated by a de novo gene predictor. As test system, the genomes of three apes (human GRCh37.57, chimp CHIMP2.1.57, and gorilla gorGor3.57) and yeast (SGD1.01.57) were used. Transcript files were obtained from the cDNA section of the corresponding genome builds.
Table 4.1: Summary of gene and domain annotation with GENSAN package. HMMER 3.0rc1 (Finn et al., 2011) was used to map the HMMs to the protein sequences with a cut-off $E \leq 10^{-3}$. The species with undercounted genes are *Plasmodium*, *Tetrahymena* *Trypanosoma* and *Leishmania*, while the species with overcounted genes is human.

<table>
<thead>
<tr>
<th>Species</th>
<th>all genes</th>
<th>zinc finger genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>genes</td>
<td>domains</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>118894</td>
<td>139016</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>28889</td>
<td>62906</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>12432</td>
<td>8752</td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>3578</td>
<td>8146</td>
</tr>
<tr>
<td><em>Aspergillus niger</em></td>
<td>8112</td>
<td>24334</td>
</tr>
<tr>
<td><em>Dictyostelium discoideum</em></td>
<td>5323</td>
<td>24496</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>64109</td>
<td>108972</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>20135</td>
<td>49974</td>
</tr>
<tr>
<td><em>Clamydomonas reinhardtii</em></td>
<td>13268</td>
<td>41576</td>
</tr>
<tr>
<td><em>Phytophthora ramorum</em></td>
<td>16701</td>
<td>53410</td>
</tr>
<tr>
<td><em>Thalassiosira pseudonana</em></td>
<td>8766</td>
<td>22006</td>
</tr>
<tr>
<td><em>Tetrahymena thermophila</em></td>
<td>2011</td>
<td>3028</td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>1439</td>
<td>3466</td>
</tr>
<tr>
<td><em>Naegleria gruberi</em></td>
<td>10748</td>
<td>28016</td>
</tr>
<tr>
<td><em>Leishmania major</em></td>
<td>4560</td>
<td>20554</td>
</tr>
<tr>
<td><em>Trypanosoma brucei</em></td>
<td>5143</td>
<td>20710</td>
</tr>
<tr>
<td><em>Trichomonas vaginalis</em></td>
<td>19251</td>
<td>49214</td>
</tr>
<tr>
<td><em>Giardia lamblia</em></td>
<td>11251</td>
<td>42324</td>
</tr>
</tbody>
</table>

However, testing the counts of all domains available in database is computationally too expensive. In order to save computation resources, 100 domains were randomly selected from the SUPERFAMILY database (Wilson et al., 2009; de Lima Morais et al., 2011), version 1.75 for the statistical analysis. HMMER 3.0rc1 Finn et al. (2011) with a cut-off $E \leq 10^{-4}$ was used to get a HMM based domain prediction in the protein sequences. In case of overlapping HMM hits, only the best-scoring match was retained. The result is a list of non-overlapping domains for each predicted protein. Here, zinc finger domains are used as an illustrative example since they form one of the most abundant classes of nucleic acid-binding domain. Other wide-spread domain families can be analyzed in the same manner. Operationally, a genscan prediction was classified as "zinc finger gene", if it contains at least one predicted C2H2 domain (SCOP family 57668).

Scatter-plots of the number of domain occurrences measured on the set of annotated transcript and on the de novo gene predictions shows a significant correlation, Figure 4.5. In contrast, an attempt to estimate the domain numbers by running the HMMER 3.0rc1 on translated genomic DNA failed miserably: only a small fractions of the known domains can be recovered. This is not surprising. Although there is a statistically significant correlation between protein domain boundaries and exon boundaries (Liu et al., 2005), about two thirds
of the annotated protein domains are interrupted by at least one intron, and on average, a domain contains 3 or 4 introns (Bhasi et al., 2009). Thus most domains are undetectable in the conceptual translations of the genomic DNA (Parikesit et al., 2010b).

The three ape species are so similar that a virtually identical complement of protein domains was expected. Even in very rapidly evolving gene families, such as the KRAB-ZNF family of transcriptional repressors, the copy numbers differences in between primates are restricted to a few percent (Nowick et al., 2010). The most extreme case are olfactory receptors (Niimura, 2009), where the number of functional copies differs by up to 25% between human and chimp due to massive gene loss (Consortium, 2008). This difference, however, will not be clearly detectable at domain level, since many of the very recent pseudogenes are expected to yield inconspicuous hits to the HMM domain models. In contrast to the expected similarity of the great apes, their transcriptome and proteome annotations differ by nearly a factor of three (Table 4.2 and Parikesit et al. (2010b)). In yeast, the correlation between domain counts from annotated transcript and GENSCAN predictions based domain annotation is excellent (Figure 4.5a and b). These differences were understood because of the quality and coverage of the transcript annotation. In the human genome, for example, a large number of annotated isoforms and alternative transcripts are annotated as a result of extensive cataloging efforts. Thus, multiple transcripts may incorporate the same genomic domain. A comparable density of data is not available for any other species, which results in an inevitable underestimation of annotated transcripts (as in the two ape genomes) (Parikesit et al., 2010b). However, Table 4.2 also shows that re-annotation by using GENSCAN or AUGUSTUS will narrow down the difference in gene and domain count between the three ape species. AUGUSTUS deserves notable mention because its gene count is closely resemble of the established RefSeq data (Lander and et al, 2001; Sequencing and Consortium, 2005; Scally et al., 2012).

There is a problem, however, with the direct use of the annotation from databases. In the human data, the majority of domains is observed more frequently in annotated transcripts than in GENSCAN predictions (Figure 4.5a). This effect is less pronounced in chimpanzee (Figure 4.5b). In Figure 4.5c, the domain counts in existing transcript are inconsistent between human and chimp. The problem comes from the biased annotation. Reannotation of genes and domains is necessary in order to reduce annotation bias in the quantitative statements. It is clearly shown in Figure 4.5d that the GENSCAN gene prediction give consistent domain counts, both in human and chimp. This is a solid evidence that domain annotations based on gene prediction could be useful in interspecies comparison of closely related species.

The discrepancy between Chimp and Human domain counts in existing transcript, and complete agreement on using gene predictors are definitely a phenomenon that observed only in multi-cellular organism. Yeast is having no problem on annotating its existing transcript and gene predictor, which is in total agreement in all of them. The discrepancy of primates genomes in this pipeline would make interspecies comparison not feasible. The inconsistent
Quantitative View of Protein Domain Distributions

Figure 4.3: Transcription Factor versus Chromatin domain plot. The occurrences of the chromatin domain are shown in the x-axis, while the occurrences of the transcription factor domains are shown in the y-axis. Blue box: Human, Kangaroo rat, Mouse, Opossum, and Fugu. Green box: Sea Squirt, Medaka, Medaka (different transcript set), Yeast, Dolphin. In the Blue box, those species are having only few chromatin-related domain but many transcription factors. Moreover, in the Green box, those designated species are having many chromatin-related domains but only few transcription factors. In this end, there are no known organisms that have a lot of both. (Parikesit et al., 2010a)

![Figure 4.3: Transcription Factor versus Chromatin domain plot.](image)

Figure 4.4: Discrepancies in the existing gene annotation. In the x-axis, The species in uppercase is the ones that extensively annotated, while the lowercases are not. The utilized RefSeq is version 42. The family classifications of the mammals are as follows: Primates (HOMO, pan, pongo, gorilla, MACACA, microcebus, otolemur, tarsius); Rodentia (MUS, RATTUS, spermophilus, and cavia); Dagonomorpha (oryctolagus and ochotona). Moreover, the Y-axis represents the total number of the existing transcript. The Figures show the total number of transcription factor transcript in each designated genome. The lines in different colors are representing different types of transcription factor domains. The transcript counts are varies from above 1000, and below 10. Thus, the count represents the widely abundant transcript in a cell, and the scarce ones as well.

![Figure 4.4: Discrepancies in the existing gene annotation.](image)
4.4 The Necessity of Domain Re-annotation

Figure 4.5: Correlation of the number of predicted protein domains. This is the inter- and intra-species comparison of the gene predictions and annotated transcript. The dots represent the domain counts of each axis. Top row: Annotated transcripts compared to de novo predicted “genes” for (a) human, and (b) chimp. Below: While domain prediction based on existing annotation yield systematic differences between human and chimp (c), congruent abundances are obtained from genscan predictions (d). Linear regression is shown as red line in panels (c) and (d). Bottom row: Annotated transcripts compared to de novo predicted “genes” for yeast. Different gene predictors (GENSCAN and GENEMARK) yield comparable results (f), shown here for yeast.
Table 4.2: Summary statistics of source data. The number of domains refers to query set of 100 randomly selected SCOP. n.d.: not determined.

<table>
<thead>
<tr>
<th>Species</th>
<th>Human Data set</th>
<th>Chimpanzee Data set</th>
<th>Gorilla Data set</th>
<th>Yeast Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RCh37.57</td>
<td>CHIMP2.1.57</td>
<td>gorGor3.57</td>
<td>SGD1.01.57</td>
</tr>
<tr>
<td>number of peptides investigated</td>
<td>number of detected domains</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transcripts</td>
<td>76592</td>
<td>34142</td>
<td>27325</td>
<td>5885</td>
</tr>
<tr>
<td>GENSCAN</td>
<td>118894</td>
<td>96615</td>
<td>113532</td>
<td>4197</td>
</tr>
<tr>
<td>AUGUSTUS</td>
<td>33507</td>
<td>29712</td>
<td>42906</td>
<td>5402</td>
</tr>
<tr>
<td>genomic translation</td>
<td>23</td>
<td>n.d.</td>
<td>n.d.</td>
<td>409</td>
</tr>
</tbody>
</table>

Domain count of interspecies comparison from existing transcript is the main motivation of the re-annotation efforts. This efforts should eventually give consistent result in every genomes. In the following section, the complete re-annotation pipeline will be explained.

4.5 Arli’s Domain Distribution Pipeline

The re-annotation method for obtaining comparable domain distributions involves several programs that we compiled to what we called Arli’s Domain Distribution Pipeline (ADD). The main elements are a gene prediction package and the HMMER domain annotation package. Both, gene predictor and HMMER are using Hidden Markov models. In Figure 4.6, the ADD pipeline is shown in a schematic diagram. The ADD pipeline’s first step is the utilization of gene predictors. Then, the predicted protein data set is used for the domain annotation. Subsequently, the annotation could be used to determine the domain co-occurrence/avoidance tendency and for the functional analyses. The single steps of the pipeline will be explained in next chapters as follows:

Chapter 5 will elaborate the utilization of gene prediction package that are used for the gene re-annotation process. The phylogenic tree of our selection of species from different clades will be presented and the feasibility of the gene annotation process will be evaluated. The performance of GENSCAN and AUGUSTUS will be compared with the RefSeq gene annotation.

Chapter 6 will explain in more detail the computation of domain occurrences. The utilized software packages and specific parameter settings will be shown. Different sources of protein domain data will be compared and their efficacy in our pipeline will be discussed.

Chapter 7 is about the computation of domain co-occurrence and avoidance. The definition of domain co-occurrence and avoidance will be given, as well as the implication for this
4.5 Arli’s Domain Distribution Pipeline

Figure 4.6: Schematic representation of Arli’s Domain Distribution (ADD) pipeline. The first step (a) is the gene prediction, utilizing GENSCAN or AUGUSTUS programs. (b) The domain annotation utilize the HMMER package with HMMs obtained from PFAM or SUPERFAMILY . (c) The domain annotation data was employed for co-occurrence/avoidance tendency. (d) The functional analysis section is responsible for pooling the domain annotation data sets into its respective functional groups.

pipeline. As an example, the Zinc Finger domain co-occurrence and avoidance tendency will be discussed in details.

Chapter 8 will describe the computation of a functional grouping of domains using GO annotations provided in the GO database. By pooling the domains to a respective functional group, a better resolution of co-occurrences/avoidance can be obtained. The co-occurrences/avoidance Zinc Finger domain co-occurrence and avoidance will be discussed in detail.
Gene prediction is the initial state of the Arli’s Domain Distribution (ADD) pipeline that was developed in this dissertation. As already mentioned in Chapter 1 and Section 4.5, ADD is our developed pipeline that reannotate protein domain by using gene predictor. The complete scheme of ADD is shown in Figure 1.3.

First, the ADD pipeline was developed with GENSCAN package, as one of the most important gene prediction packages in the earlier genome project (Burge and Karlin, 1998; Parikesit et al., 2010b, 2011c). As GENSCAN did not allow sufficient development of the independent training set, a more feasible alternative was sought. We subsequently used the AUGUSTUS package, because of its increasing popularity in the gene annotation project (Stanke et al., 2006; Parikesit et al., 2011c). Both GENSCAN and AUGUSTUS packages use HMM method for their gene finding algorithm. The result of our pipeline was compared with the existing protein complement.

The most complete information on the protein complement can be inferred from the genome sequence. In fact, only two thirds of the predicted human proteins have been directly observed by experiments so far (Nilsson et al., 2010). For most of the less-studied species, on the one hand, the set of predicted proteins in the current genome annotations is far from being complete. For example, the number of annotated transcripts varies by more than a factor of three even between great ape genomes (Parikesit et al., 2010b).

The accumulation of transcriptomics data in a few well-studied organisms such as human, mouse, or fruitfly, on the other hand, leads to an increasing number of annotated splice variants and transcripts with alternative start sites, and thus to an increasing number of overlapping protein variants (Schrimpf et al., 2009). It was argued that the large ascertainment biases in current protein databases make these data effectively impossible to
use for a quantitative comparison of protein domain abundances across species (Cappadona et al., 2012; Parikesit et al., 2010b, 2011c). Instead, utilization of de novo gene predictions to obtain quantitatively comparable estimates, (Figure 1.3) was proposed. It was shown that a simple general-purpose gene finder such as GENSCAN and AUGUSTUS (Burge and Karlin, 1997; Burge, 1998; Burge and Karlin, 1998; Stanke and Waack, 2003; Stanke, 2003) already yields plausible numbers.

5.1 Gene Annotation

The key feature of ADD pipeline is to cover the most representative genomes from the evolutionary diverse clades. A total of 18 species with sequenced genomes covering the entire phylogenetic range of the eukaryotes were considered, (Figure 5.1). Table B.1 shows the web link of the annotated genomes that is used in this pipeline.

---

**Figure 5.1:** Phylogenetic distribution of the species considered in this work following (Baldauf, 2008), showing the disputed deepest nodes unresolved. The species name and genome assembly: Homo sapiens (hg19); Drosophila melanogaster (BDGP5.13); Caenorhabditis elegans (WS200); Schizosaccharomyces pombe (EF1); Aspergillus niger (CADRE); Dictyostelium discoideum (DDB); Oryza sativa (OSV6.1); Arabidopsis thaliana (TAIR9.55); Chlamydomonas reinhardtii (Chlre4); Phytophthora ramorum (Phyra1); Thalassiosira pseudonana (Thaps3); Tetrahymena thermophila (ttal_oct2008); Plasmodium falciparum (PlasmoDB-7.0); Naegleria gruberi (Naegr1); Leishmania major (Lmj_20070731_V5.2); Trypanosoma brucei (Tb927_May08_v4); Trichomonas vaginalis (TrichDB-1.2); Giardia lamblia (WBC6); Stram.: Stramenopiles; Alveol.: Alveolata; Diplom.: Diplomonada. Multicellular species are marked by a black dot, unicellular ones with a white dot. The gray dot marks the slime mold, which is a social amoeba that has ability to alternate between unicellular and multicellular forms (Eichinger and et al., 2005).
In this section, gene predictions were performed using GENSCAN (Burge and Karlin, 1997; Burge, 1998; Burge and Karlin, 1998). Following Parikesit et al. (2011b), we split long chromosomes into overlapping fragments of about 500 kb to accommodate GENSCAN restriction on input length. GENSCAN was used with the default parameters. Protein sequences were extracted directly from the GENSCAN predictions. Duplicate predictions in the overlaps between fragments were removed. Although this procedure in general yields good results, as shown previously for mammals and yeast (Parikesit et al., 2010a), care must be taken in case of unusual genome structures. In the case of polycistronic mRNAs, as in the case of the kinetoplastids (Leishmania and Trypanosoma), a tendency to overcount co-occurrences was expected since polycistrons are not correctly split into individual functional units. Short scaffolds, as in the case of the Tetrahymena data, on the other hand, lead to underestimates. The extreme A + T content of Plasmodium, furthermore may account for the relative small number of predicted genes and the low number of reliably annotated domains (Coulson et al., 2004; Lu et al., 2007). The GENSCAN gene predictions are summarized in Table 5.1.

Moreover, Table 5.1 exposes some anomalies in the GENSCAN gene prediction in comparison with RefSeq annotation. Some noticeable over and underestimation seem to exist. The most noticeable overestimation is from the human genome, at the unreasonable 118894 counted genes. This estimate is not plausible, since the estimated gene fraction in human falls 36703 genes (Lander and et al., 2001). This could be argued as a sign of false positive genes. Contrary, the Plasmodium and Tetrahymena genome shows striking underestimation, because the RefSeq annotated genes are much higher than that number in Table 5.1 (Gardner and et al., 2002; Eisen and et al., 2006). Moreover, because GENSCAN was not trained for genomes that dominated with single-exons, such as Trichomonas vaginalis and Giardia lamblia, it is expected that the overcounting in Trichomonas is false positive as well, while the undercounting in Giardia is false negative. The remaining part of this chapter aims at discussing and solving this problem.

5.3 AUGUSTUS

AUGUSTUS (Stanke and Waack, 2003; Stanke et al., 2006, 2008) was used for gene prediction, because the package has gained popularity in genome annotation projects and because it can be trained for applications to a given genome with known cDNAs. Moreover, the availability of polycistrons and transsplicing mechanisms in certain genomes explained in Section 2.1 could possibly hampering the result of the GENSCAN based predictions. That’s why the trained gene approach is expected to overcome the limitation of the non trained approach that could not perform well in the species with strange genome architecture.
Table 5.1: Summary of gene annotation. It gives the results from GENSCAN and AUGUSTUS with both online (web-based) and offline (local) training methods and the contents of the RefSeq database. For the species with the equal result of Offline and Online method, the columns are joined together. See sect. 5.1 for full species names, and information for the phylogenetic distribution of the 18 investigated species could be found in Figure 5.1. List of Abbreviations: Onl. (AUGUSTUS online), Offl. (AUGUSTUS offline), and Refs. (Refseq). For more details, see the main text.

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene Total</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Giardia</td>
<td>4357 5178</td>
<td>11251 6583</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichomonas</td>
<td>61750 60924</td>
<td>19251 60815</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trypanosoma</td>
<td>7874 9696</td>
<td>543 10192</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leishmania</td>
<td>9451 4560</td>
<td>9155</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naegleria</td>
<td>16792 16443</td>
<td>10748 16620</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium</td>
<td>6043 1439</td>
<td>5512</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tetrahymena</td>
<td>21650</td>
<td>2011 24725</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thalassiosira</td>
<td>10428 10528</td>
<td>876 10958</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phytophthora</td>
<td>17154 16292</td>
<td>16701 15743</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlamydomonas</td>
<td>15111 13268</td>
<td>14488</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>27945 20135</td>
<td>25498</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oryza</td>
<td>62327 63693</td>
<td>64109 62709</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictyostelium</td>
<td>12904 12595</td>
<td>5323 12646</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspergillus</td>
<td>9866 8112</td>
<td>10785</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizosaccharomyces</td>
<td>4783 3578</td>
<td>4824</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caenorhabditis</td>
<td>22902 12372</td>
<td>21175</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drosophila</td>
<td>14217 28889</td>
<td>13601</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo</td>
<td>33507 118894</td>
<td>36073</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To carry out our analysis, both "off-line" (local) and "on-line" (web-based) trained models were prepared as described in the AUGUSTUS tutorial (Stanke, 2011). For several species, default training sets are provided on the AUGUSTUS website. We used default parameters for the AUGUSTUS package. For this data set, there was not difference between local and web-based training. For the remaining species, cDNAs are available in GenBank. Redundancies were removed with a PERL script. The FASTA sequences were used and their headers were cleaned from meta-characters and gaps. Models were trained both "off-line" and using the pipeline offered at the AUGUSTUS website. For the applications, AUGUSTUS was configured to generate only non-overlapping protein-coding genes. The predicted protein sequences are part of the AUGUSTUS output. Using bed-tools, it was verified, that no overlapping sequences were contained in the output (Quinlan and Hall, 2010). Moreover, due to its inapplicability in other organisms than human, Galdieria, Toxoplasma, and Caenorhabditis genomes, the UTR prediction was disabled (Stanke, 2004).

The results of AUGUSTUS are compiled in Table 5.1 together with the RefSeq (release 53) genes for each of the 18 species.

In this respect, compared with GENSCAN, AUGUSTUS could produce a much more acceptable annotation count due to its resemblance to RefSeq counts. The under-and overcounting
that was exposed in the GENSCAN annotation is substantially reduced or even not present in AUGUSTUS package. It could be observed in the AUGUSTUS gene count for human, 33507 genes, which is much closer to the RefSeq count. The GENSCAN undercounting of Plasmodium and Tetrahymena has already solved in AUGUSTUS annotation as well. The problems with single exons genome (Trichomonas and Giardia) are already resolved. The undercounting in Trichomonas and overcounting in Giardia are not present in AUGUSTUS annotation, as the result is now closer to the RefSeq gene count. We argued, that AUGUSTUS could give better approximation of gene counts than GENSCAN given its ability to train datasets using the latest assembly of the gene model. However, extra care must be taken, because in some genomes, AUGUSTUS is having an “offline” and “online” method of prediction. The count difference between “online” and “offline” came from the duration of iteration of the gene model training process. AUGUSTUS online method has a longer iteration period than the offline. It has been suggested that the iteration period for the offline method should be shut off after one day. However, the online method could extend well beyond one day. Therefore, the “online” version was chosen to finish the predictions, because it covers more gene annotation.
The comparison of the results between the AUGUSTUS gene prediction and the RefSeq gene inventories agrees rather well in some species, whereas in others there are substantial
Figure 5.2: Comparison of gene predictions for the 18 species. For each annotated species, a venn diagram is shown for both online and offline the raw output of the gene predictions (Dilts, 2012). RefSeq annotation is shown in red, AUGUSTUS prediction with online and offline trained models are shown in blue and green, respectively. For the species with only green circle of AUGUSTUS annotations, it means that there is no difference between online and offline counts because they are using the same trained model. Corresponding numbers could be found at Table 5.1. For more details, see the main text

differences, depending on the various degree of completeness of the gene annotation, (Figures 5.2). In most of the genomes, the total number of annotated RefSeq is always roughly
equal to the total number of annotated \texttt{AUGUSTUS} genes. This shows that the \texttt{AUGUSTUS} gene prediction could annotate the \texttt{RefSeq} genes as well. However, it is shown in the human genome that the \texttt{AUGUSTUS} annotation have much more annotated genes than \texttt{RefSeq}. This is due to the existence of isoforms and hypothetical proteins.

\textbf{5.4 The Performance of GENSCAN and AUGUSTUS Gene Predictors}

The existing gene annotation analysis shows that there are serious discrepancies among different genomes. These could be due to the different level of curation efforts on each of genome. Moreover, this situation must be corrected by developing new semi-automatic annotation method. Thus, our gene predictor based pipeline, ADD, was developed to cope with this challenge. Moreover, \texttt{RefSeq} annotation is not available everywhere. So, it is important to have a pipeline that can also run on new genomes. The gene prediction package incorporated in ADD could predict a gene based method on the structural model, such as exon-intron boundaries. This would eventually correct the discrepancies in the existing genome annotation.

Moreover, it was shown that quantitative comparative analyses are possible if they are based on predictions of trainable gene predictor approaches such as \texttt{AUGUSTUS}. The training phase is necessary given that it deals with particular artifacts introduced by peculiarities of the genome structure, as explained in Section 2.1. Untrained tools such as \texttt{GENSCAN}, for instance, have problems to recognize the protein boundaries in polycistronic transcripts of kinetoplastids, and experiences difficulties with extreme A/T contents, or lack sensitivity e.g. in very intron-poor genomes. Such effects are largely fixed by species-specific training as provided by \texttt{AUGUSTUS} package. However, in the current version of \texttt{AUGUSTUS}, only limited amount of the species training set is provided. As explained in section 5.3, unavailable species data set must be prepared with training procedures. The ADD pipeline was developed by combining both \texttt{GENSCAN} and \texttt{AUGUSTUS} methods. Although such a procedure could eventually provide a robust training set, it is likely that in the next version of \texttt{AUGUSTUS}, a more complete training data set will be provided.
Protein domains can be described as Hidden Markov Models (HMMs) (Eddy, 1996; Durbin et al., 1998; Eddy, 1998) and utilized as powerful tool for protein domain identification (Ochoa et al., 2011; Terrapon et al., 2009, 2012). One main objective of this research is to annotate protein domains in a wide-range of species. This would require the development of an inter-species correlation test. The description of the test will be exposed in the following paragraph.

6.1 Goal of Annotation on the Gene Prediction

The comparative and functional analysis of three regulatory mechanisms was performed: (1) transcriptional regulation by transcription factors, (2) post-transcriptional regulation by miRNAs, and (3) chromatin regulation across all domains of life. All of these mechanisms are evolutionarily old and passed through several major innovations. Single domain distributions and domain co-occurrences from SUPERFAMILY domain annotations was calculated (Gough and Chothia, 2002).

Functional annotation of domains from GO and protein domain descriptions was integrated into our comparative analysis (Parikesit et al., 2010a; Hunter et al., 2009). The investigation of the strategies to construct the inventory of protein domains that avoid the bias arising from discrepancies in gene annotation was focused (Parikesit et al., 2010b; Schrimpf et al., 2009). While it would certainly be desirable to obtain a complete set of protein domains encoded in any given genome, this is not computationally feasible at present (Parikesit et al., 2010b). Thus, the goal here is more moderate: satisfaction with estimates that are consistent between different genomes and thus allow a quantitative comparison (Parikesit et al., 2010b).
6.1.1 Results from the AUGUSTUS Predicted Genes with PFAM Annotation

In order to compute the occurrence of genes with at least one PFAM annotated domain of the two training modes of AUGUSTUS were compared with each other and with the RefSeq annotation. Their overlaps were computed with bed-tools and lucidchart was used to create Venn diagrams of the overlaps displayed in Figure 6.1 (Quinlan and Hall, 2010; Dilts, 2012).

The entire Pfam database version 26.0, comprising 33672 domain models as well as the entire collection of 9821 Hidden Markov Models (HMMs) provided by the SUPERFAMILY database (version 1.75) were used. In both cases we used HMMER3.0rc1 (Eddy, 2011) with an E-value threshold of $E \leq 10^{-3}$ to map the HMMs to the predicted amino acid sequences as well as the RefSeq protein annotations (Parikesit et al., 2011c).

In order to test the quality of gene predictions, sub-collections of protein sequences were compared with at least one mapped Pfam domain between the gene prediction methods and RefSeq database. The results are shown in Figure 6.1. Overall, the online-trained AUGUSTUS predictions have the best coverage of the manually curated RefSeq and are hence used as data basis for subsequent quantitative analysis (Parikesit et al., 2011c).
6.1 Goal of Annotation on the Gene Prediction

- **Homo sapiens**
- **Naegleria gruberi**
- **Oryza sativa**
- **Trypanosoma brucei**
- **Trichomonas vaginalis**
- **Giardia lamblia**
- **Leishmania major**
- **Plasmodium falciparum**
- **Tetrahymena**
- **Thalassiosira pseudonana**
Protein Domain Annotation Using HMMs

Figure 6.1: Comparison of gene with at least one domain in 18 species. For each species, a Venn diagram was shown for both the raw output of the gene predictions and for the subset of proteins with at least one matching Pfam model. RefSeq is shown in red AUGUSTUS prediction with online and offline trained models are shown in blue and green, respectively (Parikesit et al., 2011c).

6.2 Comparison of SUPERFAMILY and PFAM Results

In the Table 6.1, the comparison of SUPERFAMILY and PFAM domain annotation is shown. The data was generated by counting all genes that contain at least one SUPERFAMILY or
Table 6.1: Summary of domain annotation. The following blocks of columns list the numbers of genes that contain at least one SUPERFAMILY (SF) or PFAM domain, respectively. Below, the phylogenetic distribution of the 18 investigates species is summarized (Baldauf, 2008). See sect. 5.1 for full species names. List of Abbreviations: Onl. (AUGUSTUS online), Offl. (AUGUSTUS offline), and Refs. (RefSeq).

<table>
<thead>
<tr>
<th>Species</th>
<th>#Genes(SF)</th>
<th>#Genes(Pfam)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giardia</td>
<td>3240</td>
<td>3265</td>
</tr>
<tr>
<td>Trichomonas</td>
<td>3278</td>
<td>3344</td>
</tr>
<tr>
<td>Trypanosoma</td>
<td>4626</td>
<td>4626</td>
</tr>
<tr>
<td>Leishmania</td>
<td>4949</td>
<td>3758</td>
</tr>
<tr>
<td>Naegleria</td>
<td>6572</td>
<td>6442</td>
</tr>
<tr>
<td>Plasmodium</td>
<td>4110</td>
<td>1189</td>
</tr>
<tr>
<td>Tetrahymena</td>
<td>2502</td>
<td>1856</td>
</tr>
<tr>
<td>Thalassiosira</td>
<td>6248</td>
<td>6145</td>
</tr>
<tr>
<td>Phytophthora</td>
<td>7384</td>
<td>7382</td>
</tr>
<tr>
<td>Chloridomonas</td>
<td>6852</td>
<td>7062</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>8088</td>
<td>11957</td>
</tr>
<tr>
<td>Oryza</td>
<td>8580</td>
<td>7527</td>
</tr>
<tr>
<td>Dictyostelium</td>
<td>6877</td>
<td>6744</td>
</tr>
<tr>
<td>Aspergillus</td>
<td>6432</td>
<td>5407</td>
</tr>
<tr>
<td>Schizosaccharomyces</td>
<td>4259</td>
<td>2532</td>
</tr>
<tr>
<td>Caenorhabditis</td>
<td>7418</td>
<td>4329</td>
</tr>
<tr>
<td>Drosophila</td>
<td>7654</td>
<td>11618</td>
</tr>
<tr>
<td>Homo</td>
<td>8908</td>
<td>31359</td>
</tr>
</tbody>
</table>

**PFAM** domain. The data shows, that in general, the **PFAM** annotation yields higher gene counts compared to **SUPERFAMILY**. However, some species have more genes annotated for **SUPERFAMILY**. These are Giardia, Trypanosoma, Leishmania, Plasmodium, Tetrahymena, and Dictyostelium. The trends of the result clearly expose, that **SUPERFAMILY** could give better annotation on the genomes of the protist, while **PFAM** is better suitable for more complex and multicellular organism such as plants and metazoa.

### 6.3 The Necessity of using Gene Predictors

The data quality problem that encountered in section 4.2 has its consequences, that existing **SUPERFAMILY** protein domain annotations cannot be used for large scale quantitative comparisons due to several sources of bias (Parikesit et al., 2010a; Sandhya et al., 2009). They are, namely:

1. Different completeness of protein annotation for different genomes, which can be seen in the existing transcript counts in Table 4.2. The transcript counts of human and chimpanzee differ by a factor of two. Moreover, for gorilla, the case is roughly the same.

2. Differences in transcript coverage. This case was already explained in Section 4.3 and it shows that for the genomes within the same family (mammals), there is a very different
Protein Domain Annotation Using HMMs

extend of transcript coverage. The genomes that are extensively used for molecular research, such as human and chimps, are having a rather consistent and similar coverage, while the scarcely annotated genomes, such as *oryctolagus* and *pongo*, are behind in their coverage.

3. Different coverage of protein domains at kingdom level. The Figure 4.3 shows, that the organisms in the same kingdom have a very different protein domain coverages.

4. Misannotations of functions, e.g., the chromodomain, a chromatin regulation domain (*SUPERFAMILY* ID 54160), is annotated as a transcription factor in SCOP. (Data in the online supplementary material).

The genomics comparative studies typically resort to testing for relative enrichment rather than considering absolute numbers of domains. In studies focusing on the evolution of regulatory mechanisms and regulatory complexity, however, absolute gene counts play an important role.

Similarly, investigation of lineage-specific variations in regulatory schemes require plausible statistics of protein domains and their combinations (Prohaska *et al.*, 2010). The HMM models of protein domains are easily searched against the translation of these ORFs and included e.g. in the *SUPERFAMILY* database. False positives in the ORF annotation pose little problem since they are very unlikely to contain recognizable protein domains (Parikesit *et al.*, 2011c, 2010b; Lander and et al, 2001). Here, it was shown that protein domains can be annotated with acceptable accuracy using *de novo* gene predictors such as *GENSCAN*. This strategy also avoids methodological biases such as the enrichment of 3'-exons in poly-A ESTs (Parikesit *et al.*, 2011c, 2010b; Costa *et al.*, 2010). Since the distributions of protein domains are of interest, comparison of *RefSeq* data with gene predictions was restricted to only those genes in which at least one *PFam* domain was annotated. For most species, this improves the congruence between the gene sets. In a few cases, however, the differences persist, as in the case of *Trypanosoma* and human, shown in Figure 6.1. In *Trypanosoma*, most of the difference is explained by annotated *RefSeq* proteins without recognizable domains. In human, the discrepancy is in part explained by *RefSeq* isoforms and in part by *AUGUSTUS* prediction without domains (Parikesit *et al.*, 2011c, 2010b).

Among the predictions with annotated *PFAM* domains in Figures 6.1, it was found, e.g. for *Leishmania*, *Tetrahymena*, and *Plasmodium* that both the online and the offline trained gene predictions have a much larger coverage than the *RefSeq* data. For *Trichomonas* and *Giardia*, the situation is reversed. This can probably be explained in part by the large number of paralogs and possible pseudogenes included in *RefSeq* in *Trichomonas*, but also indicated as lack of sensitivity of the gene predictor for the two parabasalids with their extremely intron-poor genomes. At the domain level, *AUGUSTUS* and *RefSeq* agree nearly perfectly e.g. in human in *Naegleria*. In general, the *RefSeq* entries missed by the gene predictor are
frequently putative pseudogenes and ORFs lacking further annotation. Since the AUGUSTUS "offline" predictions overall yield the most inclusive data set, these predictions are used below for all statistical analysis of domains compositions (Parikesit et al., 2011c, 2010b).

One of major ascertainment biases in the analysis of large scale evolutionary patterns of functional domains are the protein domain databases themselves. Recent studies reported the innovation of a large number of domain innovation events within both the green plants (Kersting et al., 2012) and the animals (Moore and Bornberg-Bauer, 2012). The number of identified clade-specific domains must be expected to depend on the depths in which the clade is studied (Segata and Huttenhower, 2011). The domain inventory is thus probably more complete in animals, fungi, and plants animals compared to most protozoan lineages. Large numbers unannotated domains of course undermine the analysis presented here since the lead to a systematic under-estimation of an organisms metabolic or regulatory capability, in particular since (Kersting et al., 2012) also reported that the novel domains in stress response and developmental innovations. A more systematic survey of so-far undescribed protein domains thus constitutes a natural next step towards a comprehensive understanding of functional evolution in the eukaryotes (Goetze et al., 2009; Parikesit et al., 2011c, 2010b).
The purpose of this part is to emphasize efforts on elucidating the protein domain content based upon the existing and predicted gene annotation. Distribution of protein domain is an extensively studied field. It was found that domain distribution always follows small-world and scale-free topologies, which stated that there is expansion of particular domain families architecture when the organism is more complex (Wuchty, 2001). It could be interesting to investigate more outcomes on the interrelationship between organism complexity and protein domain distribution. Here, the protein domain distribution in Transcription Factors (TFs) was analyzed on sampled species. The combination of de novo gene prediction and subsequent HMM-based annotation of SCOP domains leads to consistent and comparable estimates of co-occurrences with acceptable accuracy. In particular, it can be utilized for systematic studies of the evolution of protein domain occurrences and co-occurrences (Parikesit et al., 2011a).

7.1 Definition of Domain Co-occurrence

Domain co-occurrence is defined as the tendency of co-existence of two or more domains in the same protein (Parikesit et al., 2010b, 2011c; Wuchty, 2001; Wuchty and Almaas, 2005). Moreover, avoidance is defined as the tendency of restraining of two or more domains in the same protein (Parikesit et al., 2010b, 2011c; Wuchty, 2001; Wuchty and Almaas, 2005). Both the co-occurrence and avoidance are two common phenomena that could be deserved in the organisms. It was found that there is a progression of domain co-occurrence networks by the increasing of evolutionarily development based on complexity, from single cellular to multi-cellular organisms (Wuchty and Almaas, 2005). It would be interesting to investigate
Further what is the tendency of domain distributions in the organisms in the light of complexity studies. The ADD (Arli’s Domain Distribution) pipeline was developed to address this challenge.

A more modest approach thus aims at tracing the distribution of protein domains comparatively. In a recent study of chromatin evolution, it was demonstrated that this is indeed feasible (Prohaska et al., 2010). More detailed insights can be gained from considering domain combinations. For instance, Itoh et al. (2007) showed that there are many animal-specific or even vertebrate-specific domain-combinations. Network analysis of domain co-occurrences, furthermore, demonstrates a growing core of combinations in multi-cellular organisms (Wuchty and Almaas, 2005).

The expectation values for each pairwise co-occurrence was calculated as follows:

### 7.1.1 Expectation Method Count

For each species, separate evaluation of the number of domain co-occurrences and the number of genes in which two domain types $x$ and $y$ co-occur was done. Here $x$ and $y$ can be either individual domains, sets of domains belonging to the same superfamily, or the collections of domains compiled into functional classes according to their GO annotations. Let $n_x$ be the total number of annotated domains belonging to group $x$. The calculation of the expected number of co-occurrences of domain $x$ and $y$ based upon the single domain occurrences and its total number of genes is given in equation (7.1).

$$E(x, y) = \frac{n_x \times n_y}{n_g}$$  \hspace{1cm} (7.1)

where:

- $n_x$ is the number of genes with domain $x$
- $n_y$ is the number of genes with domain $y$
- $n_g$ is the total number of genes.

This can be computed for GENSCAN and AUGUSTUS predictions, with SUPERFAMILY and PFAM annotation. The expectation value is then compared with the number of genes $n(x, y)$ in which $x$ and $y$ co-occur (Parikesit et al., 2010b). However, the estimate in equation (7.1) does not account for biases arising from the non-uniform distribution of domains over genes. Let $n_d(i)$ be the number of domains predicted for protein $i$, and let $n_d = \sum_{i} n_d(i)$ be the total number of domains. Then the number of domains $x$ that occur in genes that also contain a domain $y$ can be estimated as

$$E(x|y) = \frac{n_x}{n_d} \sum_{i : y \in i} (n_d(i) - 1)$$  \hspace{1cm} (7.2)
where the sum runs over all genes \( i \) that contain a domain belonging to group \( y \). The alternative estimate was obtained, by exchanging \( x \) and \( y \) in Equation (7.2). These expectations were compared with the number of empirically observed co-occurrences \( n(x,y) \) (Parikesit et al., 2011a,c). Moreover, the correction factor of \(-1\) in the equation came from the assumption, that no domain occurrence of 1 should be counted for its significance.

### 7.1.2 Significance Cut-off

**Domain co-occurrence.** From Equation (7.1) , if \( E(x,y) > n(x,y) \) then avoidance of domains is observed. On the other hand, if \( E(x,y) < n(x,y) \) then co-occurrence is preferred (Parikesit et al., 2010b). However, the overly simplified equation (7.1) was corrected, by incurring a more comprehensive significance cut-off definition for equation (7.2). We speaks of co-occurrence of \( x \) and \( y \) if \( n(x,y) \gg \max\{E(x|y),E(y|x)\} \) and of avoidance if \( n(x,y) \ll \min\{E(x|y),E(y|x)\} \). The statistical significance of an observed difference between \( n(x,y) \) and the values of \( E(x|y) \) and \( E(y|x) \), respectively, is determined under the assumption that \( n(x,y) \) is drawn from a Poisson distribution.

**Poisson distribution.** The poisson distribution is an independent probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event (Haight, 1967). It is used to determine if the observed count significantly deviate from its expectation. More precise, a poisson distribution with mean \( E(x,y) \) is used (Parikesit et al., 2011a,c; Wooldridge, 1999). In figure 7.1, the poisson distribution of domain co-occurrence and avoidance is shown. The equation (7.3) was utilized for tendency of co-occurrence

\[
P = \text{Poisson}(X, X_0, 1) \tag{7.3}
\]

Domain avoidance happens when \( X > X_0 \). The equation (7.4) was utilized for the tendency of avoidance .

\[
P = 1 - \text{Poisson}(X - 1, X_0, 1) \tag{7.4}
\]

The co-occurrence/avoidance domain pair will be counted as significant, if \( P < 0.05 \)

**Randomize the assignment of annotations to domains.** This simulation was conducted to see whether our ADD pipeline of co-occurrence/avoidance make sense. The assignment of functional annotation to domains was randomized by using the `shuf` Linux command for random permutation (Linux, 2012). The simulation was run exclusively on annotation of human gene prediction by AUGUSTUS . Both domain annotations SUPERFAMILY and PFAM, were tested for their insignificance. It is shown in Figure 7.2, that no significant result apply. This means that the 'test of test' is working, and the ADD pipeline is working
Figure 7.1: The graphic of Poisson distribution (Triola, 2003). The y axis is the probability mass function, which is defined as a function that gives the probability that a discrete random variable is exactly equal to some value. The x axis is the index k, the number of occurrences. The function is only defined at integer values of k. $X < -X_0$ is defined as domain co-occurrences, while $X > X_0$ is defined as domain avoidance. For more details, see the main text. Figure modified from (Triola, 2003), Slide number 16 and 17.

too.
7.1 Definition of Domain Co-occurrence

A. Human AUGUSTUS - SUPERFAMILY randomized annotation

B. Human AUGUSTUS - PFAM randomized annotation

Figure 7.2: Human Genome Randomized Annotation. On A and B, the functional annotation on human gene predictions was randomized and reassigned. It was found that there is no significant result at all in the histogram, due to the significance level above 10%.
7.2 Tendency of Domain Distribution

As an application, we considered seven major classes of DNA-binding domains of TFs: zinc-finger (znf), leucine-zipper, winged-helix, bromo, brct, krab and hmg-box (hmg). Znf, leucine-zipper, winged-helix, and hmg are DNA-binding domains. It was found that different types of DNA-binding domains systematically avoid each other throughout the evolution of eukaryotes. In contrast, DNA binding domains belonging to the same superfamily readily co-occur in the same protein. The domain co-occurrence was determined for znf with other non-DNA-binding domains, namely wd40, phd, ring, and tpr. In these cases, high numbers of co-occurrences was expected but observed significantly fewer than expected. This also indicates avoidance (Parikesit et al., 2011a). A systematic analysis of co-occurrences and potential reasons for avoidance will be presented. Based on the published methodology, more domain co-occurrences for significant and biologically meaningful avoidance were investigated. Domain co-occurrences computed from the de novo predictions (GP) were compared with domain co-occurrences recorded in the SUPERFAMILY database (Wilson et al., 2009) (SF) for the 18 species.

Typically, studies of this type are based on existing protein annotations derived primarily from genomic sequence data. Popular data sources are, e.g., the protein annotation compiled in KEGG or ENSEMBL. Annotated protein domains from PFAM (Finn et al., 2006) were used in (Itoh et al., 2007). The studies (Yang and Bourne, 2009; Prohaska et al., 2010) are based on the SUPERFAMILY database (Wilson et al., 2009), whose HMM models in turn are based on the SCOP (Structural Classification of Proteins) domain definitions (Andreeva et al., 2008). Both the protein annotation and the collections of domains, however, suffer from substantial biases (Parikesit et al., 2010a):

- The knowledge of protein domains is by far not complete, although most protein domains in well-studied model organisms are evolutionarily very old, while innovation of protein domains at the same time is a relatively infrequent phenomenon (Bornberg-Bauer et al., 2010; Zmasek and Godzik, 2011). The majority of "plant-specific" DNA binding domains, for instance, originated much earlier then the comparably recent expansion into the diverse gene families present in higher plants (Yamasaki et al., 2008). Unrecognized domains thus have to be attributed in many cases to insufficient sensitivity of the domain annotation procedure. Non-globular segments of proteins, in particular trans-membrane regions and signal peptides, furthermore have a hydrophobic bias leading to problematic domain models and subsequently to completely wrong function assignments inherited from these domain models (Wong et al., 2010).

- Domains are typically annotated on protein sequences stored in sequence databases. These "protein models" in turn are the result of computational procedures that combine the genomic DNA sequence, EST and cDNA data, and homology-based
7.2 Tendency of Domain Distribution

predictions. Differences in the amount of available experimental evidence can lead to dramatic ascertainment biases. The number of annotated domains in SUPERFAMILY 1.73, for example varies by more than a factor of four within eutherian mammals (64,225 domains in human versus 14,748 in the alpaca) although one would expect these species to have a very similar gene complement.

The data in Table 7.1 shows a good overall correlation between the domain counts as reported by the SUPERFAMILY database and those computed from the GENSCAN predictions, although counts can deviate largely in some species. For instance, in Trypanosoma brucei 148 zinc-fingers was detected using gene predictions compared to only 6 annotated in SUPERFAMILY annotation. Investigation was also conducted for two additional families of DNA binding domains, namely the leucine zippers (SUPERFAMILY ID 57979) and the "high mobility group" (HMG) domains (SUPERFAMILY ID 47095). It was observed of only very few candidate co-occurrences with other DNA binding domains in the species listed in Table 7.1 (the co-occurrences between leucine-zipper and winged-helix and one between HMG and winged-helix). Inspection of these five cases revealed that four of them are clear artifacts of GENSCAN, which predicts a fusion protein. The last candidate, human LARP1B, is predicted by GENSCAN to have an additional internal exon containing a leucine-zipper domain. More likely, however, GENSCAN stumbled across a retro-pseudogene deriving from FOSL1 located in an intron of LARP1B. Conversely, SUPERFAMILY reports the co-occurrence of leucine-zipper and zinc-finger in some isoforms of the paralogous human ATF2 and ATF7 genes, which are not found in our genscan-based approach (Parikesit et al., 2010b). As an application of genome-wide domain counts, the co-occurrences of four major types of DNA binding domains (zinc-fingers, leucine-zipper, HMG-box domains, and winged-helix domains) was investigated. It was found a strong and statistically highly significant anti-correlation of the four different domains. In contrast, evolutionarily related DNA binding domains readily co-occur in DNA binding proteins. It will be interesting to investigate whether a similar avoidance can be observed among other evolutionarily unrelated protein domains that share a common molecular function (Parikesit et al., 2010b). In the next part, the comparison between the empirical observations of gene-predicted domain annotations from the ADD pipeline (GP) and the existing SUPERFAMILY annotation (SF) will be posed. The expectations of GP \( E(GP) \) and SF \( E(SF) \) will be used to determine the significance of the empirical observation. After mapping efforts of GO terms 'Transcription factor activator' into SUPERFAMILY database, following Kirsten (2007), It was found 300 TFs SUPERFAMILY ID. The complete domain list and query data could be found in the online supplementary material. The results of domain co-occurrences and avoidance tendency will be presented in the following section. The legends of 18 species are as follow:
## Protein Domain Distribution

<table>
<thead>
<tr>
<th>Legend:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G.la= Giardia lamblia</td>
<td>C.re= Chlamydomonas</td>
</tr>
<tr>
<td>T.va= Trichomonas vaginalis</td>
<td>A.th= Arabidopsis thaliana</td>
</tr>
<tr>
<td>T.br= Trypanosoma brucei</td>
<td>O.sa= Oryza sativa</td>
</tr>
<tr>
<td>L.ma= Leishmania major</td>
<td>D.di= Dictyostelium</td>
</tr>
<tr>
<td>N.gr= Naegleria gluberi</td>
<td>A.ni= Aspergillus niger</td>
</tr>
<tr>
<td>P.fa= Plasmodium falciparum</td>
<td>S.po= Schizosaccharomyces pombe</td>
</tr>
<tr>
<td>T.th= Tetrahymena</td>
<td>C.el= Caenorhabditis elegans</td>
</tr>
<tr>
<td>T.ps= Thalassiosira pseudonana</td>
<td>D.me= Drosophila melanogaster</td>
</tr>
<tr>
<td>P.ra= Phytophthora ramorum</td>
<td>H.sa= Homo sapiens</td>
</tr>
</tbody>
</table>

The histograms of the computed domain pairs are shown in Figure 7.3. It is mainly dominated with the domain pair of Zinc Finger domain, with the exception of Bromo-Phd pair. In Figure 7.3c, there is a tendency of co-occurrence in bromo-Phd pair. The bromodomain (SUPERFAMILY ID 47370) is a protein domain that recognizes acetylated lysine residues such as those on the N-terminal tails of histones. This recognition is often a prerequisite for protein-histone association and chromatin remodeling. The PHD finger (Plant Homeo Domain) was discovered in 1993 as a \( Cys_4 - His - Cys_3 \) motif in the homeodomain protein HAT3 in Arabidopsis thaliana (SUPERFAMILY ID 57903). There is a tendency of co-occurrence in bromo-Phd pair. It shows up primarily in Giardia lamblia (SF and GP), Drosophila melanogaster (SF and GP), Oryza sativa (GP), Caenorhabditis elegans (SF), Dictyostelium (SF), Phytophthora ramorum (SF), Thalassiosira pseudonana (SF), and Chlamydomonas (GP). The efficacy of GENSCAN prediction is underestimated in Caenorhabditis elegans (SF), Dictyostelium (SF), Phytophthora ramorum (SF), and Thalassiosira pseudonana (SF) because they have SF entries and no or few GP co-occurrences.

### 7.2.1 The Tendency of Zinc Finger Domain Co-occurrences

To investigate the suitability of gene predictions for the assessment of domain co-occurrences, two very abundant classes of DNA binding domains were selected: ZNF and winged-helix domains. If the two human domain types were distributed randomly, it would be expected to count about 17.8 co-occurrences, estimated from the data in the SUPERFAMILY (30712 transcripts, of which 1324 contain a ZNF domain and 414 have a winged-helix domain using equation (7.1)). Surprisingly, not a single co-occurrence between these two domains is observed in the SUPERFAMILY data from the existing annotation and AUGUSTUS - SUPERFAMILY data in any species, even though both domains are conserved throughout the Eukaryota, (Table 7.1).
7.2 Tendency of Domain Distribution

Figure 7.3: The tendency of domain co-occurrences and avoidance of GENSCAN -SUPERFAMILY annotation. The $x$ axis is the utilized genome, while the $y$ axis is the total number of domain co-occurrences. (a) The znf-ring pairs showed a strong tendency of avoidance. (b) The Znf-tpr pairs showed a strong tendency of avoidance. (c) There is a tendency of co-occurrence in bromo-phd pair. (d) There is a tendency of co-occurrence in krab-znf pair. (e) The bromo-znf pairs showed a tendency of co-occurrences. (f) The znf-wd40 pairs showed a tendency of co-occurrence. (g) The znf-wing pairs showed a tendency of co-occurrence. More information about a,b,d-g is explained in the main text at the Section 7.2.1 and 7.2.1.
Table 7.1: Domain occurrences and co-occurrences of ZNF and winged-helix domains. The table shows the number of domains (Dom.), the number of "genes", i.e., (g) is for GENSCAN and (a) is for AUGUSTUS predictions that contain the domain (Genes), and for comparison the number of genes that contain the domain in SUPERFAMILY (SF). For species marked with *, multiple entries from different strains or variants in the SUPERFAMILY database exist, and SF values tend to over-count in these cases.

<table>
<thead>
<tr>
<th>Species</th>
<th>Dom. (g)</th>
<th>Genes (g)</th>
<th>Dom. (a)</th>
<th>Genes (a)</th>
<th>SF</th>
<th>Dom. (g)</th>
<th>Genes (g)</th>
<th>Dom. (a)</th>
<th>Genes (a)</th>
<th>SF</th>
<th>Genes (g)</th>
<th>Genes (g)</th>
<th>Genes (a)</th>
<th>Genes (a)</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giardia lamblia</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>16</td>
<td>13</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Trichomonas vaginalis</td>
<td>23</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>100</td>
<td>98</td>
<td>7</td>
<td>7</td>
<td>89</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Trypanosoma brucei</td>
<td>156</td>
<td>148</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>34</td>
<td>32</td>
<td>12</td>
<td>12</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leishmania major</td>
<td>29</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>50</td>
<td>27</td>
<td>12</td>
<td>11</td>
<td>23</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Naegleria gruberi</td>
<td>20</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>67</td>
<td>45</td>
<td>34</td>
<td>33</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tetrahymena</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thalassiosira pseudonana</td>
<td>15</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>145</td>
<td>138</td>
<td>104</td>
<td>101</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clamydomonas</td>
<td>81</td>
<td>46</td>
<td>15</td>
<td>15</td>
<td>34</td>
<td>80</td>
<td>75</td>
<td>50</td>
<td>47</td>
<td>62</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>18</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>48</td>
<td>44</td>
<td>20</td>
<td>20</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>284</td>
<td>224</td>
<td>43</td>
<td>43</td>
<td>307</td>
<td>151</td>
<td>146</td>
<td>112</td>
<td>110</td>
<td>443</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dicyostelium</td>
<td>21</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>12</td>
<td>42</td>
<td>37</td>
<td>25</td>
<td>25</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aspergillus niger</td>
<td>64</td>
<td>51</td>
<td>14</td>
<td>14</td>
<td>34</td>
<td>68</td>
<td>65</td>
<td>44</td>
<td>43</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe</td>
<td>34</td>
<td>24</td>
<td>11</td>
<td>11</td>
<td>38</td>
<td>43</td>
<td>41</td>
<td>26</td>
<td>26</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>58</td>
<td>27</td>
<td>36</td>
<td>36</td>
<td>144</td>
<td>15</td>
<td>14</td>
<td>68</td>
<td>66</td>
<td>165</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>853</td>
<td>301</td>
<td>53</td>
<td>53</td>
<td>322</td>
<td>126</td>
<td>122</td>
<td>96</td>
<td>94</td>
<td>152</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>5090</td>
<td>1048</td>
<td>447</td>
<td>387</td>
<td>1324</td>
<td>274</td>
<td>256</td>
<td>182</td>
<td>172</td>
<td>414</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
In the *GENSCAN*-based analysis, it was detected co-occurrences of ZNF and winged-helix domains only in the clades Kinetoplastida (*Leishmania* and *Trypanosoma*) and in *Phytophthora*. Upon closer inspection, these can be identified as artifacts. The exclusive usage of one of the two types of DNA binding domains is statistically highly significant. In human, for instance, it was expected 11.7 co-occurrences (5090 ZNF and 274 winged-helix domains in 118894 *GENSCAN* predictions) while none is observed ($p < 10^{-5}$). This indicates a selective pressure against their co-occurrences.

Therefore, it was concluded that the major types of DNA binding domains, and possibly other evolutionarily unrelated domains of similar function, strongly avoid each other in Eukarya. In contrast, domains with complementary functions readily co-occur with each other. A good example are zinc-fingers and the "Krüppel associated box" (KRAB) domain. The KRAB domain is a small (75 AA) protein domain (*SUPERFAMILY* ID 109640) that functions as a transcriptional repressor and is predicted to act via protein-protein interactions. It appears in a highly prolific family of evolutionarily very young transcription factors. Among the species listed in Table 7.1, it appears only in human. Moreover, 446 domains in 421 "genes" were detected, in agreement with the literature (Nowick *et al.*, 2010). In contrast to the winged-helix domain, however, it readily combines with zinc-finger domains: 351 *GENSCAN* predictions (i.e., a third) of the 1048 ZNF proteins and 5/6 of the KRAB domain proteins belong to the KRAB-ZNF family, again in good agreement with the literature (Nowick *et al.*, 2010).

It was emphasized that it is impossible in practice to devise a fair benchmark for domain co-occurrence counts since the ground truth depends on the complete knowledge of all transcripts, even if one settles for the definition that two particular protein domains co-occur if they appear together in at least one protein-coding transcript. Therefore, it is better for resorting to compare counts between closely related species for which we can plausibly expect to obtain similar numbers (Parikesit *et al.*, 2010b).

In easy cases, such as yeast, where the transcript structure is simple and data coverage is excellent, gene prediction and transcript annotation yield nearly identical results. For large mammalian genomes, on the other hand, estimates of domain numbers depend strongly on transcript coverage, while gene predictions yield numbers that are consistent among closely related species (Parikesit *et al.*, 2010b). The investigation suggests that the biases and artifacts in the *GENSCAN* predictions are small compared to the numerous problems of annotation-based approaches. In particular, a very small number of false positive co-occurrences was arising from the incorporation of additional introns and the erroneous prediction of fusion proteins (Parikesit *et al.*, 2010b).
Zinc Finger Domain Avoidance

After testing the available domain pair combination, only Znf-ring and Znf-Tpr shows significant avoidance in at least one species. The Figures in this section and the following one will present the extent of the avoidances and co-occurrences significance. In many case, it was observed systematically fewer domain co-occurrences than expected, i.e., there is a selection pressure causing the domains to ‘avoid’ each other. In fact, this is the case with most — but not all — combinations of distinct DNA binding domains. In Oryza sativa $E(GP) \ll E(SF)$, because SF has more annotated individual domain than GP (Parikesit et al., 2011a). Both Figures 7.3a and b expose a scarcity in the GP and SF domain annotation. Only human yields an abundant domain avoidance. The detailed observation on the data is shown in their respective figures. In Figure 7.3a, the znf-ring pairs showed a strong tendency of avoidance. Znf is the zinc finger protein C2H2 (SUPERFAMILY ID 57667). In molecular biology, a RING (Really Interesting New Gene) finger domain is a protein structural domain of zinc finger type which contains a $\text{Cys}_3\text{HisCys}_4$ amino acid motif which binds two zinc cations (SUPERFAMILY ID 57850). The znf-ring pairs showed a strong tendency of avoidance in the Homo sapiens (SF). In Figure 7.3b, the Znf-tpr pairs showed a strong tendency of avoidance. Tetratricopeptide-like repeats (tpr) are found in a numerous and diverse proteins involved in such functions as cell cycle regulation, transcriptional control, mitochondrial and peroxisomal protein transport, neurogenesis and protein folding (SUPERFAMILY ID 48453). The znf-tpr pairs showed a strong tendency of avoidance. The efficacy of GENSCAN prediction is hampered in Homo sapiens (SF) because it has SF entries and fewer GP co-occurrences.

Zinc Finger Domain Co-occurrences

After testing the available TF domain pair combination, only Krab-Znf, Bromo-Znf, Znf-Wd40, and Znf-Wd40 show significant co-occurrence in at least one species. In some cases, however, a positive correlation between distinct DNA binding domains is observed. A well-studied example is the co-occurrence of KRAB domain and ZNF domains in a large group of primate-specific transcription factors (Nowick et al., 2010; Parikesit et al., 2010b). The Figures 7.3c, f, and g show that only the human genome has clear resolution of domain Co-occurrences. In Figure 7.3d, there is a tendency of co-occurrence in krab-znf pair. KRAB domain is a category of transcriptional repression domains present in approximately 200 human zinc finger protein-based transcription factors (SUPERFAMILY ID 109640). There is a tendency of co-occurrence in krab-znf pair only in Homo sapiens (SF and GP). Krab-znf co-occurrence are happening exhaustively in Homo sapiens (Nowick et al., 2010) In Figure 7.3e, the bromo-znf pairs showed a tendency of co-occurrences shown in Caenorhabditis elegans (GP). The hypothetical protein existence in Caenorhabditis elegans was found. That’s
due it has no SF co-occurrence, and abundant GP co-occurrences. In Figure 7.3f, the \textit{znf-wd40} pairs showed a tendency of co-occurrence. The \textit{wd40} domain is a short structural motif of approximately 40 amino acids, often terminating in a tryptophan-aspartic acid (W-D) dipeptide (\textit{SUPERFAMILY} ID 50978). The \textit{znf-wd40} pairs showed a tendency of co-occurrence in \textit{Leishmania major} (GP). In Figure 7.3g, the \textit{znf-wing} pairs showed a tendency of co-occurrence. The winged helix-turn-helix (wHTH) motif is formed by a 3-helical bundle and a 3- or 4-strand beta-sheet (wing), and this structure is capable to bind DNA (\textit{SUPERFAMILY} ID 46785). It is shown in \textit{Phytophthora ramorum} (GP) and \textit{Leishmania major} (GP). Information about gene fragments existence are already mentioned in Section 7.2.

### 7.2.2 Limitation of The tendency of Domain Distribution

Although a plethora of annotation data is available in publicly accessible databases for most of the published genomes, quantitative comparisons remain difficult. Consequently, comparative studies typically resort to testing for relative enrichment rather than considering absolute numbers of domains. In studies focusing on the evolution of regulatory mechanisms and regulatory complexity, however, absolute gene counts play an important role. The previous investigations suggested that the biases and artifacts, caused by \textit{de novo} gene prediction methods such as \textit{GENSCAN}, are small compared to the numerous problems of annotation-based approaches. In particular, a very small number of false positive co-occurrences was arising from the incorporation of additional introns and the erroneous prediction of fusion proteins (Parikesit \textit{et al.}, 2011a; Nielsen, 1999). The combination of \textit{de novo} gene predictors and subsequent HMM-based annotation of \textit{SCOP} domains in the predicted peptides leads to consistent estimates with acceptable accuracy that in particular can be utilized for systematic studies of the evolution of protein domain occurrences and co-occurrences (Parikesit \textit{et al.}, 2011a).

Protein domains are not randomly combined in functional proteins. Statistically significant avoidance was observed, if the TF domain paired with other non DNA-Binders (RING and TPR domains). On the other hand, it was found more co-occurrences than expected for certain combinations of TF and non-TF domains (e.g. bromo-phd), between distinct types of TF domains (e.g. in the combinations bromo-znf and znf-wing) as well as for combinations of DNA binding domains (e.g. krab-znf). The general trends are in most cases detected consistently based on \textit{de novo} genome predictions (GP) and from \textit{SUPERFAMILY} annotation databases (SF).

Avoidance and preferential co-occurrence, however, are only observable in genomes with sufficiently large numbers of proteins, in particular multicellular plants and animals. In most species with small genomes the expected numbers of domain co-occurrences is already below one so that a selection pressure for domain avoidance cannot be detected.
7.3 The Limitation of Individual Domain Pairs Annotation

In (Parikesit et al., 2010a), therefore, it was proposed to bypass existing genome annotations and to estimate domain occurrence data by combining de novo gene prediction with HMM-based domain annotation of the predicted protein structures instead. It was found that the number of domains found by this procedure correlates very well with the annotation compiled in the SUPERFAMILY database for both human and yeast. Furthermore, consistent estimates are obtained for closely related species such as the apes. This implies that cross-species comparisons are more meaningful when using a consistent de novo annotation by Arli’s Domain Distribution (ADD) pipeline than based on currently available protein databases. It was noted that false positives of the gene prediction step are not much of a problem for our purposes since the predicted amino acid sequences do not match valid protein domain models. False negatives, on the other hand, affect our results. Hence, completeness was traded for a relatively unbiased annotation so that estimates of domain content are consistent between different genomes. Taken together, this allows quantitative comparisons of domain-occurrences and co-occurrences at least at a statistical level.

As a first application of this approach, the investigation on the co-occurrences of four major types of DNA binding domains (zinc fingers, leucine-zipper, HMG-box domains, and winged-helix domains) was conducted and it was observed a strong and statistically highly significant anti-correlation of the four different domains. In contrast, evolutionarily related DNA binding domains readily co-occur in DNA binding proteins (Parikesit et al., 2011a). In many genomes, in particular in the rather compact genomes of simple unicellular eukaryotes, however, the total number of genes and domains that can be annotated is too small for a meaningful statistical evaluation. Several combinations of protein domains show specific tendencies to either systematically avoid each other or to co-occur preferentially in proteins. In the examples studied so far, avoidance appears to be conserved among those major Eukaryotic clades where the effect is detectable. Signals for preferential co-occurrence can arise from recent proliferation by gene duplication as in the case of the primate-specific krab-znf family of transcription factors (Nowick et al., 2010; Parikesit et al., 2010b).

The limitation was addressed by pooling related domain models derived from data for different phylogenetic groups, albeit at the expense of losing resolution regarding structural and functional differences among domains belonging to the same family or superfamily. There does not seem to be an easy remedy for the ascertainment biases when currently available databases are used.
CHAPTER 8

Analysis of Functional Annotation

This chapter is elaborating on the analysis of functional annotation which successfully gives a fine-grained data resolution for the domain annotation. Integration of the GO database which annotates the functional annotation of genes/proteins, and domain databases is crucial to standardize the domain annotation procedure in Arli’s Domain Distribution (ADD) pipeline.

8.1 Pooling the Domain Model

Domain databases contain thousands of distinct domain models. Few domains thus appear a sufficiently large number of times to allow for a quantitative statistical analysis of their occurrences. The HMMER 3.0rc1 package of Finn et al. (2011) was used to map the HMMs to the protein sequences with the cut-off \( E \leq 10^{-3} \). Thus, the data was pooled by functional categories (Parikesit et al., 2011c). Version 1.75 of the SUPERFAMILY database offers a ”Structural Domain Functional Ontology” providing functional and phenotypic annotations of protein domains at the superfamily and family levels (de Lima Morais et al., 2011).

Since any protein can be annotated with multiple functions, it is clear that membership in GO annotation classes does lead to a partition of the set of protein domains into functional groups. The Pfam annotation is already integrated into InterPro2GO database, providing a mapping from Pfam domains to GO ontology terms (The Gene Ontology Consortium, 2000; Punta et al., 2012).

In this work the following seven functional groups were used:

- bN \( \text{binding of nucleic acids} \): GO:0003676 at superfamily level.
bP binding of proteins with potential nuclear localization: GO:0005515 superfamily level.

rC regulation of chromatin GO:0016568 at superfamily level.

rC* regulation of chromatin as determined in (Prohaska et al., 2010), comprising a combination of family and superfamily level. This set only generated for SUPERFAMILY annotation.

rB regulation of binding: GO:0051098 at superfamily level.

rE regulators of enzymatic activity: GO:0050790 at superfamily level.

mS metabolism of saccharides: GO:0005976 at superfamily level.

The four functional groups bN, bP, rC, and rB encapsulate major modes of regulation. Both bN and bP play an important role for gene regulation by transcription factors and are among the most abundant GO classes, while rC focuses on chromatin-based epigenetic regulation. The choice of the two variants of chromatin-associated domains rC and rC* is motivated by the previous work on the co-occurrence of protein domains that can act as readers, writers, and erasers of histone modification (Prohaska et al., 2010), which revealed changes in the co-occurrence patterns within this group. The domain groups rE and mS were intended as a form of controls that a priori and are not expected to correlate in a particular way with either nucleic acid or protein binding domains (bN, bP).

8.2 Test of Domain Co-occurrences

The membership of a domain in the functional groups was annotated. From the co-occurrences of domains in predicted proteins and the map of domains to functional GO classes it is straightforward to obtain the number $n(C, D)$ of co-occurrences of the functional classes. The corrected formulation for the domain pooling into functional classes is shown in Equation 8.1 and 8.2.

$$n(C) = \sum_{x \in C} n_x$$  \hspace{1cm} (8.1)

$$n(D) = \sum_{x \in D} n_x$$  \hspace{1cm} (8.2)

$n(C, D)$ is the empirical observance of Domain co-occurrences of $C$ and $D$ and defined in Equation 8.3.

$$n(C, D) = \sum_{x \in C} \sum_{y \in D} n_{xy}$$  \hspace{1cm} (8.3)

In Equation 8.4, The $n(C, D)$ was corrected for the fact that the same domain $x$ can be a member of both $C$ and $D$ by counting these cases with a weight of 1/2.
8.2 Test of Domain Co-occurrences

\[ \text{if } \text{n}(C, D) = X \text{ then } \text{n}(C, D) = \frac{X}{2} \]  

(8.4)

Table 8.1: Overlaps between the 7 functional groups defined in the text. This is the overlapping of all \textit{SUPERFAMILY} domains of a functional groups with domains from the other groups. The overlapping of the \textit{SUPERFAMILY} domain is exposed in the count.

<table>
<thead>
<tr>
<th></th>
<th>bN</th>
<th>bP</th>
<th>rC</th>
<th>rC*</th>
<th>mS</th>
<th>rB</th>
<th>rE</th>
</tr>
</thead>
<tbody>
<tr>
<td>bN</td>
<td>112</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>bP</td>
<td>4</td>
<td>118</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>rC</td>
<td>4</td>
<td>6</td>
<td>25</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>rC*</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>27</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>mS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rB</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>rE</td>
<td>6</td>
<td>21</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 8.1 shows the overlapping of \textit{SUPERFAMILY} domains between functional groups. Moreover, the overlapping between rE and bP does not signify a correlation in domain distribution. It is just showing whether each individual domain ID in one functional group is available in the others. Pooling the domain model into designated GO terms will eventually improve the statistical power of the domain annotation.

8.2.1 Performance of \textsc{Genscan} - \textit{SUPERFAMILY} Annotation

The results of the co-occurrence analysis at the level of \textit{GO} classes are summarized in Figure 8.1 for the complete set of domains. Some interesting global patterns were observed. With the exception of the functional classes rE (regulation of enzymatic activity) and rC (regulation of chromatin in a narrow sense) there is no pattern of conserved avoidance. In fact, most other combinations of domain functions are at least weakly positively correlated. However, it is found, that among the multicellular organisms, \textit{Oryza} has the most domain pair avoidance tendencies, while human has the least tendency.

With respect to the phylogenetic distribution of co-occurrence patterns, the most interesting observation is a trend towards wide-spread avoidance in particular in multicellular plants, and — to a lesser extent — also in animals. Among unicellular species, only \textit{Trichomonas} and \textit{Phytophthora} show similar patterns of functional avoidance. The lack of significant signals is at least in part explained by the small number of proteins that can be annotated by \textit{ab initio} methods.

In Figure 8.1, it is shown as well, that \textit{Plasmodium} could not yield any co-occurrence significance, so it is assumed that there are no co-occurrences within its genome. This could
be due to the inability of GENSCAN to annotate the extreme A + T rich content of Plasmodium. Its high gene prediction count could be the reason of the massive diversification of domain functions that led to a strong tendency of avoidance. Zinc finger proteins are one of the largest single classes of proteins (Klug, 2010). It is also known, that zinc finger proteins have very diverse functional features such as DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding (Laity et al., 2001). The recent work on zinc finger engineering is also known to have potential diagnostic and therapeutic features, due to its ability to bind human telomeric DNA (Laity et al., 2001).
Figure 8.1: Summary of domain co-occurrences in 18 eukaryote genomes. The observed empirical domain co-occurrences were computed with Equation 8.1 and 8.2, and corrected with Equation 8.4. Colors indicate the statistical significance of co-occurrence $n(C, D) \gg E(C|D)$ (red) and of avoidance $n(C, D) \ll E(C|D)$ (blue). Significance levels on individual comparisons are shown in three levels of color saturation for $p < 0.001$, $0.001 \leq p < 0.01$, and $0.01 \leq p < 0.1$, respectively. Abbreviations: bN: binding of nucleic acids; bP: binding of proteins with potential nuclear localization; rC: regulation of chromatin; rC*: regulation of chromatin as determined in (Prohaska et al., 2010), comprising a combination of family and superfamily level; rB: regulation of binding; rE: regulators of enzymatic activity; mS: metabolism of saccharides; H.sa: Homo sapiens; D.me: Drosophila melanogaster; C.el: Caenorhabditis elegans; S.po: Schizosaccharomyces pombe; A.ni: Aspergillus niger; D.di: Dictyostelium discoideum; O.sa: Oryza sativa; A.th: Arabidopsis thaliana; C.re: Chlamydomonas reinhardtii; P.ra: Phytophthora ramorum; T.ps: Thalassiosira pseudonana; T.th: Tetrahymena thermophila; P.fa: Plasmodium falciparum; N.gr: Naegleria gruberi; L.ma: Leishmania major; T.br: Trypanosoma brucei; T.va: Trichomonas vaginalis; G.la: Giardia lamblia

Figures 8.1 shows that the mutual relationships of a many, but certainly not all, GO classes are observed coherently across the major groups of Eukarya. Due to the large differences in genome size and domain numbers it makes little sense to compute a summary statistic by...
adding up the counts of occurrences across species: such data would be dominated by the large,
gene-rich multicellular organisms. Instead we employed a simple voting procedure, associating
scores of $+3$, $+1$, $-1$, and $-3$ only with the two most significant levels of co-occurrence and
avoidance, respectively. Figure 8.2 displays these scores averaged over the 18 species for all
the proteins.

![Figure 8.2: Summary of co-occurrence data. See text for details.](image)

It was found that most of the domain GO-classes are at least weakly positively correlated,
in part reflecting the fact that the protein domains can have promiscuous functions, in
part possibly also because the domains investigated here are mostly involved in binding and regulatory processes. Surprisingly, the only combination that shows strong avoidance across all data sets is regulation of chromatin vs. regulation of catalytic activity (rC:rE). This effect is not visible in comparison to the set rC* of domains associated with chromatin-regulation. The latter in particular contains also enzymatic domains such as kinases and phosphatases involved in chemical modifications of histones (Prohaska et al., 2010).

To our surprise, there is no systemic anti-correlation of the domains involved in saccharide metabolism (mS) and regulation of enzymatic activity (rE), respectively, with the binding and chromatin associated domains. For the mS group, correlations with functional classes are weak, while rE co-occurs readily with binding domains but avoids the core set of chromatin associated domains (rC). In retrospect, the positive correlation of rE and bP makes sense as regulators of enzymatic activity have reason to bind to enzymes. This also explains the co-occurrence with the rC* set, which contains in particular also histone modifying enzymes. We have at present no good explanation, however, why there is co-occurrence with nucleic acid binding.

### 8.2.2 Comparison of SUPERFAMILY - PFAM Annotation

In Figure 8.3 we observe a systematic avoidance of functionally distinct GO-classes of protein domains in AUGUSTUS annotation. Satisfactorily, the patterns obtained from PFAM and
SUPERFAMILY annotations are largely consistent. Not surprisingly, we find fewer significant relations in the SUPERFAMILY data due the much smaller number of domains.

However, there are some interesting facts that deserve to be mentioned. The abundant positive correlation in both the SUPERFAMILY and PFAM annotation is a reversed trend compared to Section 8.2.1. In the latter Section, the GENSCAN annotation is dominated by negative correlation while that is not the case for AUGUSTUS. It is assumed, that this could happen due to the different gene models utilized by both gene prediction tools. AUGUSTUS is assumed to have a better predictor, due to its customized training models for each different species. Moreover, it could be assured, that the positive correlation abundance in AUGUSTUS is a more valid data representation. The domain co-occurrences’ scarcity is, surprisingly, present as well in some other genomes. In the SUPERFAMILY annotation, this happens with the Tetrahymena genome. The scarcity of this genome is present as well in the PFAM annotation, but to a more limited extend. It is assumed, that this is due to the short scaffold in the Tetrahymena genome that could lead into underestimates.

As shown in Figure 8.3, the main exceptions for the significant relations are the co-occurrences bN-rB, rC-rB, and bP-rE. The latter is not unexpected, since regulators of enzymatic activity (rE) can be expected to act by protein-protein binding (bP). The positive correlations between nucleic acid binding domains (bN) and chromatin associated domains (rC) with domains involved in the regulation of binding deserve further investigation. It is consistent with the intimate link of both DNA and RNA binding with chromatin regulation reported in Prohaska et al. (2010).

An interesting phenomenon is also observed for the human genome of both SUPERFAMILY and PFAM annotations mode. The strong co-occurrences of bN-rB, rC-rB, and bP-rE are missing, and replaced with a strong tendency of avoidance. This could be due to the complicated nature of human transcription factor and chromatin regulation, that causes the complex diversification of domain function, that resulted in domain avoidance. However, there are noticeable pattern differences in PFAM and SUPERFAMILY annotation. In PFAM annotation, it could be seen that the mS-rB pattern is totally missing while in the SUPERFAMILY annotation, it is clear that the pattern is a co-occurrence. At present, there is no clear explanation for the cause of this phenomenon. Moreover, in PFAM table, it is clear that for human genome there is not even a single red cell and this signifies the lack of tendency of co-occurrence. This case is not observable for SUPERFAMILY due to the weak tendency of avoidance in mS-rB. The explanation could be that PFAM annotation definitely shows a clear domain diversification for complex functional features as the strong tendency of avoidance over-ruled the co-occurrences.

Then, for the average summary of co-occurrence data on the left side of Figure 8.3, it is clearly shown that the tendency of the positive correlation is confirmed. However, it is seen as well, that the positive correlation is weaker for SUPERFAMILY compared to PFAM
Figure 8.3: Summary of AUGUSTUS co-occurrences patterns of major functional classes of protein domains across the Eukaryotes. The estimated obtained from Pfam-domains (l.h.s.) are qualitatively consistent with those from SUPERFAMILY-domains (r.h.s). The left panels shows the data separately for each species, the smaller panels in the right summarize the co-occurrence patterns across the 18 species. Blue rectangles indicate statistically significant avoidance between functional classes of protein domains, red indicates co-occurrence. The saturation of the color denotes the significance levels $p < 0.001$ (saturated color), $0.001 \leq p < 0.01$ (intermediate), and $0.01 \leq p < 0.1$ (pale). Entries that show neither avoidance or co-occurrence at a significance level of at least 10% remain white. Abbreviations : bN: binding of nucleic acids; bP: binding of proteins with potential nuclear localization; rC: regulation of chromatin; rB: regulation of binding; rE: regulators of enzymatic activity; mS: metabolism of saccharides; H.sa: Homo sapiens; D.me: Drosophila melanogaster ; C.el: Caenorhabditis elegans ; S.po: Schizosaccharomyces pombe; A.ni: Aspergillus niger; D.di: Dictyostelium discoideum ; O.sa: Oryza sativa ; A.th: Arabidopsis thaliana; C.re: Chlamydomonas reinhardtii; P.ra: Phytophthora ramorum; T.ps: Thalassiosira pseudonana; T.th: Tetrahymena thermophila; P.fa: Plasmodium falciparum; N.gr: Naegleria gruberi; L.ma: Leishmania major ; T.br: Trypanosoma brucei ; T.va: Trichomonas vaginalis ; G.la: Giardia lamblia

8.2.3 Comparison of RefSeq SUPERFAMILY - PFAM Annotation

The RefSeq annotation is utilized as a comparison with the AUGUSTUS annotation in Section 8.2.2. Figure 8.4 shows the co-occurrence patterns of the eukaryotes. It could be seen that the functional annotation pair of mS-rB is missing. It is also observed that the human genome has more insignificant functional annotation pairs in the RefSeq than the AUGUSTUS annotation. Moreover, the tendency of avoidance in the RefSeq annotation of the
Figure 8.4: Summary of RefSeq co-occurrences patterns of major functional classes of protein domains across eukaryotes. The estimated patterns obtained from Pfam-domains (l.h.s.) are qualitatively consistent with those from SUPERFAMILY-domains (r.h.s).

rice genome is strongly observed compared with the AUGUSTUS annotation.

In the Schizosaccharomyces pombe genome, only a weak tendency of co-occurrence is observable with intervals of insignificant functional annotations. Moreover, the RefSeq annotation detected more insignificant tendency in Trichomonas, and the tendency of avoidance for AUGUSTUS is not detected. There are stronger avoidance tendencies in the RefSeq annotation of Arabidopsis than in the AUGUSTUS annotation.

Interestingly, the tendency of functional annotation pair conservation in RefSeq is not as strong as in AUGUSTUS. There is a weak avoidance tendency in bN-bP, and bN-rE. There are weak tendencies of co-occurrence in bN-rB, bP-rC, and bP-rE. The rE-rB patterns in SUPERFAMILY are almost lost, and in the bP-rB, many patterns are already lost. rC-bP pair in AUGUSTUS -PFAM annotation shows strong avoidance, while in RefSeq it shows co-occurrence. rC-bN pair in AUGUSTUS -PFAM annotation shows strong avoidance, while in RefSeq it has no significance co-occurrence. This argue for a discrepancy between AUGUSTUS and RefSeq annotations.

8.3 Zinc Finger Protein Annotation

In Figure 8.5, the investigation was conducted in order to know to what extent the occurrence and co-occurrence of other domains is influenced by the additional presence of a zinc finger domain. Surprisingly, it was found that patterns of positive or negative correlation among
domain functions are enhanced within zinc finger proteins. In fact, also much more significant deviations from the expectation were found even though the sample size was much smaller than in the functional annotation pairs analysis (Figure 8.1). In particular, it was observed that domain avoidance is most common within multicellular organisms, where they affect in particular the two groups of nucleic acid and protein binding domains. It was suspected that this statistical pattern derives from recent rapid expansions of certain protein families. An example would be the mammalian-specific KRAB-ZNF protein comprising hundreds of closely related transcription factors (Nowick et al., 2010).

In Figure 8.5, it is shown that several organisms, in particular *Schizosaccharomyces*, *Tetrahymena* and *Plasmodium*, have only few zinc finger genes, so that a global statistical analysis of this protein family cannot provide meaningful results. Some genomes, such as *Naegleria* and *Trichomonas*, are having only limited amount of detected co-occurrences. At present there is no good explanation for the wide-spread avoidance among other domain functions in the many zinc finger genes of *Trypanosoma*.

In contrast to the data set comprising all domain pairs, we observe much less coherence among the domain classes in zinc finger proteins. On the other hand, we observe that the clade-specific patterns become more pronounced in the zinc finger data set. This indicates that the evolutionary trends within this group of proteins is dominated by lineage-specific
8.4 The Consistency of SUPERFAMILY and PFAM Domain Annotation

In general, very little variation in the number of domains per protein was observed. A significant increase is found in human and fruit-fly only. It is unclear, however, whether this a true effect or an artifact arising from a bias in the PFAM database. Iyer et al. (2008) describe a difference in the complexity of chromatin proteins between Diplomonads and Dicristates on the one hand, and between Alveolates and Stramenopiles on the other hand. The data do not show such a systematic difference for proteins containing an rC domain. bP/rC pair seems to have a trend to co-occurrence in RefSeq but more avoidance in the AUGUSTUS data set, in particular in the Pfam data. Overall, the RefSeq patterns are showing weaker signal
than AUGUSTUS. These conditions argue that the annotation-based approach really suffers from problems.

The improved coverage and accuracy of the gene prediction procedure has a major impact on the observed domain co-occurrences. In an earlier study using the non-trainable GENSCAN gene predictor, similarly wide-spread functional avoidance was observed only for the large genomes of multicellular organisms (Parikesit et al., 2011c). At least a moderate positive correlation was found for most other genomes. In the light of the present data, i.e., a much larger set of annotated domains as well as a substantially improved set of underlying gene predictions, these co-occurrences are largely identified as artifacts. The rC* dataset was dropped from the tested functional classes, because the current version of the SUPERFAMILY to PFAM mapping still could not reproduce enough data set for the PFAM annotation. Moreover, it is very clear, that in the big picture, SUPERFAMILY and PFAM domain annotations show a consistent result, and have a similar domain pattern repertoire. It should be noted, however, that PFAM has a more complete coverage, due to its massive sequence alignment collection.

Differences between the RefSeq-based and the AUGUSTUS-based co-occurrence data are observed in particular for some of the combinations involving chromatin regulating (rC) domains, in particular with binding of nucleic acids (bN), binding of proteins (bP), and regulation of binding (rB) domains. There does not seem to be a simple reason for these differences. One possible source of the discrepancy are multipe isoforms, more precisely differences in the average number of overlapping isoforms between the difference groups of domain combinations. An unusually high number of RefSeq isoforms for a particular domain combination might be detected as co-occurrence in the transcript-based data. In contrast, the AUGUSTUS data use overlap-free gene predictions. A closer inspection of the overlapping transcripts shows that the there is some variation depending on species and domain combination. These variations, however, do not provide a straight forward explanation of the differences. It appears that the observed differences must be attributed to ascertainment biases in the RefSeq annotation. The histograms on the average gene overlaps in RefSeq annotation are shown in Figures D.
CHAPTER 9

Conclusion and Outlook

Protein domains become the natural level of description of protein evolution in particular when very large evolutionary time-scales are of interest. The distribution of protein domains is an informative fingerprint of metabolic and regulatory capabilities of an organism. However, to determine the domain repertoire, a solid knowledge on the organism gene structure must be elucidated. Although a plethora of domain annotation data are available in publicly accessible databases for most of the published genomes, quantitative comparisons remain difficult due to dramatic differences in domain annotation methodology and data coverage.

Broad cross-species comparisons are dependent upon unbiased estimates of the number and genomic distribution of protein domains. Thus ascertainment biases that can arise from large differences in the coverage of gene annotation and from the use of very specific domain models need to be avoided or at least reduced as much as possible. Here we have investigated, therefore, to what extent it is feasible to compare patterns of functional protein classes across all major groups of Eukarya based on automatic de novo gene annotation and pooling of domain-models into larger functional classes.

In both respects, substantial improvements should be feasible for future, more comprehensive studies: in particular, it appears promising to combine transcript-based gene annotation with trained, instead of general-purpose, ab initio gene prediction. It is expected that such an extension will increase the accuracy of domain estimates in particular in genomes with unusual structure such as ciliates or kinetoplastids. The incomplete and potentially biased set of gene models available already in early stages of genome annotation projects can be expected to provide sufficient training data for our purposes. A reduction of the phylogenetic bias of domain models, on the other hand, will also require the development of a
solid theoretical framework to inter- and extrapolate protein domain models well beyond the phylogenetic range in which the domain was annotated and hence was available for constructing the HMM.

This dissertation is trying to determine the tendency of domain avoidance or co-occurrences in genetic regulation of eukaryotic genomes. The earlier developed pipeline by Prohaska et al. (2010) was for annotating existing gene and domain annotation. The pipeline was referred to as a SUPER script. In this step, the domain occurrence of transcription factors and chromatin regulation could be elucidated. However, after deeper investigation upon its phylogenic properties, it was found that the domain occurrence of the transcription factors indeed did not agree with their occurrence. The notable example is the domain distribution of human and dolphin, which are very diverse. To this end, it could be concluded that the existing gene and domain annotation is not feasible for inter-clade comparisons (Parikesit et al., 2010a).

Moreover, in the (Parikesit et al., 2010b), the more sophisticated Arli’s Domain Distribution (ADD) pipeline was developed. The main difference between ADD pipeline and SUPER script is that ADD is using de novo gene prediction and Hidden Markov domain models, while SUPER script is using the existing gene and domain annotation. The existing genome annotation was downloaded, and with it, the SUPERFAMILY domain set. The GENSCAN program to predict genes was employed, and the program HMMER was utilized for annotating the domain occurrence and co-occurrences of the domain pair set towards the predicted genes of the representative clades. We were able to elucidate the avoidance and co-occurrence tendency of transcription factor. However, some problems still persist; for instance in the protists, the domain co-occurrences are totally insignificant (Parikesit et al., 2011a). This trend is observed in almost every single domain pair, e.g. in Krab-Znf pair where only human shows a significant co-occurrence. To this end, using just a domain pair data set, the sufficient data visualization for the whole tested clades are deemed to be insignificant.

The effort to solve the problem above was done by using GO functional annotation for annotating domain distribution. By this end, it is clear that a more significant data resolution could be obtained (Parikesit et al., 2011c). In this respect, the ADD pipeline could map almost every single species. However, some species are still having a insignificant or under-represented annotation. This could be demonstrated for Plasmodium falciparum, which hardly has any significant domain co-occurrences in the matrix, and similar cases also are present in Tetrahymena, which is under-represented. The domain co-occurrence pattern is not conclusive due to the difficulties of GENSCAN for annotating genomes with weird gene architectures, such as Trichomonas (single-exons). To this end, GENSCAN annotation pipeline was confirmed as insufficient to provide statistically meaningful domain annotation.

In the final step of the ADD pipeline development, a more sophisticated gene prediction method was used. The AUGUSTUS package was chosen, because it could be trained for the
data of the new genome project, and its popularity within the gene prediction packages is rising (Parikesit et al., 2011c). As expected, the specialized training set for different species could give a more significant result in the data computation. The whole clades have been covered with significant co-occurrences, and the domain pattern could reveal, that the transcription factor and chromatin regulation functional group are correlated. In the multi-cellular organisms with large genomes and large gene families, however, there is a strong signal of avoidance between several functional groups of protein domains, see Figure 8.1. This may be a result of the expansion and diversification of large families of paralogous genes and their use for specific tasks in the regulation of cellular processes. Furthermore, there are substantial differences in the domain co-occurrence patterns of distant lineages, emphasizing the importance of lineage-specific histories and constraints. The available training phase of AUGUSTUS that has been explained in section 5.3 has successfully improved accuracy of the gene prediction. Whole genomes from representative clades have been annotated, and the domain annotation resolution was much better. To conclude, it is not surprising that PFAM gives more domain annotation coverage than SUPERFAMILY due to its massive sequence collection.

In the short term, the ADD pipeline will be utilized to compute larger data sets for obtaining a stronger biological conclusion. Moreover, the ADD pipeline will be improved, to estimate the protein domain content directly from genomic data, thus having the potential to greatly facilitate phylogenomic investigations. Accurate domain inventories can be an important source for generating phylogenetic information (Yang et al., 2005b), in particular for employing ”deep phylogeny” applications. The presence/absence patterns of protein domains were recently used for instance to place the Strepsiptera as a sister group of beetles in insect phylogeny (Niehuis et al., 2012).

Moreover, a novel description on how to envisage the bias of domain annotation is on sight. In the domain repertoire of SUPERFAMILY and PFAM annotations, what is overlapping on the species level can be observed in the domain evolution, e.g. domain gain and loss. Meanwhile, what is non-overlapping on the functional level can be observed by its specific domains.
APPENDIX A

Supplementary Data

The online supplementary data is available in this following link: http://www.bioinf.uni-leipzig.de/~arli/supplement.html
APPENDIX B

Gene Prediction
<table>
<thead>
<tr>
<th>No</th>
<th>Organism</th>
<th>Genome Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Giardia lamblia*</td>
<td><a href="http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AACB02">http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AACB02</a></td>
</tr>
<tr>
<td>8</td>
<td>Thalassiosira pseudonana*</td>
<td><a href="http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AAFD02">http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AAFD02</a></td>
</tr>
</tbody>
</table>
APPENDIX C

Protein Domain Annotation

C.1 Single Domain Occurrences Annotation in the SUPERFAMILY of Transcription Factors
Table C.1: The Single Domain Co-Occurrences Annotation of **GENSCAN -SUPERFAMILY**. The parameters for **HMMER**, **GENSCAN**, and **SUPERFAMILY** were taken from (Parikesit et al., 2010b)

<table>
<thead>
<tr>
<th>Species</th>
<th>Genes</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>SF</td>
<td>Genes</td>
</tr>
<tr>
<td>SF</td>
<td>Genes</td>
<td>SF</td>
</tr>
<tr>
<td>SF</td>
<td>Genes</td>
<td>SF</td>
</tr>
</tbody>
</table>

**Ribon(59066) | Paj(10491) | Helicase(5027) | Dorn(44769))**

<table>
<thead>
<tr>
<th>Species</th>
<th>Genes</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>SF</td>
<td>Genes</td>
</tr>
<tr>
<td>SF</td>
<td>Genes</td>
<td>SF</td>
</tr>
<tr>
<td>SF</td>
<td>Genes</td>
<td>SF</td>
</tr>
</tbody>
</table>

- **Guardia lambi**
- **Trichomonas vaginalis**
- **Trypanosoma brucei**
- **Leishmania major**
- **Naegleria gruberi**
- **Plasmodium falciparum**
- **Tetrahymena**
- **Nakagami pseudoana**
- **Phytophthora ramorum**
- **Clamydomonas**
- **Arabidopsis thaliana**
- **Organ zealota**
- **Daphniscus* sp.**
- **Aspergillus niger**
- **Schizosaccharomyces pombe**
- **Ctenocephalid triton**
- **Drosophila melanogaster**
- **Homo sapiens**
### Single Domain Occurrences Annotation in the SUPERFAMILY

#### Chromo\[54165\]
- Trypanosoma brucei
- Leishmania major
- Naegleria gruberi
- Plasmodium falciparum
- Tetrahymena
- Phytophthora ramorum
- Arabidopsis thaliana
- Oryza sativa
- Dictyostelium
- Aspergillus niger
- Schizosaccharomyces pombe
- Caenoharbditis elegans
- Drosophila melanogaster
- Homo sapiens

### Genes
- Cyclopilin[50892]
- SF
- Dom
- Genes
- Mbtrep[89299]
- SF
- Dom
- Genes
- Leucine[57979]
- SF
- Dom
- Genes
- Fbox[81381]
- SF
- Dom
- Genes
- Znf[57667]
- SF
- Dom
- Genes
- Krab[109640]
- SF
- Dom
- Genes
- Zhou[766\]
- SF
- Dom
- Genes
- Lemine[51079]
- SF
- Dom
- Genes

### Table 1: Single Domain Occurrences Annotation in the SUPERFAMILY of Transcription Factors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila melanogaster</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>126</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>274</td>
<td>256</td>
<td>414</td>
<td>446</td>
<td>421</td>
<td>649</td>
<td>5090</td>
</tr>
<tr>
<td>Caenoharbditis elegans</td>
<td>16</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Aspergillus niger</td>
<td>16</td>
<td>14</td>
<td>17</td>
<td>13</td>
<td>17</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe</td>
<td>48</td>
<td>48</td>
<td>18</td>
<td>16</td>
<td>18</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Dictyostelium</td>
<td>42</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
</tbody>
</table>
## Protein Domain Annotation

<table>
<thead>
<tr>
<th>Species</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>133</td>
<td>86</td>
<td>79</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><em>Tetrahymena</em></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>105</td>
<td>96</td>
<td>125</td>
<td>684</td>
<td>557</td>
<td>448</td>
<td>46</td>
<td>37</td>
<td>49</td>
</tr>
<tr>
<td><em>Phytophthora ramorum</em></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>236</td>
<td>169</td>
<td>180</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>228</td>
<td>186</td>
<td>119</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>12</td>
<td>7</td>
<td>7</td>
<td>391</td>
<td>384</td>
<td>224</td>
<td>29</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td><em>Aspergillus niger</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>15</td>
<td>11</td>
<td>94</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>37</td>
<td>37</td>
<td>450</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td><em>Bromo</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>24</td>
<td>18</td>
<td>192</td>
<td>130</td>
<td>104</td>
<td>12</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td><em>Phytolecanium parasitica</em></td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>34</td>
<td>23</td>
<td>18</td>
<td>334</td>
<td>240</td>
<td>158</td>
<td>16</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td><em>Tetrahymena</em></td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>16</td>
<td>12</td>
<td>12</td>
<td>367</td>
<td>230</td>
<td>142</td>
<td>17</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td><em>Archeoglobus occultus</em></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>31</td>
<td>29</td>
<td>34</td>
<td>286</td>
<td>238</td>
<td>284</td>
<td>12</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>22</td>
<td>22</td>
<td>39</td>
<td>263</td>
<td>233</td>
<td>314</td>
<td>16</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td><em>Dictyostelium</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>76</td>
<td>15</td>
<td>230</td>
<td>162</td>
<td>101</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><em>Aspergillus niger</em></td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>8</td>
<td>6</td>
<td>240</td>
<td>162</td>
<td>84</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>9</td>
<td>9</td>
<td>170</td>
<td>133</td>
<td>121</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>9</td>
<td>25</td>
<td>41</td>
<td>35</td>
<td>221</td>
<td>4</td>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>33</td>
<td>21</td>
<td>36</td>
<td>312</td>
<td>246</td>
<td>261</td>
<td>30</td>
<td>28</td>
<td>55</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>79</td>
<td>57</td>
<td>126</td>
<td>537</td>
<td>421</td>
<td>532</td>
<td>149</td>
<td>114</td>
<td>145</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
<th>Dom</th>
<th>Genes</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Trichomonas vaginalis</em></td>
<td>16</td>
<td>18</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>11</td>
<td>137</td>
<td>525</td>
<td>421</td>
<td>465</td>
</tr>
<tr>
<td><em>Tetrahymena</em></td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>24</td>
<td>23</td>
<td>54</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td><em>Leishmania major</em></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>16</td>
<td>57</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td><em>Naegleria gruberi</em></td>
<td>11</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>17</td>
<td>78</td>
<td>46</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>267</td>
<td>2</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td><em>Thieliaasoma pseudomana</em></td>
<td>54</td>
<td>21</td>
<td>24</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>30</td>
<td>29</td>
<td>69</td>
<td>63</td>
<td>38</td>
<td>27</td>
</tr>
<tr>
<td><em>Phytolecanium parasitica</em></td>
<td>98</td>
<td>39</td>
<td>42</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>53</td>
<td>23</td>
<td>107</td>
<td>88</td>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td><em>Clamydomonas</em></td>
<td>81</td>
<td>64</td>
<td>33</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>38</td>
<td>25</td>
<td>91</td>
<td>61</td>
<td>39</td>
<td>32</td>
</tr>
<tr>
<td><em>Archeoglobus occultus</em></td>
<td>83</td>
<td>49</td>
<td>107</td>
<td>17</td>
<td>17</td>
<td>22</td>
<td>27</td>
<td>26</td>
<td>542</td>
<td>194</td>
<td>160</td>
<td>237</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>88</td>
<td>51</td>
<td>99</td>
<td>22</td>
<td>18</td>
<td>24</td>
<td>30</td>
<td>28</td>
<td>610</td>
<td>232</td>
<td>189</td>
<td>227</td>
</tr>
<tr>
<td><em>Dictyostelium</em></td>
<td>12</td>
<td>8</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>16</td>
<td>134</td>
<td>48</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td><em>Aspergillus niger</em></td>
<td>22</td>
<td>12</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>43</td>
<td>17</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>30</td>
<td>17</td>
<td>19</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>48</td>
<td>19</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>6</td>
<td>3</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>185</td>
<td>4</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>69</td>
<td>40</td>
<td>65</td>
<td>9</td>
<td>8</td>
<td>11</td>
<td>15</td>
<td>19</td>
<td>151</td>
<td>29</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>130</td>
<td>74</td>
<td>237</td>
<td>26</td>
<td>21</td>
<td>77</td>
<td>109</td>
<td>105</td>
<td>613</td>
<td>51</td>
<td>34</td>
<td>66</td>
</tr>
</tbody>
</table>
### C.1 Single Domain Occurrences Annotation in the SUPERFAMILY of Transcription Factors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Giardia lamblia</td>
<td>44</td>
<td>7</td>
<td>14</td>
<td>61</td>
<td>25</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Trichomonas vaginalis</td>
<td>21</td>
<td>16</td>
<td>20</td>
<td>162</td>
<td>70</td>
<td>80</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Trypanosoma brucei</td>
<td>42</td>
<td>38</td>
<td>56</td>
<td>137</td>
<td>108</td>
<td>137</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leishmania major</td>
<td>16</td>
<td>12</td>
<td>18</td>
<td>69</td>
<td>82</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Naegleria gruberi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tetrahymena</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>334</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Podospora australis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phytophthora ramorum</td>
<td>56</td>
<td>29</td>
<td>70</td>
<td>148</td>
<td>191</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chlamydomonas</td>
<td>54</td>
<td>44</td>
<td>43</td>
<td>106</td>
<td>140</td>
<td>27</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>56</td>
<td>29</td>
<td>70</td>
<td>148</td>
<td>191</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>36</td>
<td>29</td>
<td>70</td>
<td>148</td>
<td>191</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dictyostelium</td>
<td>21</td>
<td>17</td>
<td>57</td>
<td>85</td>
<td>92</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>27</td>
<td>22</td>
<td>65</td>
<td>195</td>
<td>350</td>
<td>51</td>
<td>18</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

109
APPENDIX D

Analysis of Functional Annotation
A. *Caenorhabditis elegans* overlapping genes

B. *Drosophila melanogaster* overlapping genes

Figure D.1: The average of *Drosophila*, and *Caenorhabditis*, overlapping RefSeq genes on *PFAM* annotation
Table D.1: Detailed and General function categories Mapping. The abbreviation of ‘m/tr’ means metabolism and transport (Vogel, 2010)

<table>
<thead>
<tr>
<th>General function</th>
<th>Detailed function</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td>Energy</td>
<td>C</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Photosynthesis</td>
<td>CB</td>
</tr>
<tr>
<td>General</td>
<td>Small molecule binding</td>
<td>HA</td>
</tr>
<tr>
<td>General</td>
<td>Ion binding</td>
<td>HB</td>
</tr>
<tr>
<td>General</td>
<td>Lipid/membrane binding</td>
<td>HC</td>
</tr>
<tr>
<td>General</td>
<td>Ligand binding</td>
<td>HE</td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td>R</td>
</tr>
<tr>
<td>General</td>
<td>Protein interaction</td>
<td>RD</td>
</tr>
<tr>
<td>General</td>
<td>Structural protein</td>
<td>ST</td>
</tr>
<tr>
<td>Information</td>
<td>Chromatin structure</td>
<td>B</td>
</tr>
<tr>
<td>Information</td>
<td>Translation</td>
<td>J</td>
</tr>
<tr>
<td>Information</td>
<td>Transcription</td>
<td>K</td>
</tr>
<tr>
<td>Information</td>
<td>DNA replication/repair</td>
<td>L</td>
</tr>
<tr>
<td>Information</td>
<td>RNA processing</td>
<td>LB</td>
</tr>
<tr>
<td>Information</td>
<td>Nuclear structure</td>
<td>Y</td>
</tr>
<tr>
<td>Metabolism</td>
<td>E- transfer</td>
<td>CA</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Amino acids m/tr</td>
<td>E</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Nitrogen m/tr</td>
<td>EA</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Nucleotide m/tr</td>
<td>P</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Carbohydrate m/tr</td>
<td>G</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Polysaccharide m/tr</td>
<td>GA</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Storage</td>
<td>GB</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Coenzyme m/tr</td>
<td>H</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Lipid m/tr</td>
<td>I</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Cell envelope m/tr</td>
<td>M</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Secondary metabolism</td>
<td>Q</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Redox</td>
<td>RA</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Transferases</td>
<td>RB</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Other enzymes</td>
<td>RC</td>
</tr>
<tr>
<td>Other</td>
<td>Unknown function</td>
<td>S</td>
</tr>
<tr>
<td>Other</td>
<td>Viral proteins</td>
<td>SA</td>
</tr>
<tr>
<td>Extra-cellular processes</td>
<td>Cell adhesion</td>
<td>MA</td>
</tr>
<tr>
<td>Extra-cellular processes</td>
<td>Immune response</td>
<td>RE</td>
</tr>
<tr>
<td>Extra-cellular processes</td>
<td>Blood clotting</td>
<td>RG</td>
</tr>
<tr>
<td>Extra-cellular processes</td>
<td>Toxins/defense</td>
<td>SB</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Cell cycle Apoptosis</td>
<td>D</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Phospholipid m/tr</td>
<td>IA</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Cell motility</td>
<td>N</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Trafficking/secretion</td>
<td>NA</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Protein modification</td>
<td>O</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Proteases</td>
<td>OA</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Ion m/tr</td>
<td>P</td>
</tr>
<tr>
<td>Intra-cellular processes</td>
<td>Transport</td>
<td>RF</td>
</tr>
<tr>
<td>Regulation</td>
<td>RNA binding m/tr</td>
<td>A</td>
</tr>
<tr>
<td>Regulation</td>
<td>DNA-binding</td>
<td>LA</td>
</tr>
<tr>
<td>Regulation</td>
<td>Kinases/phosphatases</td>
<td>OB</td>
</tr>
<tr>
<td>Regulation</td>
<td>Signal transduction</td>
<td>T</td>
</tr>
<tr>
<td>Regulation</td>
<td>Other regulatory function</td>
<td>TA</td>
</tr>
<tr>
<td>Regulation</td>
<td>Receptor activity</td>
<td>HD</td>
</tr>
<tr>
<td>N/A</td>
<td>not annotated</td>
<td>NONA</td>
</tr>
</tbody>
</table>
C. *Homo sapiens* overlapping genes

Figure D.2: The average of *Homo sapiens* overlapping RefSeq genes on *PFAM* annotation
List of Abbreviations

A ......................... Arachea
ADD ....................... Arli’s Domain Distribution
B ....................... Bacteria
bN ....................... Binding of Nucleic Acids
bP ....................... Binding of Proteins
ChAp .................. Chromosomal Architectural Protein
CRM ..................... cis-regulatory modules
Dac ..................... De-acetylation
Dme ..................... De-methylation
DNA .................... Deoxy Ribonucleic Acids
Dph ..................... e-phosphorylation
Dub ..................... De-ubiquitination
E[AO] ................... Chromalveolata
E[BK] ................ Basal eukaryots and kinetoplastids
EC ..................... Enzyme Comission
EF ..................... Fungi
EM ..................... Metazoa
EST .................... Expressed Sequence Tag
EV ..................... Viridiplantae
HMM .................... Hidden Markov Model
KRAB ................... Krüppel associated box
Mac .................... acetylation modifier
miRNA ................ Micro Ribonucleic acid
Mme .................... Methylation modifier
Mph .................... Phosphorylation modifier
mRNA ................. messenger Ribo-Nucleic Acids
mS .................... Metabolism of Saccharides
Mub .................... ubiquitination modifier
NRDB ................... Non Redundant Database
ORF .................... Open Reading Frame
R– ...................... Reader of an unmodified side chain
Rac ...................... acetylation reader
rB ....................... Regulation of Binding
rC ....................... Regulation of Chromatin
rC* ...................... Our set of Chromatin Regulation
rE ....................... Regulators of Enzymatic Activity
Rme ..................... Methylation reader
Rph ..................... Phosphorylation reader
Rub ..................... ubiquitination reader
snRNP ................... small nuclear ribonucleoprotein particles
TFBS ................... transcription-factor binding sites
ZNF ..................... Zinc Finger
List of Figures

1.1 Alternate splicing .................................................. 5
1.2 Correlation of the number of protein domains ..................... 5
1.3 Workflow for the estimation of domain abundance data .......... 6

2.1 Cells, Chromosome, Histones, and DNA .......................... 8
2.2 Traditional and Modern view of the Genetic Information Flow ..... 9
2.3 Translation initiation ................................................. 10
2.4 Promoter in eukaryotes ............................................. 11
2.5 The Scheme of RNA Splicing .......................................... 12
2.6 The Scheme of Splice Sites .......................................... 13
2.7 The Sequence Logo of the Vertebrate Splice sites ................. 14
2.8 The nuclear and mitochondrial genetic codes are similar but not identical .... 15
2.9 Translation initiation ................................................. 16
2.10 Polycistronic transcription and processing of mRNAs in trypanosomes ..... 17
2.11 Transsplicing ......................................................... 18
2.12 The Protein Structures ............................................. 21
2.13 Transcription Factors and Zinc Fingers domain .................. 24

3.1 Scheme on The Markov chains ...................................... 28
3.2 Scheme on the hidden markov model ............................... 29
3.3 HMG-box and WD40 Domains ....................................... 34
3.4 Scheme of PcrA Helicase Protein and its Domain .................. 34
3.5 Protein Domain as independent functional unit .................... 35
3.6 Scheme of Domain Combination .................................... 36
3.7 Scheme on SUPERFAMILY database annotation .................... 38
3.8 Scheme on PFAM database annotation .......................... 39
3.9 The distribution of domain functions ............................. 42
4.1 Chromatin Regulation Distribution matrix ....................... 46
4.2 Transcription Factor Domain versus Total Protein Plot ............ 46
4.3 Transcription Factor versus Chromatin domain lot .................. 50
4.4 The discrepancy in the existing gene annotation ................... 50
4.5 Correlation of the number of protein domains ..................... 51
4.6 Scheme on Arli’s Domain Distribution (ADD) pipeline .............. 53
5.1 Phylogenetic Distribution of the Species Considered in this Work ... 56
5.2 Comparison of gene predictions for the 18 species ................. 61
6.1 Comparison of gene with at least one PFAM domain for the 18 species ... 66
7.1 Poisson Distribution ............................................. 74
7.2 Human Genome Randomized Annotation .......................... 75
7.3 The Tendency of Domain Co-occurrences and Avoidance ............ 79
8.1 Summary of domain co-occurrences in 18 eukaryote genomes ......... 89
8.2 Summary of co-occurrence data .................................. 90
8.3 Summary of AUGUSTUS co-occurrences patterns of major functional classes of protein domains across the Eukaryotes .................. 92
8.4 Summary of RefSeq co-occurrences patterns of major functional classes of protein domains across the Eukaryotes .................. 93
8.5 Summary of domain co-occurrences of functional classes of protein domains in zinc finger proteins .......................... 94
D.1 Drosophila and Caenorhabditis overlapping genes ................. 112
D.2 Homo sapiens overlapping genes ................................ 114
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The Intron statistics</td>
<td>19</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of gene and domain annotation with Genscan and hmmmer</td>
<td>48</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary statistics of Domain source data</td>
<td>52</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of gene and domain annotation</td>
<td>58</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary of domain annotation</td>
<td>67</td>
</tr>
<tr>
<td>7.1</td>
<td>Domain occurrences and co-occurrences of ZNF and winged-helix domains</td>
<td>80</td>
</tr>
<tr>
<td>8.1</td>
<td>Overlaps between the 7 functional groups defined in the text</td>
<td>87</td>
</tr>
<tr>
<td>8.2</td>
<td>Summary of genes with at least one zinc finger domain with Genscan and Augustus annotation. HMMER 3.0rc1 Finn et al. (2011) was used to map the HMMs to the protein sequences with the cut-off $E \leq 10^{-3}$.</td>
<td>95</td>
</tr>
<tr>
<td>B.1</td>
<td>The web link of the genome repositories that were used</td>
<td>104</td>
</tr>
<tr>
<td>C.1</td>
<td>The Single Domain Co-Occurrences Annotation of Genscan - Superfamily</td>
<td>106</td>
</tr>
<tr>
<td>D.1</td>
<td>Mapping between detailed and more general function categories</td>
<td>113</td>
</tr>
</tbody>
</table>


130


Publications

JOURNALS:


Books:


Conferences:

Regional Seminar of Advances in Tropical Genomics: Conservation and Sustainable Utilization of Tropical Biodiversity (Presenter)
Parikesit AA: Utilization of Modern Protein Domain Annotation Technique
12/2011: Bogor, West Java, Indonesia

Asia Pacific Bioinformatics Network’s 10th Incob-1st ISCB Asia Joint Conference (Presenter)
Parikesit AA: In silico modification of suberoylanilide hydroxamic acid (SAHA) as potential inhibitor for class II histone deacetylase (HDAC)
12/2011: Kuala Lumpur, Malaysia

26th German Conference on Bioinformatics (Poster Presenter)
Parikesit AA: Evolution of domain co-occurrences: some striking results
09/2011: Weihenstephan, Germany
4th Berlin Summer meeting: Computational and Experimental Molecular Biologist meet (Poster Presenter)
Parikesit AA: Quantitative Comparison of Selected Genomic-wide Protein Domain Co-occurrences in 18 Species
06/2011; Berlin, Germany

Leipzig Research Festival (Poster Presenter)
Parikesit AA: Quantitative Measurement of Genome-wide Protein Domain Co-occurrence of Transcription Factors
12/2010; Leipzig, Germany

25th German Conference on Bioinformatics (Presenter)
Parikesit AA: Quantitative Comparison of Genomic-Wide Protein Domain Distributions
09/2010; Braunschweig, Germany

4th ESFF Conference (Poster Presenter)
Parikesit AA: Protein domain co-occurrences reveal functional changes of regulatory mechanisms during evolution
04/2010; Dresden, Germany

BOMPFFÜNEWERER CONSORTIUM SEMINAR:

27th TBI Vienna Winter Seminar (presenter)
Parikesit AA: Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions
02/2012 Bled, Slovenia

9th Leipzig Bioinformatics Autumn seminar (presenter)
Parikesit AA: Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions
10/2011 Décín, Czech Republic

26th TBI Vienna Winter Seminar (presenter)
Parikesit AA: Quantitative Measurement of Genome-wide Protein Domain Co-occurrence of Transcription Factors
02/2011 Bled, Slovenia

8th Leipzig Bioinformatics Autumn seminar (presenter)
Parikesit AA: Domain Prediction in Eukaryote Genome
10/2010 Décín, Czech Republic
25th TBI Vienna Winter Seminar (presenter)
Parikesit AA; Domain Distribution of Transcription Factor
02/2010 Bled, Slovenia

7th Leipzig Bioinformatics Autumn seminar (presenter)
Parikesit AA; Domain Cooccurence Distribution of Genetic Regulation Activation from an Evolutionary Perspective
10/2009 Děčín, Czech Republic

COLLABORATIONS:

Prof. Usman Sumo Friend Tambunan, Department of Chemistry, Faculty of Mathematics and Science
University of Indonesia; Depok, Indonesia
Bibliographische Daten

Evolutionary Analysis of the Protein Domain Distribution in Eukaryotes
(Evolutionäre Analyse der Protein-Domänen-Verteilung in Eukaryotes)
Parikesit, Arli Aditya
Universität Leipzig, Dissertation, 2012
134 Seiten, 44 Abbildungen, 305 Referenzen
Selbständigkeitserklärung


(Arli Aditya Parikesit)