Department of Physics, Chemistry and Biology

Ph.D. Thesis

Characterization of protein families, sequence patterns, and functional annotations in large data sets

Anders Bresell

Thesis No. 1159

INSTITUTE OF TECHNOLOGY
LINKÖPING UNIVERSITY

IFM Bioinformatics
Department of Physics, Chemistry and Biology
Linköping University
SE-581 83 Linköping, Sweden

Cover art by Anders Bresell 2008, showing images from Paper I–V; Helical wheel of the RBL (background, Paper II), user interface of OAT (top left, Paper III), Relative amino acid residue distribution (top right, Paper IV), MAPEG trimer (lower left, Paper I), *i*-score curves (middle right, Paper V) and RBL structure (bottom right, Paper II).

During the course of the research underlying this thesis,
Anders Bresell was enrolled in Forum Scientium, a multidisciplinary doctoral programme at Linköping University, Sweden.

Bresell, Anders
ISBN: 978-91-85523-01-6
ISSN: 0345-7524
Linköping studies in science and technology. Dissertation, No. 1159

Printed by LIU-TRYCK, Linköping, Sweden, 2008.

Ph.D. Thesis
Thesis No. 1159

Characterization of protein families, sequence patterns, and functional annotations in large data sets

Anders Bresell

Supervisor: **Bengt Persson**
IFM - Bioinformatics
Linköping University
S-581 83 Linköping, Sweden

Opponent: **Inge Jonassen**
Department of Informatics and Computational Biology Unit
Bergen Center for Computational Science
University of Bergen
N-5020 Bergen, Norway

Linköping, 15 February, 2008

Abstract

Bioinformatics involves storing, analyzing and making predictions on massive amounts of protein and nucleotide sequence data. The thesis consists of six papers and is focused on proteins. It describes the utilization of bioinformatics techniques to characterize protein families and to detect patterns in gene expression and in polypeptide occurrences. Two protein families were bioinformatically characterized - the membrane associated proteins in eicosanoid and glutathione metabolism (MAPEG) and the Tripartite motif (TRIM) protein families.

In the study of the MAPEG super-family, application of different bioinformatic methods made it possible to characterize many new members leading to a doubling of the family size. Furthermore, the MAPEG members were subdivided into families. Remarkably, in six families with previously predominantly mammalian members, fish representatives were also now detected, which dated the origin of these families back to the Cambrium "species explosion", thus earlier than previously anticipated. Sequence comparisons made it possible to define diagnostic sequence patterns that can be used in genome annotations. Upon publication of several MAPEG structures, these patterns were confirmed to be part of the active sites.

In the TRIM study, the bioinformatic analyses made it possible to subdivide the proteins into three subtypes and to characterize a large number of members. In addition, the analyses showed crucial structural dependencies between the RING and the B-box domains of the TRIM member Ro52. The linker region between the two domains, denoted RBL, is known to be disease associated. Now, an amphipathic helix was found to be a characteristic feature of the RBL region, which also was used to divide the family into three subtypes.

The ontology annotation treebrowser (OAT) tool was developed to detect functional similarities or common concepts in long lists of proteins or genes, typically generated from proteomics or microarray experiments. OAT was the first annotation browser to include both Gene Ontology (GO) and Medical Subject Headings (MeSH) into the same framework. The complementarity of these two ontologies was demonstrated. OAT was used in the TRIM study to detect differences in functional annotations between the subtypes.

In the oligopeptide study, we investigated pentapeptide patterns that were over- or under-represented in the current de facto standard database of protein knowledge and a set of completed genomes, compared to what could be expected from amino acid compositions. We found three predominant categories of patterns: (i) patterns originating from frequently occurring families, e.g. respiratory chain-associated proteins and translation machinery proteins; (ii) proteins with structurally and/or functionally favored patterns; (iii) multicopy species-specific retrotransposons, only found in the genome set. Such patterns may influence amino acid residue based prediction algorithms. These findings in the oligopeptide study were utilized for development of a new method that detects translated introns in unverified protein predictions, which are available in great numbers due to the many completed and ongoing genome projects.

A new comprehensive database of protein sequences from completed genomes was developed, denoted genomeLKPG. This database was of central importance in the MAPEG, TRIM and oligopeptide studies. The new sequence database has also been proven useful in several other studies.

Acknowledgments

This turned out to be one of the longest acknowledgments I have ever seen, and I started to suspect that other people actually had done the job for me. However, that is not true. There is only one person (besides me) that was of ultimate importance to accomplish this thesis, that is Bengt! The rest of you were not necessary. Nevertheless, I still want to give you my deepest and sincere acknowledgment (some are not named, others are forgotten).

I have had the best time here at IFM and Linköpings University and the biggest acknowledgment is of course given to my supervisor Professor Bengt Persson. You have been the best and most perfect supervisor I could ever dream of. We started from scratch here in Linköping and it has been pure fun to be a part of your group. I wish you the best in your continued work and hope we will be given the opportunity to meet both socially and professionally in the future.

Next to Bengt, my fellow Ph.D. students Jonas and Joel are the most important people during my time as a Ph.D. student. I have enjoyed our scientific discussions every day, ranging from direct implementation details to conceptual philosophies. We have shared experience and listen to each others frustration over imperfect software and databases (usually those that are free and cost us nothing :-)). Roland and Kristofer, you were also important as long as you were here. Especially Roland, who put up with my trivial questions on statistics.

A big thank goes to Janosch for the very inspiring collaboration on Ro52 and the TRIM proteins. It was a true 'twinning' project in the spirit of Forum Scientium, one of the kinds where you really reach further if you work together and exchange experience rather than dividing the work load between each other. I am convinced, you are going to be a very successful researcher, no doubt about it.

A special thank to Bo Servenius at AstraZeneca R&D, Lund, for the work on OAT and introducing me to the corporate world, it was a truly inspiring environment. I want to thank Ralf Morgenstern and his colleagues at Karolinska Institutet for the collaboration on the MAPEG family. I also thank the co-authors of the TRIM manuscript.

Furthermore, I want to acknowledge the PhD Programme in Medical Bioinformatics at Karolinska Institutet for all the interesting and useful courses I had the oppertunity to attend. I also supervised several master students, it was always a joy to see my wildest ideas being tested by others. Thank you Andreas, Jenny, Martin and Patrik.

A big thank you to Jan-Ove, our system administrator. Your technical support has been fantastic and you were perfect for the job. I also want to acknowledge NSC for their support on our computer cluster Dayhoff. Thank you Kerstin and Ingegärd for the administrative support. In general, service establishments in this country are sometimes confused with bureaucracy. You are all truly service minded and that is something I appreciate a lot. Also a small thank you to my assigned mentor Leif Johansson, I am glad we only needed to talk once a year...

One seldom has the opportunity to thank people, and I feel that this is my chance to acknowledge those that have not directly contributed to my thesis but have played inspiring roles:

First of all the people behind the Engineering Biology program, several of them are central figures at the department. I don't know if you ever have been thanked for this, but I want to acknowledge this fantastic Life Science program and all the inspiring teachers that have participated. It was/is a perfect background for a bioinformatician. Another important person is Patrick Lambrix, who introduced me to the field of bioinformatics during my undergraduate studies. Thanks also to Olof Karlberg, who convinced me to follow the path of bioinformatics, I haven't regretted it a second. I also want to acknowledge the senior LiU staff members Timo Koski, Jesper Tegnér and José M. Peña for inspirations as academic researchers and teachers. I also want to thank Arne, my math and physics gymnasium teacher, for inspiring me in the disciplines of natural science.

A big thank you to Stefan Klintström for your work with the graduate school of Forum Scientium. Forum has been one of those components during my time here that not only provided inspiration and lots of joyful courses and study visits, but also has been like a safety net for the Ph.D. students, and you Stefan are its heart and its driving force. I am not sure that people outside Forum understand your importance for Forum. I thank you and all the fellow students in Forum Scientium for all joyful events.

The time here at IFM have not always been about work. Thank you all guys who played football and floorball with me. Jonas and Joel, for all those coffee brakes of shooting eight-ball. Joel and Roland, without all that nice coffee I wouldn't be a particular nice person (or effective researcher either). Lunchklubben, what a geeky way of deciding where to go for lunch. Hobbynight-gang, I miss those weekly events of trying out hobbies that could be combined with drinking 'follubas'. The gamers, it has never been more exciting to watch live-scores on the web, I really though we were going to be rich. The whisky-gang, so many malts, so much money, you have provided me with an expensive hobby. Ooh, how I long for that Port Ellen and that cask bottle of The Ileach.

At last, a special thanks to my favorite mother and my favorite father. The support and comfort you have given me during the childhood and my time at the university, it has been invaluable. Dad, if I ever become a farther to a son, I want to be like you. Thank you Dad, for playing soccer with me all those years and driving me to practice and the weekend games. Mom, the same goes for you, if I ever become a mother I would like to be like you. Thank you for all the LEGO®! I couldn't have had better parents, I love you!

Thank you Lisa, You are the best, not only for being my desperate housewife the last couple of months... I have met them all and you ARE the best, I love you!

Papers

I. **Anders Bresell**, Rolf Weinander, Gerd Lundqvist, Haider Raza, Miyuki Shimoji, Tie-Hua Sun, Lennart Balk, Ronney Wiklund, Jan Eriksson, Christer Jansson, Bengt Persson, Per-Johan Jakobsson and Ralf Morgenstern.
Bioinformatic and enzymatic characterization of the MAPEG superfamily.
FEBS J. 2005 Apr;272(7):1688–1703.
Anders Bresell performed the data collection and analysis of bioinformatics related data. Anders Bresell and Bengt Persson have both contributed to the bioinformatics related sections of the manuscript.

II. Janosch Hennig, **Anders Bresell**, Martina Sandberg, Klaus D. M. Hennig, Marie Wahren-Herlenius, Bengt Persson and Maria Sunnerhagen.
The fellowship of the RING: The RING-B-box linker region (RBL) interacts with the RING in TRIM21/Ro52, contributes to an autoantigenic epitope in Sjögren's syndrome, and is an integral and conserved region in TRIM proteins.
Accepted in J. Mol. Biol., 2 January 2008.
Anders Bresell performed the data collection and analysis of bioinformatics related data. Anders Bresell has written the bioinformatics related sections of the paper. Bengt Persson supervised the bioinformatics analysis.

III. **Anders Bresell**, Bo Servenius and Bengt Persson.
Ontology annotation treebrowser: an interactive tool where the complementarity of medical subject headings and gene ontology improves the interpretation of gene lists.
Appl Bioinformatics. 2006;5(4):225–236.
Anders Bresell performed data collection, computer programming, analysis and prepared the manuscript. Bo Servenius initiated the study and Bengt Persson continued to supervise the study. Both Bo and Bengt contributed with ideas and inputs on the manuscript.

IV. **Anders Bresell** and Bengt Persson.
Characterization of oligopeptide patterns in large protein sets.
BMC Genomics. 2007 Oct 1;8(1):346.
Anders Bresell performed the data collection, computer programming and analysis. Bengt Persson initiated, planned and supervised the study. Both authors contributed with ideas and wrote the manuscript.

V. **Anders Bresell** and Bengt Persson
Using SVM and tripeptide patterns to detect translated introns.
Submitted to BMC Bioinformatics, 7 December 2007.
Anders Bresell has initiated the study, developed the applications, analyzed the data and written the manuscript. Bengt Persson has supervised the study and contributed with ideas and inputs on the manuscript.

VI. **Anders Bresell** and Bengt Persson.
GenomeLKPG: A comprehensive proteome sequence database for taxonomy studies.
To be submitted.
Anders Bresell was responsible for writing source code, compiling data and preparing the manuscript. Bengt Persson has participated in preparation of manuscript.

Contents

Chapter 1

Introduction

In Chapter 1, I will describe fundamental biological concepts (section 1.1) that are needed to understand the methods, results and discussions in the thesis. The aim of the rest of the chapter is to put the field of bioinformatics into a perspective, but already here you can think of bioinformatics as the science of storing, processing, analyzing and making predictions on biological information in general, and molecular biology in particular.

1.1 Fundamental molecular biology concepts

The central dogma of molecular biology [1] describes the flow of biological information from deoxyribonucleic acid (DNA) via ribonucleic acid (RNA) to the synthesized protein (Figure 1.1). It is a very simplified view of the informational processes in molecular biology. Nevertheless, it is an appropriate level for systemizing sequence analysis. DNA can be viewed as the blueprint of an organism; it is a sequence of the nucleotide bases adenine (A), cytosine (C), guanine (G) and thymine (T). Every cell in an organism has the same DNA sequence. The DNA is divided into a set of chromosomes, all stored in the nucleus of the cell. Each chromosome has a set of genes which encodes the proteins. The genes are often divided into coding regions (*exons*) and non-coding regions (*introns*).

To make a protein the cell first copies the gene, by transcribing the gene region into an RNA sequence. Transcribed RNA is processed in order to remove the introns. The mature transcript is denoted mRNA. The protein is synthesized in the ribosome by generating a new type of sequence consisting of 20 different types of amino acid residues. The residues in the protein-coding region of an mRNA are encoded by a sequence of nucleotide triplets, denoted codons. The four letter code of nucleotides can form 64

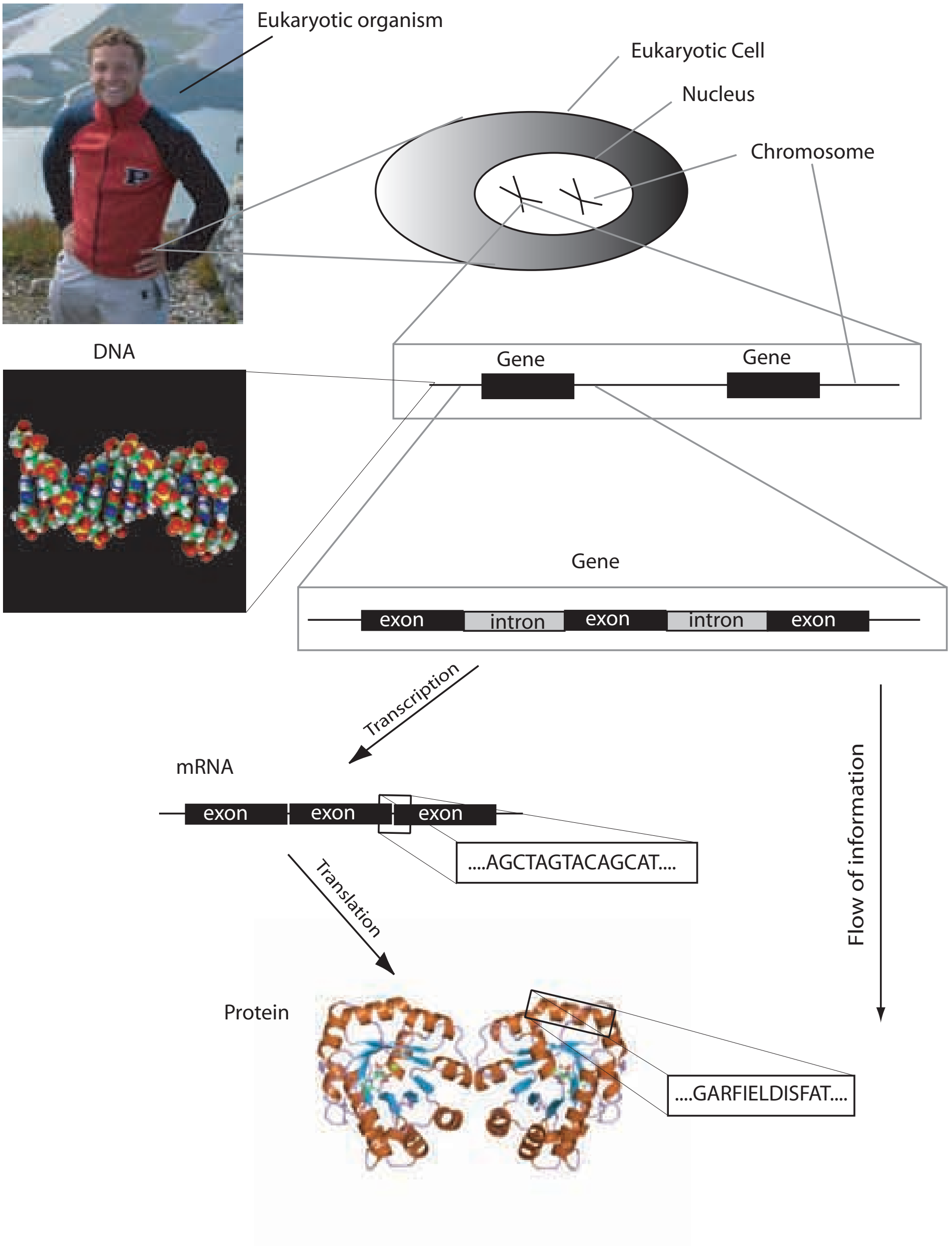


Figure 1.1. The central dogma of molecular biology. Describes the flow of information from DNA (the blue print) via mRNA (copies of instructions) to the protein (the functional entity). The information only flows in one direction.

different three-letter codons. The translation code is degenerated since only 20 codons would be needed to encode all residue types. Consequently each amino acid residue is represented by one or more codons. After its synthesis, the protein sequence is folded into a functional entity, having a specific enzymatic activity or structural property. These descriptions of the fundamental concepts in molecular biology are an oversimplification of the processes going on in the cell; there are many exceptions and additions to what is outlined here. Furthermore, this thesis will mainly focus on higher eukaryotes, and parts of the description given here can not be applied on prokaryotic or archaeal domains of life.

1.2 Bioinformatics in general

Bioinformatics and the sister disciplines of computational biology and systems biology often overlap in their definitions. Some make a distinction between *bioinformatics*, which is technique-driven, and *computational biology*, which is hypothesis-driven as exemplified by the National Institute of Health (NIH) definitions [2]:

> Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
>
> Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Personally, I prefer the definition from the online encyclopedia, Wikipedia [3]

> Bioinformatics and computational biology involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level [4].

Bioinformatics is a multidisciplinary field and one of its challenges is to combine the different disciplines in an effective way. This is a non-trivial task as biology is a rather inexact science in comparison to mathematics and computational sciences. Exceptions frequently occur in biology and this is a complicating fact when we represent biological entities in terms of mathematical or computer models. In order to make the models usable in a computational sense, one often needs to simplify or make assumptions that not always describe the biological concept correctly. However, there is no way that we can handle the huge amounts of biological information without using modern computational techniques.

1.2.1 History of bioinformatics

Bioinformatics tasks have been pursued long before the field got its name. In the early days of sequencing (1950s–70s), protein and nucleotide sequence alignments were built by hand. At that time computers cost a fortune. However, we have seen a major shift in the field of molecular biology the last two decades. The exponential increase of biological data (Figure 1.2) can be ascribed new efficient technologies. Along with this followed an increased demand for efficient processing of the large data amounts. Fortunately, there was also an increased ratio between performance and cost of computer hardware. Hence, the field of bioinformatics took a central role in modern Life Science, which is illustrated with the exponential increase of bioinformatics related research in Figure 1.2.

One important factor that has made bioinformatics popular is the open source mentality. The majority of molecular databases can be used freely (at least for non-commercial use). Furthermore, many bioinformatics applications are also free for academic use. This results in that many researchers around the globe can access the data over the internet and benefit from software and algorithmic developments.

1.3 Bioinformatics of different kinds

The field of bioinformatics can be subdivided into several areas, all with individual scopes and aims. In this section, a few of them will be outlined.

1.3.1 Knowledge management

The explosion of biological data, and sequence data in particular, puts new demands on storing data and making it usable for researchers in all corners of the world. The semi-structured text files (often referred to as flat files) were popular in the pioneering era and combined two tractable features; they were readable by humans and they could easily be analyzed with text parsing scripts. However, when the amount of data and the set of different file formats increases two problems arise; the data fields in a flat file are

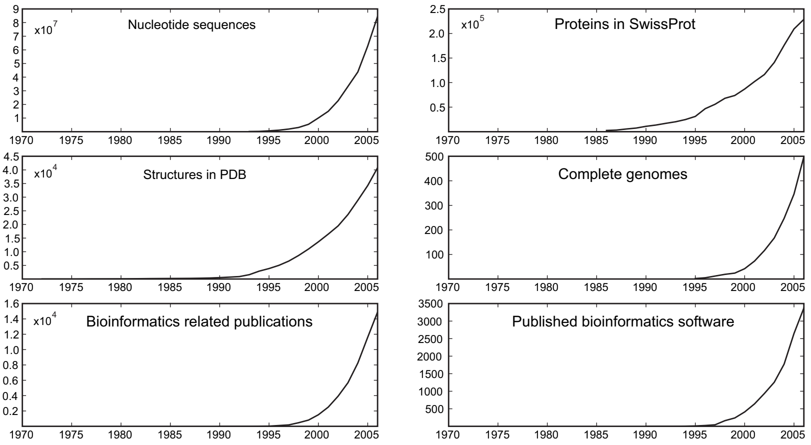


Figure 1.2. The increase of molecular biological data and bioinformatics research. The number of EMBL nucleotide sequences [5] and SwissProt proteins [6, 7] is retrieved via SRS [8–11]. The number of PDB structures [12] is obtained from PDBs own statistics [13]. The number of complete genomes is obtained from GOLD [14, 15]. In conjunction with the exponential increase of data, the bioinformatics related research also grow exponentially. Here illustrated in terms of scientific publications (PubMed [16] search term:bioinformatics) and software (PubMed [16] search term:bioinformatics + MeSH [17] term:software).

not indexed, which results in linearly increasing search time and each data bank needs its own parsing scripts. One solution to this was the Sequence Retrieval System (SRS) [8, 9], which puts an index on top of the flat files. SRS also efficiently integrated the various sources of biological data and presented the user with a common web-based search interface. In recent years a set of 'Nice Views' was introduced, which represents the search hits graphically [10, 11]. In addition, new file formats were supported (e.g. eXtensible Markup Language (XML)). However, the major improvement was the integration of analysis tools, in particular the European Molecular Biology Open Software Suite (EMBOSS) [18, 19], which facilitates basic analysis done by bioinformaticians on a daily basis (see section 1.3.2).

A second effort to handle the flat file data is Entrez at National Center for Biotechnology Information (NCBI) [20]. Entrez is a compilation of data banks but does not have the analysis tools integrated to same extent as SRS. The major benefit of Entrez is the use of precalculated relations between nucleotide sequences, proteins, structures and literature entries. However, the informational content is mostly the same between Entrez and SRS.

Newly developed data resources have taken a more modern approach by building their data structures on relational databases instead of flat files.

Such an approach comes with several advantages; i) it is much easier to keep the data non-redundant, as each type of data is stored in individual tables, ii) an index can be automatically assigned to any field in any table, iii) queries can be formulated to answer any question related to any combination of the data using Structured Query Language (SQL) and iv) backup and incremental updates become easier. This comes of course with a small drawback; the raw data format is not easily read by humans, but this can be addressed by building a user interface on top of the database. Additionally, some expertise in database creation and management is needed.

Ontologies – Towards a common language

To make data interchangeable between research groups and to facilitate computational analysis, it is important to define the various concepts used in the community. However, this is especially problematic in biology, as the field is not very exact and the number of exceptions is much greater compared to many other fields such as computer science, mathematics or physics. A way of coping with this inexactness is to use ontologies or controlled vocabularies. One example of this is Medical Subject Headings (MeSH) [17], which is used for associating descriptive keywords and concepts to entries in Medical Literature Analysis and Retrieval System Online (MEDLINE) [21], a data bank of scientific articles in the domain of biomedicine which is freely accessible through PubMed [16, 20]. The hierarchical structure of MeSH is subject oriented and suited for finding articles of particular interest. However, it is not suited for detailed description of a biological process or protein function. To this end, a more detailed organization and naming convention of enzymes was introduced by Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), denoted Enzyme Nomenclature. This system arranges the enzymes in a four level hierarchical structure based on the chemical reactions that they take part in.

At the dawn of genome projects, massive amounts of new genes and proteins emerged and there was no systematic way of describing the biological processes and functions. There was a long tradition of using non-descriptive names, such as Sonic the hedgehog gene, named after a video game character. Furthermore, parallel discoveries of the same gene in different species lead to a range of synonyms. Michael Ashburner once summarized the problem by stating:

> *Biologists would rather share their toothbrush than share a gene name* [22].

To overcome this he founded the the Gene Ontology Consortium (GOC) aiming at standardizing the language that was used to describe functions of genes and gene products [23]. The Gene Ontology (GO) was initially used by the model organism projects and made it possible to compare protein functions and cellular processes between different species. GO has become a *de facto* standard for annotation vocabularies and has set the gold standard for Open Biological Ontologies (OBO) [24], a collection of ontologies within biomedicine. In addition, GO is implemented in many analysis tools where its impact on whole genome gene expression (e.g. microarrays) has been substantial [25].

1.3.2 Sequence analysis

The central dogma of molecular biology [1] describes the flow of biological information from DNA via mRNA to protein, Figure 1.1. It is a very simplified view of the informational processes in molecular biology (see discussion in section 1.3.5), but is serves as basis for systemizing sequence analysis. The three components are described in more detail in section 1.1. Anyhow, sequence analysis can be done on all three levels, all having their pros and cons. The reason why this thesis is focused on the protein level is that the proteins are the major group of cellular molecules performing most of the functions of the cell. The DNA and mRNA molecules can be seen as information carriers that do not possess any enzymatic activity on their own. If we on the other hand would like to understand how, when and why the proteins are expressed and how they are regulated, we need to also analyze the nucleotide sequences. The information we work on is linear (i.e. sequence of letters) and the principles are the same for the methods we can apply, regardless of which part of the central dogma we perform the sequence analysis on. It is only the types of questions we want to answer and the interpretation of the results that will be different.

One principle idea of sequence analysis is based on evolution, where important biomolecules are conserved among organisms. If the protein function is conserved then the protein sequence must be conserved, hence lessons learnt from one protein can be inferred to another if the sequences in matter are sufficiently similar. The sequence similarity can also be seen on the nucleotide level. However, at the nucleotide level large regions of the sequence are not coding for the protein and hence we would have more noise in the data. For many years, the non-coding nucleotide regions were thought of as ”junk DNA”, but in recent years these regions have been suggested to be important in regulatory processes and have been given much attention [26–28].

1.3.3 Structure analysis

Although much can be learnt from sequence analysis, there are numerous limitations of using the one-dimensional information exclusively. For instance, we can not determine exactly how substrates, inhibitors, ligands, co-factors etc interact with the protein. Neither is it possible to analyze stability, fold or effects of mutations using the sequence only. To perform such analyses, we need the structure (three-dimensional coordinates of all atoms in the protein). Obtaining the structure is a relatively hard task and available techniques such as X-ray diffraction, nuclear magnetic resonance (NMR) or electron microscopy usually require that the protein is stable and can be isolated in sufficient concentration. The number of available structures in comparison to known protein sequences is very low (see Figure 1.2). Aiming at filling the gap of proteins without structures and developing cost-effective methods, various ongoing efforts (denoted structural genomics) try to establish structures in high-throughput manner [29–31]. The success of the structural genomics field promises much for the future and the importance of the structural biology can never be stressed enough. Nevertheless, proteins without known structure will dominate the field of bioinformatics due the cost and time associated with obtaining the structure in comparison to determining the DNA and amino acid residue sequences. Although we will analyze some protein structures later in the thesis, the emphasis will be on sequence analysis.

1.3.4 Text mining

The ever increasing number of publications in biomedicine provides an impressive resource for extracting knowledge. However, the share amount of new scientific literature (see Figure 1.2) makes it impossible to keep abreast of all developments; therefore, automated means to manage the information overload are required [32]. The challenge of analyzing this kind of information is that the natural language used in articles is not appropriate for processing *in silico*. To address this task, the field of text mining has been applied to the discipline of molecular biology. Numerous successful studies have shown that much can be learnt [33] and text mining have recently become popular in the field of systems biology [32].

1.3.5 Systems biology

Systems biology can be described as a philosophy, where the gene-centered analysis has been put aside in favor of a holistic view. The principle idea is that we can no longer focus on only one or few genes or proteins, we need a systematic view in order describe and understand the interplay of genes, proteins, metabolites and states of the cell or tissues. The ultimate goal may be described as determining the dynamic network of biomolecular interactions and the field of systems biology can be summarized by the quote of Uwe Sauer and colleagues:

> The reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge ... the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models [34].

Usually this involves an iterative process by integrating methods from quantitative analysis (levels of mRNA, proteins and metabolites), determining types and quantity of component interactions and combining heterogenous data of various kinds (e.g. experimental data, databases, text mining and computational models). Consequently, the requirements for performing these types of studies in terms of time, costs, instruments, biological samples and expertise are huge. Furthermore, if all the above requirements are fulfilled, these kind of studies usually struggles with the problem that the number of data points is far greater than the number of observations. Hence, it is difficult to establish which components give rise to what effect. Nevertheless, I personally agree that in time this is the path we must follow, and the methods in systems biology developed today will be of major importance in the future. Still, the field of systems biology depends on availability of detailed knowledge on the gene and protein level and we can not give up the gene-centered research for many years. Hence, traditional research is still of major importance and with mutual contributions form the field of genetics, molecular biology, bioinformatics and systems biology, we will obtain a better understanding of the complexity of the living cells.

Chapter 2

Aim

The aim of the thesis was to:

- Characterize the MAPEG superfamily in order to detect new members and determine significant sequence patterns of the subfamilies of MAPEG.
- Characterize the TRIM protein family with focus on identifying structural features important to determine the N-terminal structure of the disease associated protein Ro52.
- Develop a tool to analyze the knowledge associated to a long list of gene and protein identifiers. The tool should be independent of the type of data the user submits in order to be helpful in (but not exclusive to) microarray and proteomics studies.
- Investigate oligopeptide motifs in large protein data sets and to characterize the extremely over- and under-represented patterns in order to understand their effects on protein structure and function.
- Develop methods that use the knowledge gained in the oligopeptide investigation with focus on detecting regions of erroneously translated introns in protein predictions from genome projects.
- Develop a comprehensive sequence database of the proteins from genome projects, which was important in the MAPEG, TRIM and oligopeptide studies.

Chapter 3

Methods

In this chapter, the fundamental principles and methods used in the thesis will be discussed. The methods will be discussed both on a general level and in aspects that are important in presentation of the results in Chapter 4. The data resources used in the thesis are also discussed.

3.1 Pairwise sequence analysis

3.1.1 Aligning two sequences

One of the most fundamental steps in sequence analysis is to compare two sequences. In order to do this we need two things – an algorithm that can identify the similarities (aligning the two sequences) and an evaluation procedure to determine which of all possible alignments that is the optimal one. The evaluation principles are more or less the same for all algorithms and they are based on a scoring system, usually in the form of a substitution matrix. The two most used scoring systems for protein sequences are BLOSUM (BLOcks of amino acid SUbstitution Matrix) [35] and PAM (Point Accepted Mutation) [36]. With these, each position in the alignment is scored dependent upon the type of amino acid exchange; identical (high score), similar (intermediate score) and mismatching (low score). The best alignment between two sequences is the one with the highest total score. However, finding the optimal alignment is a non-trivial task as two sequences of lengths n and m, respectively, have nm possible alignments. The complexity becomes far greater in reality as we also need to handle insertions or deletions (*indels*) in either of the two sequences in order to model the changes that can occur during the molecular evolution of a sequence. Without limiting the number of insertions and deletions, which are represented by gaps in the alignment, the number of combinations of two sequences is infinite. In order to punish the introduction of gaps in alignment algorithms, two additional parameters are used; the gap opening cost (large and negative) and the gap extension cost (small and negative). These parameters make it possible to explore the full search space but at the same time discontinue search paths that accumulate a large negative alignment score.

Global alignment

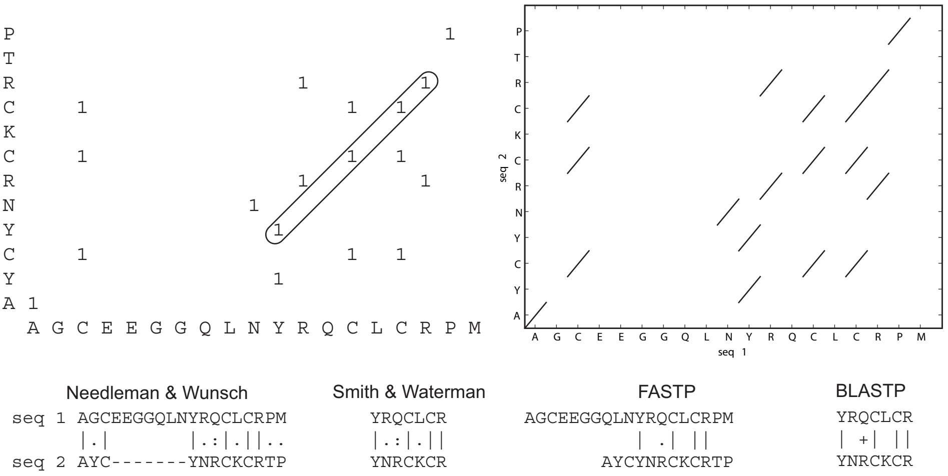
The global alignment approach tries to optimize the alignment of the shorter (length n) of two sequences over the length of the longer one (length m). Global alignment procedures have a high risk of introducing artificial gaps, as they maximize the number of matching positions without taking into account if they are adjacent or not. This method performs well for aligning sequences of similar length, e.g. when aligning sequences of similar fold. In the pioneering work of Needleman and Wunsch [37], a strategy called dynamic programming was introduced, which detects the best *path* through a two-dimensional array representing all similarities between all positions of the two sequences. It has a running time of $O(nm)$ while allowing gaps. The dynamic programming approach set the golden standard for how to solve the alignment problem and most algorithms are a modification of the Needleman and Wunsch algorithm.

Local alignment

Local alignment procedures prioritize to match a local segment and usually do not report unaligned flanking regions. This is useful when analyzing whether two sequences have one or more common domains. The Smith and Waterman algorithm [38] was one of the first local alignment procedures and is still the most accurate [39]. It is an analog of the Needleman and Wunsch algorithm but has a designated gap penalty function, which favors local regions rather than maximizing the overall number of matching positions on the full length. The Smith and Waterman algorithm also has a running time of $O(nm)$.

3.1.2 Finding homologs in a sequence database

In traditional sequence analysis, one usually aims at finding all homologs of a specific sequence. The principle idea of finding homologs is to align the query sequence to every sequence in the database, and those sequences with sufficiently high alignment scores are considered homologous. The



```
   Needleman & Wunsch          Smith & Waterman            FASTP                 BLASTP
seq 1 AGCEEGGQLNYRQCLCRPM          YRQCLCR          AGCEEGGQLNYRQCLCRPM          YRQCLCR
       |.|       |.:|.||..         |.:|.||                |  .|  ||             |  +|  ||
seq 2 AYC-------YNRCKCRTP          YNRCKCR              AYCYNRCKCRTP             YNRCKCR
```

Figure 3.1. A dot plot example. The residues of the two proteins are indicated on each axis. The positions that are identical between the two sequences are indicated with ”1” in the left plot and as short diagonals in the right plot. The result of the different alignment algorithms are given at the bottom. The symbol ”|” indicates identity and the symbols ”.”, ”:” and ”+” indicate similarity.

methods presented so far will scale badly, where m will be the total number of residues in the sequence database. These exhaustive search methods are usually not an attractive approach due the growth of sequence data (Figure 1.2) that currently outperforms the increase of computer performance [40] (cf. Moore’s Law [41]). Instead, the field of bioinformatics has to rely on heuristics (educated guesses). Newer methods use some kind of initial simplified comparison to detect regions or positions in the two sequences that can be used as seeds to find a good alignment. This initial step restricts the search space, but can by no means guarantee that the best solution is found as most alternative solutions never becomes evaluated. Still heuristics has proven very useful in bioinformatics.

The initial comparison can be illustrated in a typical dot plot, shown in Figure 3.1, where the matching regions are illustrated with diagonal fragments. An *indel* can be thought of as an offset between two adjacent diagonal fragments.

FASTA

The first considerable speed improvement in local alignment methods was the FASTP algorithm [42]. This algorithm developed for protein sequences was later generalized to include also nucleotide sequences, denoted FASTA

[43, 44]. In this method only a small fraction of possible *paths* is evaluated. The seed diagonals are chosen by compiling a lookup table of offsets and the number of matching positions. The lookup table can be built by looking at individual positions (ktup=1) or dipeptides (ktup=2), where the latter is considerably faster but at the cost of decreased sensitivity. Each diagonal fragment is scored without gaps according to a scoring scheme and by default only the five best diagonals are selected for further optimization, which bounds the search space drastically. The optimization is a modification of the Needleman and Wunsch algorithm and only takes the surrounding residues of the diagonal fragment into account (16 by default for ktup=2 and 32 for ktup=1) by expanding the seed diagonals and rescoring them according to a selected scoring scheme. The best diagonal in the example of Figure 3.1 is enclosed by a line and has an offset of 7 in sequence 1 and starts in Y and ends in the last R. The optimized FASTP alignment is shown at the bottom of Figure 3.1, which illustrates the region of the best seed diagonal, analogously with the Smith and Waterman alignment, but with the additional expanded flanking residues. The benefit of FASTP is the speed it gains by evaluation only a small set of the diagonals while the decrease in sensitivity is rather small.

BLAST

Basic Local Alignment and Search Tool (BLAST) [40] is a more popular alignment method in comparison to FASTA. It uses a similar approach by selecting only a small fraction of diagonal elements to be evaluated. BLAST uses a look-up table of words (of typically length 3 for proteins) with a score greater than the threshold T, according to a selected scoring scheme. If two hits are on the same diagonal and within a distance A, they are used as seeds (or high-scoring segment pairs (HSPs)) in an extension procedure using a modified Smith and Waterman algorithm. Therefore, the principle difference versus FASTP is that in BLAST the diagonal seeds are made out of similar consecutive residues (words) which are not necessarily identical as in FASTP. The T and A parameters can be set so that very few HSPs need to be extended, and consequently a substantial increase in speed is gained to a relatively small loss in sensitivity. The BLAST algorithm is a bit faster and have about the same sensitivity as FASTP with ktup=2 [39]. However, as the size of sequence database increases exponentially, the speed consideration has been the major factor making the BLAST algorithm the most popular homology search tool.

Scores and E-values

So far we have only discussed different alignment procedures and not said much about the scores. When comparing one sequence aligned to either of two other sequences, the sequence pair with the highest alignment score consists of the two most closely related proteins. However, when is a score sufficiently high to be the result of an evolutionary conserved sequence? In addition, which of two alignments with similar scores but of different lengths contains the most conserved motif? In order to determine whether a score is good or just may be caused by chance, we need some statistical procedure to assess its reliability. The E-value is a statistical measure that estimates the chance of having a certain score S or higher just by chance when taking the size and amino acid residue composition of the query and database into account [45]. An E-value of 0.001 for S is to be interpreted as 1 of 1000 alignments would have the score S or higher just by chance. Consequently, if only 15 hits have a higher score than S, it is very likely that none of them is a result by chance. Most sequence search tools estimate the E-value by normalizing the score S by the scoring scheme, gap penalties and size of query and database; it works quite well as a rule of thumb [39].

3.2 Multiple sequence analysis

A protein family is defined as a group of related proteins, and usually they also have the same or similar function. In this section, we will expand the reasoning of pairwise sequence analysis to multiple sequence analysis. While pairwise analysis can be applied on the problem of finding homologs, it cannot (by itself) be used to draw conclusions for a family of sequences. To this end, we need to take all the members of a family into account in order to characterize sequence signatures caused by similarity in fold, membrane topology, residues participating in active site or ligand binding etc. One objective of protein family characterization on the sequence level is to determine which regions that are important for the protein family (conserved by some property) and which regions that are not. By aligning all member sequences in one big alignment, we can detect regions that are important in all or a majority of the sequences, rather than looking at similarities between only pairs of sequences.

3.2.1 Multiple sequence alignment

The time complexity of finding the optimal multiple sequence alignment (MSA) is $O(m^n)$, where m is the typical length of a sequence in the protein family and n is the number of members to align. This can be done by extending the idea of the two-dimensional dynamic programming procedure into an n-dimensional, collectively referred to as *simultaneous* alignment methods. Following the reasoning of pairwise alignment methods, it comes natural that these exhaustive alignment methods are not at all practical and the need for good heuristics is even more important in MSA algorithms. The alternative to *simultaneous* methods is *progressive* methods, which can collectively by described by the following steps; i) determine a distance matrix by some pairwise similarity measure (e.g. pairwise alignment score), ii) determine a *guidance tree* from the distance matrix and iii) iteratively align one sequence to the others by starting with the two most similar sequences (taken from the *guidance tree*) and ending with the evolutionary most distant sequences. The basic idea of the *progressive* approach is that by starting with the most confident data, we minimize the chance of introducing errors that will propagate through the iterative process.

ClustalW

In 1994 one of the first and most popular MSA methods was published, denoted *ClustalW* [46]. It had its roots in the *progressive* Feng and Doolittle method from 1987 [47]. The two major issues of the *progressive* approach are the local minimum problem and the choice of parameter settings. The focus in *ClustalW* was on the latter where the parameter weights are adjusted based on empirical knowledge and biological reasoning, during the progress of the alignment algorithm. The justification of this procedure is that as long as identities are dominating the alignment most fixed scoring schemes will find a sufficiently accurate solution. However, when only few identities are present the importance of non-identical positions becomes imminent and another scoring scheme would be more appropriate. Furthermore, gaps will not occur randomly and are more likely to occur in regions without regular secondary structure elements (i.e. in loops). The algorithm also gives lower gap opening costs for existing gaps and higher gap opening costs to the surrounding residues in order keep the existing gaps and not introduce new ones. The adjusted weights are derived from the branch length of the *guidance tree*. The guidance tree in *ClustalW* is built by the neighbor-joining method [48], which is good at estimating individual branch lengths and coping with unequal evolutionary rates in different lineages. Taken together, this procedure tends to keep regions of biological importance aligned and introduce gaps only in regions that are less critical for fold or function.

Program	Year	Accuracy		Time (s)
		Simprot	BAliBASE	
ClustalW [46]	1994	0.789	0.702	22
Dialign2.2 [51]	1999	0.755	0.667	53
T-Coffee [52]	2000	0.836	0.735	1274
POA [53]	2002	0.752	0.608	9
Mafft FFT-NS-2 [54]	2002	0.839	0.701	1
Muscle [55]	2004	0.830	0.731	4
Mafft L-NS-i [50]	2005	0.865	0.758	16
ProbCons [56]	2005	0.867	0.762	354
Dialign-T [57]	2005	0.775	0.670	41
Kalign [58]	2005	0.803	0.708	3

Table 3.1. Table shows the evaluation of different MSA methods. The Simprot [59] and BAliBASE [60] are two evaluation sets for assessing the quality of MSA algorithms, where BAliBASE consists of manually currated MSAs of naturally occurring proteins and Simprot consists of fictive sequences based on an evolutionary model. The running times are normalized to Mafft FFT-NS-2, which is the fastest of the methods. Details can be found in [49]. *ClustalW* was for long the golden standard but in recent years many new methods have been published which are both faster and more accurate.

Newer methods

From the mid-nineties until 2002, *ClustalW* [46] was one of very few alternatives on the market. Its popularity set the golden standard but in recent years many new algorithms have become available and most of them are both faster and more accurate [49]. A summary of their performance is shown in Table 3.1, where Mafft L-NS-i [50] is the best method of those with short calculation times. However by the time of the analysis in Paper **I**, for which *ClustalW* was used, these newer software packages were not yet published or proven to be of any significant improvement in comparison to *ClustalW*.

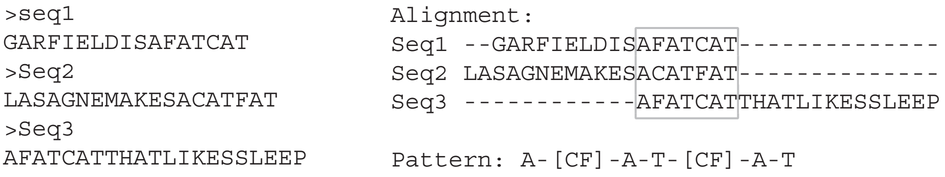
3.2.2 Evolutionary analysis

Analysis from a molecular evolutionary perspective, e.g. the development of a protein family, is complicated because we seldom have historical samples. Instead, we need to rely on reconstructions from the current sequence data. A typical approach to characterize a protein family from an evolutionary perspective is to use an MSA and calculate a distance matrix of it. This distance matrix can be used as basis for constructing an evolutionary tree. Much can be said about evolutionary methods and their interpretation but in this thesis it will only be discussed on a very general level. Two members of a protein family, which once was a single gene, can be the result of two events; either a speciation (the separation into two different species) or a duplication within a species. If a speciation has occurred, the two genes are referred to as *orthologs* and if a duplication has occurred, they are referred to as *paralogs*. The information about orthologs can be used to date a protein family back to the speciation. If members of a protein family have orthologs only in mammals we can say that the protein family is not older than the separation of mammals and birds (about 310 million years ago) [61], which is a quite recent event in an evolutionary perspective. On the other hand, if orthologs are found in both eukaryotes and bacteria (which were separated about a billion years ago) [62] it is very old and is probably a part of a more fundamental biological process.

In Paper **I**, we use the Neighbor-Joining method [48] implemented in *ClustalX* [63]. This method weights the lengths of the branches and it is possible to obtain reliability estimates of the branching order by applying a bootstrap procedure [64]. Bootstrapping is a computationally intensive method, which in phylogenies involves sampling the columns from the MSA with replacement. Hence, each bootstrap sample (tree) will have a different set of sampled sites, of which some sites might be sampled twice or more while others are not included. This procedure thus gives a number of different trees, each based upon a different data set. Branches (or groups) are considered confident if they have the same leaves in most of the bootstrap trees and are usually determined by a threshold of 90% or 95%.

3.2.3 Finding homologs using multiple sequences

The problem of detecting homologs to a protein family using a pairwise alignment technique is that our query sequence might not be the best representative. Probably no single member is good enough to represent all the sequences belonging to the family, and a single search strategy would lead to bias towards finding only hits similar to our query sequence. When detecting members of a family it would be better to weight properties that are in common for all members to be more important, properties that are observed occasionally should be considered less important and properties that never have been observed should be ignored. Such a strategy can be applied if the information in the MSA of the protein family can be implemented in the homology search algorithm.



```
>seq1                       Alignment:
GARFIELDISAFATCAT           Seq1 --GARFIELDISAFATCAT-------------
>Seq2                       Seq2 LASAGNEMAKESACATFAT-------------
LASAGNEMAKESACATFAT         Seq3 -----------AFATCATTHATLIKESSLEEP
>Seq3
AFATCATTHATLIKESSLEEP       Pattern: A-[CF]-A-T-[CF]-A-T
```

Figure 3.2. Prosite patterns. From the MSA of this sequence family it is possible to build a pattern that detects all members. Patterns of known active sites or motifs of ligand binding sites is found in the Prosite database [65]

Patterns and profiles

A simple form of including data from several sequences is to use *patterns* or regular expressions. An example of a family of three sequences is illustrated in Figure 3.2. From the MSA we can build a pattern that finds all three sequences using the signature A-[CF]-A-T-[CF]-A-T, where the brackets represent a set of allowed residues (either C or F). This method is used in the Prosite database [65–67], which also handles arbitrary residues (X), exclude residues ({R} = not R), repeats, repeat intervals and any combination of these. This generalization of the query is more sensitive than using a pairwise alignment search strategy, but it is limited by the fact that if we have not yet seen a certain possible residue at a position we will miss those members also in the feature. Furthermore, it is difficult to determine an appropriate level of detail of the pattern that is both specific and sensitive.

A more sophisticated approach is to weight the preferred residues at each position. From an MSA it is possible to derive a scoring matrix for the protein family by scoring the residues at each position by the number of times they have been observed. A further improvement is to include substitution matrix information and background frequencies, which results in a family specific scoring scheme, denoted *profile*. This approach has proven to be orders of magnitude better than pairwise alignment algorithms [68] and the increased sensitivity makes new unknown family members more likely to match. This kind of sequence models are also found in the Prosite database [65].

Aligning a sequence to a profile is not very different from performing a pairwise alignment. Roughly, the only difference is that instead of a substitution matrix of $20x20$ residues we use a scoring matrix of $Lx20$ where L is the length of the family. Further improvements may be obtained by allowing position-specific gap cost, analogously to *ClustalW*(section 3.2.1).

Profiles are preferably used for modeling domains where the there is a built-in length constraint. However, patterns are sufficient (or better) for capturing the biological signal in a few situations. These signals usually consists of 10–20 conserved residues at different positions in the sequence as for catalytic sites of enzymes, sites for attachment of prosthetic groups, metal ion binding sites, cysteine bridges and regions involved in binding another molecule (ADP/ATP, GDP/GTP, calcium, DNA and protein). Profiles and patterns can be used together, where the profile can be used to detect the domain and the patterns can be used to locate the active site within the domain.

PSI-BLAST

The requirement for performing profile-based search procedures is an MSA compiled of relevant sequences. The retrieval of seed sequences for the MSA is usually based on standard pairwise homology search with a query sequence. This protocol was implemented in an automated fashion in Position-Specific Iterated (PSI)-BLAST [40]. The method performs an ordinary BLAST search as a first step. The resulting hits above a certain cutoff are used in building an MSA and an accompanying profile. The profile (with increased sensitivity) is used in a new search generating a new hit list. Again, the second hit list is used in building a new profile. This iterative process is pursued until the algorithm converges and no new sequences are found. The algorithm can be run in a semi-automatic mode were the user can intervene with the hit list by including members below cutoff and exclude non-members above cutoff in order to increase specificity in the following iteration. Every profile iteration takes just a little more time than an ordinary BLAST search. The total amount of iterations leads to a total running time which is many times more than an ordinary BLAST search and the method is therefore sometimes not applicable to a problem for the same reasons as for the Needleman/Wunsch and Smith/Waterman algorithms (section 3.1.1). However, the gain of sensitivity usually justifies the use of it.

Hidden Markov Model (HMM)

The Prosite profiles described above are intuitive and the reasoning is based on empirical knowledge. The effectiveness relies on expert knowledge about the family and the handcrafted design approach, which can be partly automated [65–67]. An alternative approach is to use HMM methods. They rely on probability theory; hence, rigorous mathematic algorithms can be applied. The technique was initially used in speech recognition but was proven very effective in modeling protein family domains [69, 70]. In this

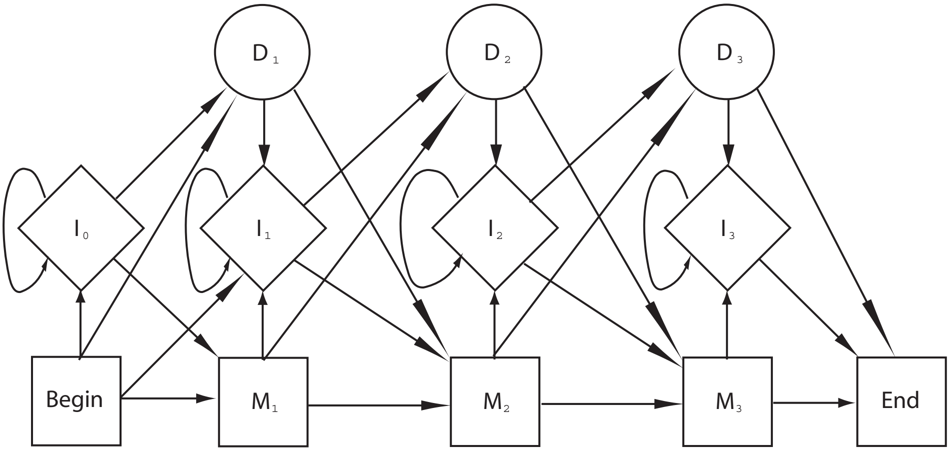


Figure 3.3. HMM topology. Squares indicate match states (modeling consensus positions in the alignment). Diamonds indicate insert states (modeling insertions relative to consensus). Circles indicate delete states (modeling deletions relative to consensus). The arrows indicate state transitions.

thesis we will refer to these profile HMMs as HMMs, although HMM generally includes a much wider scope than presented here.

The Prosite profile model is a matrix representation of residue frequencies and gap costs for a domain. An HMM models the same information by a graphical representation of the states and transitions that fits the sequence to the model, Figure 3.3. The match state is a direct analog of the column (or position) in the MSA and the insert states are used for "extra" residues in sequence to be fitted relative to the consensus sequence. The delete states are used to *stretch* a sequence in order to fit it to the model, i.e. when gaps need to be inserted relative to the consensus sequence. A transition (arrow) represents the move from one state to another in the topology graph. Within each state all possible events are given a probability (e.g. a match state has assigned probabilities for all the 20 amino acid residues). The transitions from one state to another is also given probabilities, typically a higher probability is given to a move from one match state to the next match state than to a move from one match state to an insert or delete state. A transition from a match state to an insert state corresponds to the gap opening cost in a profile (or substitution matrix) and the transition from one insert state to itself corresponds to the gap extension cost. The HMM method can handle any type of scenario of fitting a sequence to the domain model, which is represented by taking different paths through the topology. However, the path with the largest product of probabilities is the optimal one and this can be calculated by dynamic programming.

Software packages are available for the different HMM-related tasks. These tasks can be to build (or train) an HMM using an MSA of a domain as input, align one or several sequences to the domain or to search a sequence database for domain regions matching the HMM. The two most widespread packages are Sequence Alignment and Modeling System (SAM) [71] and HMMER [72], the latter is used in Paper **II**. HMMER is also the package utilized in the Pfam domain database (see section 3.3.2).

When searching a sequence database with Prosite profiles and HMMs, both present the hitlist in terms of scores and E-values. The performance of Prosite profiles and HMMs is dependent on the quality of the input MSA. If the MSA has badly aligned regions, noise or errors will be included in the model and severely affect the method negatively.

3.2.4 Discrimination of sequences

The single and multiple sequence analyses described so far are used to collect similar sequences and to extract their common properties. The aim and methods are slightly altered for protein superfamilies (i.e. proteins with similar fold that not necessarily have the same function) and for multi-domain protein families. Once the members are aggregated, it is possible to gain additional information by detecting discriminating features that distinguish the different subfamlies. This can be done by modeling the sequences as vectors, where each element is a numerical feature derived from its characteristics in the respective position in the MSA. The numerical features can be of various types, e.g. Kyte and Doolittle hydrophobicity values [73] (Paper **I**), secondary structure scores, amphipathicity (Paper **II**) and conservation scores.

Principal Components Analysis (PCA)

Principal components analysis PCA [74] is a method for classifying and discriminating by reducing dimensions and the method is part of a discipline called multivariate data analysis (MVDA). The fundamental idea is to represent data in a lower dimension while keeping the most descriptive information (Figure 3.4). By mathematical terms, it is defined as an orthogonal linear transformation using the variance in each variable to retrieve a new coordinate system. The first principal component is the vector in the multidimensional input space that describes the data best in a least square sense. All principal components will go through the average point.

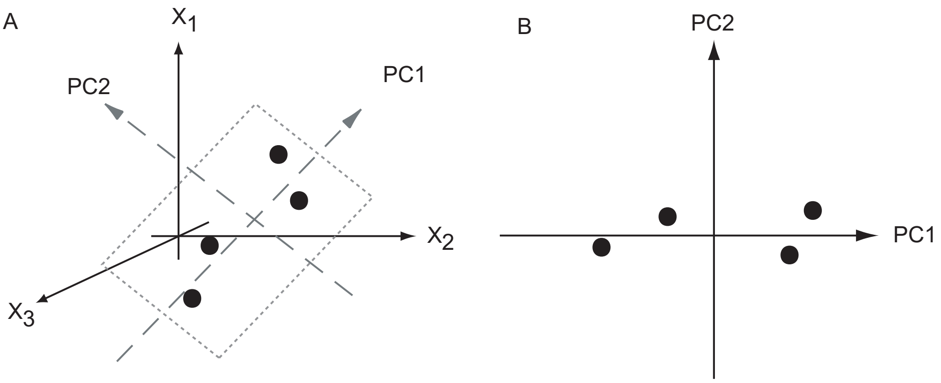


Figure 3.4. Principal Components Analysis (PCA). (A) The three-dimensional coordinate system x_1, x_2, x_3 and the four data points. **(B)** The new coordinate system in two dimensions using the first and second principal components (PC) as axes.

Each observation (data point) is then projected onto the first principal component. The projection distance on the first principal component can be thought of as the residual error if the data is represented in only one dimension. The second principal component is selected to be orthogonal to the first and a two-dimensional plane can be formed with the first and second principal components as axes. The data can be easily viewed and interpreted in this new coordinate system. Clusters of observations in the new coordinate system represent subclasses and the discriminative features can be mapped by the weights of the principal components back to the input space. Any number (less then or equal to the dimensionality of the input space) of principal components can be used and the more principal components we generate the better they will fit the data. However, using more than three principal components makes it difficult to visualize and the risk of over-fitting the data is increased for every additional component, as we may include information of non-representative members. The latter problem can be solved using cross-validation. It is recommended to scale and mean-center the input data in order to obtain good separability and to avoid bias of certain variables.

Novel methods using principal components analysis (PCA) in protein sequence analysis were presented in the mid-nineties [75, 76]. PCA is not of the same central importance in bioinformatics as HMMs and Support Vector Machines (SVMs), which usually are better at capturing biologically important features. However, PCA can be very useful in explorative analyses (unsupervised clustering) when nothing is known about possible

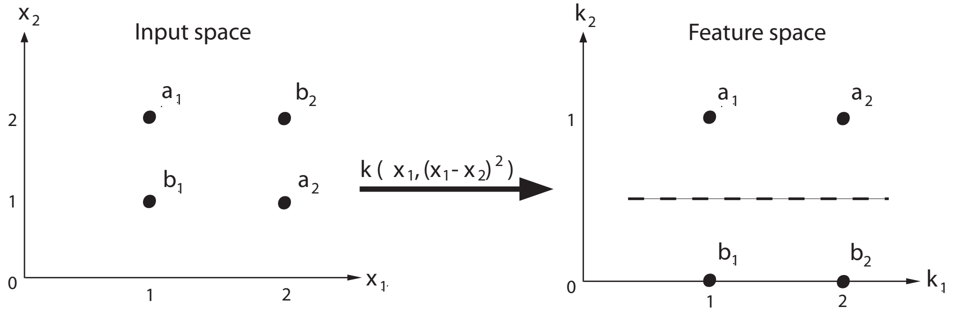


Figure 3.5. Using SVM to separate classes. In input space the two classes (**a** and **b**) is not linearly separable. By mapping the input data via the kernel function $\langle x_1, (x_1 - x_2)^2 \rangle$, the two classes become linearly separable in feature space.

subfamily classification. In paper **II**, we used unaligned sequences represented by vectors of their hydrophobic moment and could by this method detect three distinct subfamilies not yet discovered with MSA-based methods.

Support Vector Machines (SVMs)

SVMs have been used in many areas of bioinformatics including protein function prediction, protease functional site recognition, transcription initiation site prediction and gene expression data classification [77]. The SVM approach is very different from that of PCA. Instead of reducing the dimensions of the input space, one usually uses a nonlinear mapping of the data into a higher dimension called *feature space*. The principle idea is that if it is not possible to do a linear separation of two classes in the input space, this will become possible if we blow up the dimensions using a kernel function (Figure 3.5). Roughly, a kernel function is a similarity measure that by some means includes the dependencies between input variables. In SVM classification we determine the optimal hyperplane in *feature space* that separates two classes, which is obtained by training on a set of examples. By knowing the class belongings in the training set, the SVM can capture the most important features (i.e. *support vectors*) that distinguish one class from the other. Besides the classification problem, SVMs can be used in regression (fitting data to a function) and to estimate a distribution (one-class problem).

The kernels used in mapping input space to feature space can be of various kinds (e.g. dot product, polynomial and radial base function) [78]. In two cases we will not benefit from moving into higher dimensions; i) when number of data points $\ll$ number of features and ii) when both number of data points and number of features are large [79]. The latter case becomes true in Paper **V**. In these two cases, we will do equally well or better with a linear kernel, i.e. it is sufficient to calculate the optimum weights of the elements of the input vector. The significantly shorter calculation times obtained with a linear kernel in comparison to nonlinear alternatives will therefore favor the simpler form.

Generalizability can be described as the ability of a model to correctly classify data that has not been used during training. If we learn many details in the training data (e.g. when almost all training examples become support vectors), we might be able to classify all training examples correctly. Nevertheless, if the data we actually want to classify only have a few common features of all the features in the input space, they will be incorrectly classified because the method emphasizes individual details rather than common properties. Therefore, we have a tradeoff between the number of support vectors (ability to classify difficult data points correctly) and generalizability (ability to classify data points correctly, which we have not yet seen).

The SVM methodology is based upon the principles of learning theory, which includes mathematical theorems on how to calculate the best separating plane as fast as possible and to obtain generalizability (avoid over-fitting) [78]. Two widely used software packages are SVMlight [80,81] and LIBSVM [82,83]. A linear version of LIBSVM is used in Paper **V**.

3.3 Databases and data resources

3.3.1 Sequence databases

The primary source of bioinformatic data is the protein and nucleotide sequences. In this thesis, all analyses are performed with the protein sequence in focus and in this section the core set of resources are described.

UniprotKB

UniProt Knowledgebase (UniProtKB) [6,7] is the universal resource of all available proteins and it is divided into two parts: SwissProt and Translated EMBL (TrEMBL). SwissProt is administrered by European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB) and is often referred to as the current *de facto* standard of protein knowledge. The SwissProt section of UniProtKB consists of only well-documented and manually curated protein sequences. Entries in SwissProt are frequently updated with respect to new findings. Hence, SwissProt sequences are often used as seeds when creating a predictive model.

TrEMBL holds the protein sequences of all protein coding sequences (CDSs) in the European Molecular Biology Laboratory (EMBL) nucleotide sequence database. Entries in EMBL can be posted by any researcher and the entries are in general posted with only limited additional information besides the sequence, and seldom or never are the entries updated. Hence, TrEMBL is SwissProt's opposite in being a non-curated repository of all available translations of nucleotide sequences.

RefSeq

NCBI's reference sequence database (RefSeq) is a resource of complete and incomplete genomic, RNA and protein sequences [84]. Its sequence content is recruited from GenBank [85] but is made non-redundant. The RefSeq staff classifies each entry according to its quality and performs manual curation. Furthermore, they perform a range of algorithmic checks on the sequences, which results are used for prioritizing sequences that need manual investigation. The RefSeq database includes all types of organism and is a good reference set for finding reliable information and can be thought of as the SwissProt counterpart for messengerRNA.

Ensembl

Ensembl is a database of eukaryotic genome-centered information [86, 87]. It started with the completion of the human sequencing project [88, 89] but is currently (October 2007) including 35 species [90]. The complete genomic sequence of an organism only consists of raw nucleotide sequences of the chromosomes, and it is made useful first after that the individual regions are annotated (assigning descriptive features to the nucleotide coordinates). Manual annotation of these regions takes enormous amounts of time, but much can be learnt from predictions of automated annotations. The Ensembl database is built by an automatic annotation pipeline [91, 92] that predicts various types of feature in the sequence. Its core focus is on genes and their transcripts. All known genes, mRNAs, ESTs and proteins are mapped down to the genomic level, which is a more reliable approach than making *ab initio* predictions by analyzing signals in the original sequence [93, 94]. A deeper discussion of the topic is found in Paper **V**. Ensembl provides a browser interface [95], which enables users to zoom in on regions of interest and view their transcriptional landscape. In addition, the database is available in either flat file format or relational MySQL database format [96] and is accessible via Perl application programming interface (API) and through the BioMart interface [97, 98], which handles batch queries in a range of various input and output formats.

3.3.2 Protein family databases

Pfam

The Pfam protein family database [99, 100] contains more than 9000 families (release 22, July 2007). The protein families are modeled according to their domains and are represented by profile-HMMs (section 3.2.3). All data are available either through the web interface, in flat file format or as a MySQL relational database [96]. Each family is created from a manually curated MSA, denoted *seed alignment*. An HMM is built on the seed alignment and the model is iteratively calibrated to determine a score value that mediates detection of the member sequences but does not result in an overlap with other families. The same setup is then used for scanning UniProtKB, which results in the *full alignment*. Each family has a documentation page summarizing the links to other family databases, external resources and relevant citations. The HMMs can be downloaded and used locally with the HMMER package [72], as in Paper **II**.

The HMM approach of Pfam is one of the most effective ways of modeling protein families but suffers from one disadvantage; the one-to-one relationship between fold, domain and function is not appropriate for all families. In order to adress this, the concept of *clans* was introduced, which groups related families in a hierarchical manner [101].

Prosite

The Prosite database [65–67] consists of profiles and patterns, both described in detail in section 3.2.3. The Prosite entries are well documented and provide a good starting point when searching information about a family.

Interpro

Various additional databases provides resources on protein families, domains and functional sites (e.g. PRINTS [102], ProDom [103], SMART [104], TIGRFAMs [105], PIRSF [106], SUPERFAMILY [107], Gene3D [108] and PANTHER [109]). These databases (including Pfam and Prosite) are all hierarchically integrated into InterPro [110], which provides a nice overview of existing types of family models. InterPro is used to annotate

UniProtKB on a regular basis via the InterProScan software [111,112]. The current release of InterPro contains more than 13 000 entries and covers over 78% of UniProtKB proteins [110]. InterProScan can be used locally (e.g. for genome annotation) or via a web interface. Furthermore, InterPro links to additional external databases, including relevant literature references, and the database is of major importance in bioinformatics.

3.3.3 Identifiers and annotation data banks

The variety of databases and levels of sequence data (organism, genomic, genes, gene names, transcripts, proteins etc) results in a myriad of identifiers. Some of the problems associated with this multiplicity are already discussed in section 1.3.1. In order to combine various types of data banks and to transfer the results from one tool to another (as in Paper **III**), we need meta data resources that can be used in mapping identifiers and to pre-compute annotations.

Every database needs internal identifiers in order to control introduction, updates and deletion of entries. These can also be used for cross-references between databases. Consistency is needed in order to accurately combine or map identifiers between different resources. One such meta resource is the International Protein Index (IPI) [113], which combines UniProtKB [6, 7], Ensembl [86] and RefSeq [84] entries. Entries in these databases that represent the same gene or protein are assigned a common IPI identifier.

Merging identifiers is not only crucial for transferring data between analysis tools; it is also of importance in annotation analysis. An annotation is a mapping between descriptive information and a gene or protein identifier. Hence, if the identifier and annotation sets are not compatible, the descriptive information cannot be retrieved.

The Gene Ontology Annotation (GOA) project [114, 115] annotates UniProtKB entries [6, 7] and includes both manual investigations and automatically derived annotations. High quality manual annotations require much time of expert personnel, while automatic approaches tend to be inexact, but can be used in high-throughput procedures. Remarkably, GOA has made an impressive effort in building a framework for retrieving high-quality automatically derived annotations using a range of reliable predictive and inferring procedures. As of July 2007, GOA released over 20 million annotations, to more than 3 million proteins. These can easily be incorporated in various databases and analysis tools, largely due to efforts like IPI [113].

3.4 Statistics and evaluation

3.4.1 The multiple comparison problem

Typical predictions in the field of bioinformatics are made on large data sets. Classical statistics is intended for small samples and the classical p-value is not always appropriate without certain precautions or adjustments. In statistical hypothesis testing, we form a null hypothesis and an alternative hypothesis. A test is called significant if we can reject the null hypothesis in favor of the alternative hypothesis. A typical criterion for a test being significant is that we can reject the null hypothesis with a calculated risk, i.e. a p-value of 5% implies that we have a 5% error of falsely rejecting the null hypothesis. This type of error is called a false positive or Type I error, denoted by α. The p-value describes the False Positive Rate (FPR) and when performing a test n=10 000 times we would expect to have 500 false positives (α x n). In bioinformatics, the features we want to predict usually occur very seldom (e.g. mutations, clustering of genes on a chromosome, participation in a few of many thousands cellular processes etc). In such cases, if only 50 of the 10 000 features are truly positive; we will make 10 false positive predictions for every positive feature that we predicted correctly. That is, if we want to detect 10 correct positive features we will have a total error of 91% among those called significant (100/(10+100)). This phenomenon is denoted the *multiple comparison problem.*

There are ways of dealing with this. One such method is the Bonferroni correction, which can be approximated by dividing the Type I error with the number of tests performed (α/n). However, this correction is much too conservative and often results in no significant features at all [116], i.e. we have neither false nor true positives for the adjusted p-value of $5 \cdot 10^{-6}$. Another method is to adjust the test in order to control the False Discovery Rate (FDR). While FPR determines the rate of making false positive in each test (i.e. the error accumulates for every performed test), the FDR is the rate for which we include false positives among those we call significant. If we have a method that allows an FDR of 5% and we want to detect 10 of the positive features, then the expected total error among those we call significant will be 5%. In Paper **III**, we use an FDR-based method called q-value [117] in order to address the multiple comparison problem when testing if a gene list is significantly enriched of annotations in a branch of GO or MeSH, where many branches are tested.

		Truly	
		positive	negative
Predicted	positive	TP	FP
	negative	FN	TN

Table 3.2. Table shows the relationship between truly positive and negative items and the prediction outcome.

3.4.2 Evaluation and assessments procedures

There are several ways to assess the performance of a prediction algorithm. In this section, we will describe a few of them that can be applied to binary classification. We define the items we want to detect as *truly positives* and the remainder are denoted *truly negatives.* We can then form the following four parameters for the outcome of the prediction algorithm on the two-class data set:

- TP - true positives, the number of correctly predicted truly positive items.
- TN - true negatives, the number of correctly predicted truly negative items.
- FP - false positives, the number of truly negative items that are predicted to be positive.
- FN - false negatives, the number of truly positive items that are predicted to be negative.

The relationships between these parameters are illustrated in Table 3.2. A good algorithm detects many positive features (it has high sensitivity or *recall*) and generates few false positives (it has high specificity or *precision*). Using the parameters above, we can form the following performance measures for recall and precision.

$$\begin{aligned} Recall &= \frac{TP}{truly\ positive} = \frac{TP}{TP+FN} \\ Precision &= \frac{TP}{predicted\ positive} = \frac{TP}{TP+FP} \end{aligned}$$

And from recall and precision we can form a joint performance measure:

$$F\text{-}score = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Generally, both recall and precision cannot be optimized at the same time. In the extreme, if the complete data set D is predicted positive we will have a maximum recall of one but simultaneously many FP and consequently a very low precision. A method that allows no FP and consequently has the maximum precision of one, will result in low recall and the method will only detect the most common positive features.

The FPR and FDR, described in section 3.4.1, can also be expressed in the terms of TP, FP, TN and FN:

$$\begin{aligned} FPR &= \frac{FP}{truly\ negative} = \frac{FP}{FP+TN} \\ FDR &= \frac{FP}{predicted\ positive} = \frac{FP}{TP+FP} \end{aligned}$$

In Paper **III**, we wanted to control the number of errors made on those predicted to be positive (controlling FDR), while in Paper **V** we expected to have many true negatives and hence used FPR to control the number of errors. However, excluding any of the TP, FP, TN and FN parameters may lead to an unbalanced estimation of the performance [118]. Two typical balanced assessment methods are accuracy and the Matthews' correlation coefficient (MCC) [119] defined by:

$$\begin{aligned} Accuracy &= \frac{correct\ pred.}{full\ data\ set} = \frac{TP+TN}{TP+FP+TN+FN} \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{aligned}$$

A predictive algorithm is often tuned in either direction to mirror the associated *cost* and *risk* of making a prediction error. This can be exemplified by an algorithm that predicts nuclear power plant failures, where we would allow many false alarms (FP) in favor of actually detecting the true power plant failures (TP). A counter example would be an algorithm that detects tax dodgers. Each suspected tax dodger needs to be checked in a costly manual investigation. If too many FP are allowed, the unpaid tax and fines of the true tax dodgers (TP) will not cover for the expenses of the investigations. In this latter case, a priority on precision would be preferred.

A frequently used assessment method is to illustrate the receiver operating characteristics (ROC) [120], which shows the *recall* as a function of FPR (Figure 3.6). Perfect performance would be if we have a recall of one and a FPR of zero, corresponding to the upper left corner. If we have no discriminative performance at all (e.g. random guessing), it would give a point on the diagonal from the lower left to the upper right, denoted *line of*

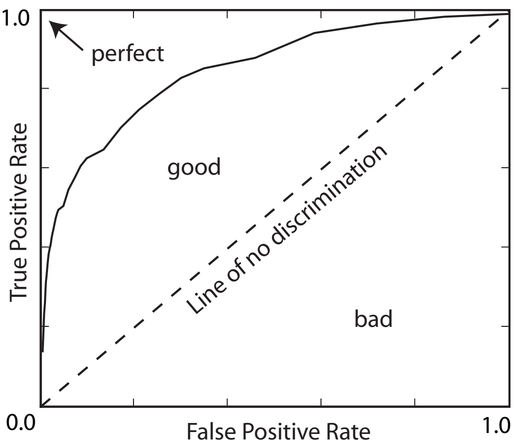


Figure 3.6. The receiver operating characteristics (ROC) curve. The plot shows recall (TPR) as function of type I error (FPR). A perfect predictor detects all true positives and no false positives, indicated in the upper right. In practice, these two categories can not be optimized simultaneously.

no discrimination. A point above the diagonal corresponds to a good classification result. The *recall* and FPR metrics are based on ratios and are therefore especially useful when the distribution of positives and negatives is skewed [121]. Due to the insensitivity of the distribution, it is possible to choose cutoffs or parameters on one data set and still expect similar performance on another data set with different distribution. In contrast, F-score, accuracy and precision metrics will change if the distribution is altered. Furthermore, the ROC is separated from the cost context. Hence, we could set the cutoff according to an accepted number of errors. In paper **V**, we use ROC on skewed data to set the cutoff based on the accepted number of error and the size of the data set.

Chapter 4

Present investigation

This chapter summarizes the findings in the papers, including some recent updates of the current knowledge. In the thesis, bioinformatic techniques have been used for characterization of protein families and detection of patterns in gene expression and in polypeptide frequencies. Papers **I**–**II** describe protein family characterization. In Paper **III** a tool to analyze gene lists was developed. This tool has been used in Paper **II**. Papers **IV**–**V** describe the investigation of oligopeptide pattern occurrences and how these can be used in detection of proteins that contain translated introns. Paper **VI** documents the genome sequence database that has been used in several studies including Papers **I**, **II** and **IV**.

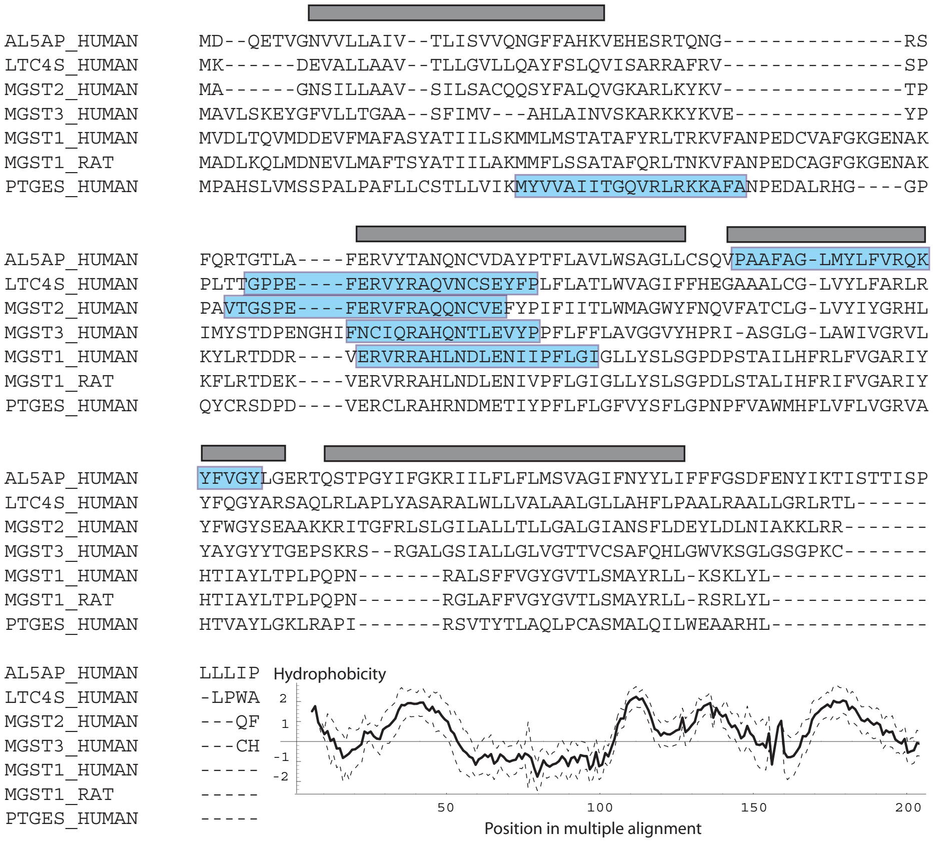
Paper I Characterization of the MAPEG superfamily

The membrane associated proteins in eicosanoid and glutathione metabolism (MAPEG) superfamily was first discovered in 1999 [122] and includes six protein families of higher eukaryotes: 5-lipoxygenase-activating protein (FLAP), leukotrine C_4 synthase (LC_4S), microsomal glutathione transferase (MGST)1–3 and prostaglandin E synthase (PGES) and additional groups of prokaryotic origin. The MSA of the human members and rat MGST1 in Figure 4.1 shows the four transmembrane alpha helices known to accommodate the common fold of the superfamily [122, 123]. The proteins are membrane bound and in eukaryotes, MAPEGs are located in the endoplasmic reticulum (ER) and/or nuclear membrane. Furthermore, glutathione transferase activity has been assigned a central property for a majority of the members.

In our study we made an extensive search in UniProtKB [6], NCBI non-redundant (known to be extensively redundant [124]) and GenomeLKPG (Paper **VI**) for new MAPEG members and we found over 130 different proteins, of which only 56 were previously listed in Pfam [99, 100]. Among the new members we found fish representatives for six of the major families of MAPEG, which dates the origin of these families back to before the occurrence of early vertebrates. We also detected two distinct prokaryotic subfamilies, featuring representative proteins from *Escherichia coli* and *Synechocystis sp.*, respectively. A third, not well-defined, cluster of bacterial proteins was also discovered. The cellular role of MAPEG in bacteria is yet to be determined (no literature on the subject is found, December 2007) and no new bacterial member has been deposited in SwissProt since the publication of Paper **I**. However, one of the biological processes assigned to eukaryotic glutathione S-transferase (GST) is to participate in detoxification of endogenous and exogenous electrophilic compounds and,

Family	Pattern
FLAP	`P-A-A-F-A-G-x(0,1)-L-x(0,1)-Y-L-x(2)-R-Q-K-Y-F-V-G-Y`
LC4S	`G-P-P-E-F-[DE]-R-[IV]-[FY]-R-A-Q-[AV]-N-[CS]-[ST]-E-Y-F-P`
MGST1	`E-R-V-R-R-[ACG]-H-x-N-D-[IL]-E-N-[IV]-[IV]-P-F-[FLV]-[AGV]-I`
MGST2	`V-[ST]-G-[APS]-[LP]-[DE]-F-[DE]-R-x-F-R-A-x(0,1)-Q-x(0,1)-N-[CNS]-[ALV]-E`
MGST3	`F-N-C-[AIV]-Q-R-[AGS]-H-[AQ]-[NQ]-x(2)-E-x(2,3)-P`
PGES	`M-Y-[AIV]-[IV]-A-[IV]-I-T-G-Q-[IMV]-R-L-R-[KR]-K-A-x-A-N`

Table 4.1. Diagnostic patterns of MAPEG members from Paper I. The patterns are described according to the Prosite pattern convention (section 3.2.3).



```
AL5AP_HUMAN     MD--QETVGNVVLLAIV--TLISVVQNGFFAHKVEHESRTQNG---------------RS
LTC4S_HUMAN     MK------DEVALLAAV--TLLGVLLQAYFSLQVISARRAFRV---------------SP
MGST2_HUMAN     MA------GNSILLAAV--SILSACQQSYFALQVGKARLKYKV---------------TP
MGST3_HUMAN     MAVLSKEYGFVLLTGAA--SFIMV---AHLAINVSKARKKYKVE--------------YP
MGST1_HUMAN     MVDLTQVMDDEVFMAFASYATIILSKMMLMSTATAFYRLTRKVFANPEDCVAFGKGENAK
MGST1_RAT       MADLKQLMDNEVLMAFTSYATIILAKMMFLSSATAFQRLTNKVFANPEDCAGFGKGENAK
PTGES_HUMAN     MPAHSLVMSSPALPAFLLCSTLLVIKMYVVAIITGQVRLRKKAFANPEDALRHG----GP

AL5AP_HUMAN     FQRTGTLA----FERVYTANQNCVDAYPTFLAVLWSAGLLCSQVPAAFAG-LMYLFVRQK
LTC4S_HUMAN     PLTTGPPE----FERVYRAQVNCSEYFPLFLATLWVAGIFFHEGAAALCG-LVYLFARLR
MGST2_HUMAN     PAVTGSPE----FERVFRAQQNCVEFYPIFIITLWMAGWYFNQVFATCLG-LVYIYGRHL
MGST3_HUMAN     IMYSTDPENGHIFNCIQRAHQNTLEVYPPFLFFLAVGGVYHPRI-ASGLG-LAWIVGRVL
MGST1_HUMAN     KYLRTDDR----VERVRRAHLNDLENIIPFLGIGLLYSLSGPDPSTAILHFRLFVGARIY
MGST1_RAT       KFLRTDEK----VERVRRAHLNDLENIVPFLGIGLLYSLSGPDLSTALIHFRIFVGARIY
PTGES_HUMAN     QYCRSDPD----VERCLRAHRNDMETIYPFLFLGFVYSFLGPNPFVAWMHFLVFLVGRVA

AL5AP_HUMAN     YFVGYLGERTQSTPGYIFGKRIILFLFLMSVAGIFNYYLIFFFGSDFENYIKTISTTISP
LTC4S_HUMAN     YFQGYARSAQLRLAPLYASARALWLLVALAALGLLAHFLPAALRAALLGRLRTL------
MGST2_HUMAN     YFWGYSEAAKKRITGFRLSLGILALLTLLGALGIANSFLDEYLDLNIAKKLRR-------
MGST3_HUMAN     YAYGYYTGEPSKRS--RGALGSIALLGLVGTTVCSAFQHLGWVKSGLGSGPKC-------
MGST1_HUMAN     HTIAYLTPLPQPN-------RALSFFVGYGVTLSMAYRLL-KSKLYL-------------
MGST1_RAT       HTIAYLTPLPQPN-------RGLAFFVGYGVTLSMAYRLL-RSRLYL-------------
PTGES_HUMAN     HTVAYLGKLRAPI-------RSVTYTLAQLPCASMALQILWEAARHL-------------

AL5AP_HUMAN     LLLIP
LTC4S_HUMAN     -LPWA
MGST2_HUMAN     ---QF
MGST3_HUMAN     ---CH
MGST1_HUMAN     -----
MGST1_RAT       -----
PTGES_HUMAN     -----
```

Figure 4.1. MSA of the MAPEG superfamily. The sequences are FLAP (AL5AP), LC_4S (LTC4S), MGST1–3 and PGES (PTGES) from human and MGST1 from rat. The common fold of the superfamily is shown by the four hydrophobic alpha helices (grey rectangles) and the corresponding hydrophobicity plot [73] (bottom). The diagnostic patterns in table 4.1 are blue boxed.

consequently, this may be true also for bacteria [125].

Structures of membrane proteins are difficult to obtain and by the time of Paper **I** only a low-resolution (6 Å) structure was available [123], which is not sufficient to determine the active sites. However, our extensive sequence analysis revealed signaturing patterns for all of the eukaryotic families (Table 4.1) and we suggested that these regions are part of the active site.

In recent years, high resolution structures have become available [126–128], which allow mapping the diagnostic patterns onto the structure. MAPEG proteins form homotrimers (Figure 4.2A) where glutathione is located between adjacent chains. Figure 4.2B shows the conserved motif of

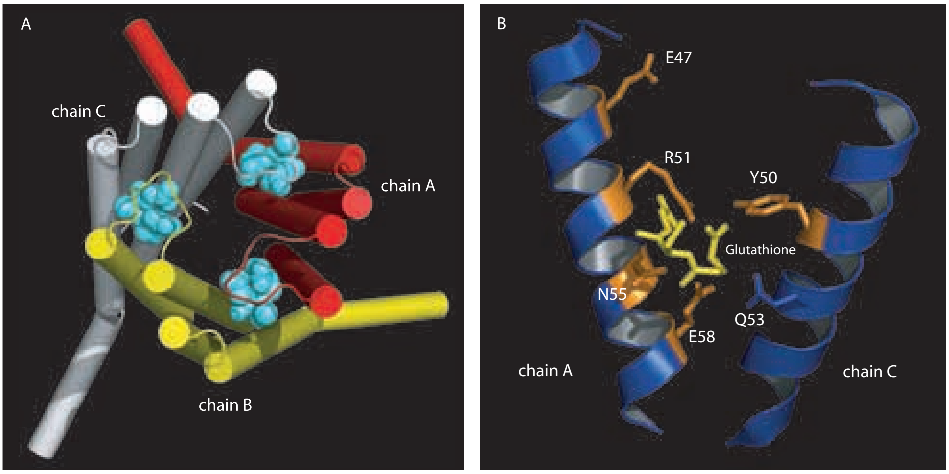


Figure 4.2. MAPEG trimer and the active site of LC$_4$S. (**A**) The structure of MAPEG shows that each monomer consists of four transmembrane helices (cylinders) and an additional helix at the C-terminus. The glutathione (cyan) at the active site is located between two adjacent monomers. (**B**) The structural elements of the diagnostic patterns of LC$_4$S with the completely conserved residues in orange. The distances between the glutathione and the neighboring side chains R51, N55 and E58 from chain A and Y50 and Q53 from chain C are between 2 and 4 Å.

LC$_4$S in detail where the majority of the completely conserved residues (orange) of the pattern are in a vicinity of 2–4 Å from the bound glutathione. The LC$_4$S pattern is located in the second helix and surprisingly, due to the organization of the monomers, both sides of the helix are part of the active site. Similarly, the pattern of MGST1 could also be mapped onto the monomeric structure (Figure 4.3A). The monomeric structure shows the transmembrane regions in red, the loops in green and the region of the pattern in blue which overlaps with glutathione.

In contrast to what has been observed for other MAPEG members, FLAP has not been shown to possess enzymatic activity or to be functionally modulated by glutathione. It is known that it activates 5-lipoxygenase via physical interaction but it is not understood how [129]. In Figure 4.3B, the homotrimer of FLAP shows the monomers colored according to Figure 4.2A. The inhibitor MK-591 binds to the region corresponding to the active site of MGST1 and LC$_4$S, while the signifying pattern of FLAP (as chosen in Paper **I**) is located on the outside of the trimer facing the hydrophobic core of the membrane.

Intriguingly, postulating that the diagnostic pattern (blue) is important

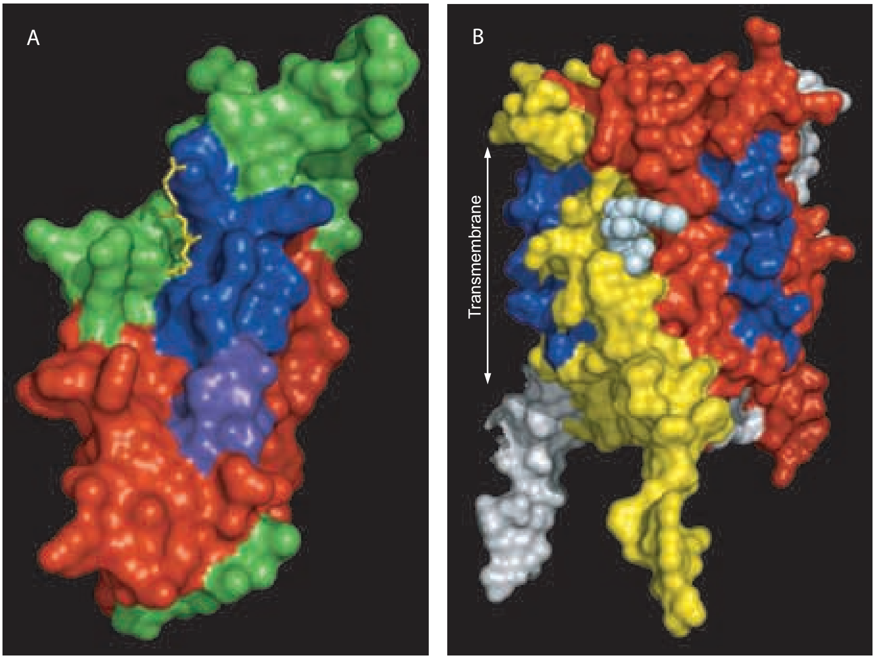


Figure 4.3. MSGT1 monomer and FLAP trimer. (**A**) The red regions of MGST1 correspond to the hydrophobic transmembrane passages while green corresponds to the cytosolic (top) and lumen (bottom) hydrophilic loops. Blue and purple shows the region of the diagnostic pattern determined in paper **I**, where the purple region indicates an overlap with the hydrophobic region. The glutathione (yellow) fits well into the pocket of the diagnostic pattern of MGST1. (**B**) FLAP does not have enzymatic activity but is known to activate 5-lipoxygenase. The process is inhibited by MK-591 which is shown as spheres (cyan). However the pocket where the inhibitor binds is not part of the diagnostic pattern (blue) of FLAP which region is faced to the membrane interior.

for the physical interaction with 5-lipoxygenase, it is interesting to investigate the possible effects on this region due to the binding of the inhibitor. In Figure 4.4, a region (orange) in close vicinity of both the diagnostic pattern and the inhibitor of FLAP is part of an alpha helix in the corresponding region of LC_4S. This region is also a conserved and specific motif (QSTPGxxFGKR) of the FLAP family, which is located immediately upchain of the diagnostic pattern. Hence, it is likely that these two adjacent sequence motifs are responsible for the interaction between FLAP and 5-lipoxygenase. If so, the abolished physical interaction might be due to the possible loss of helical structure upon binding the inhibitor MK-591.

Our suggestion that the diagnostic patterns presented in Paper **I** are

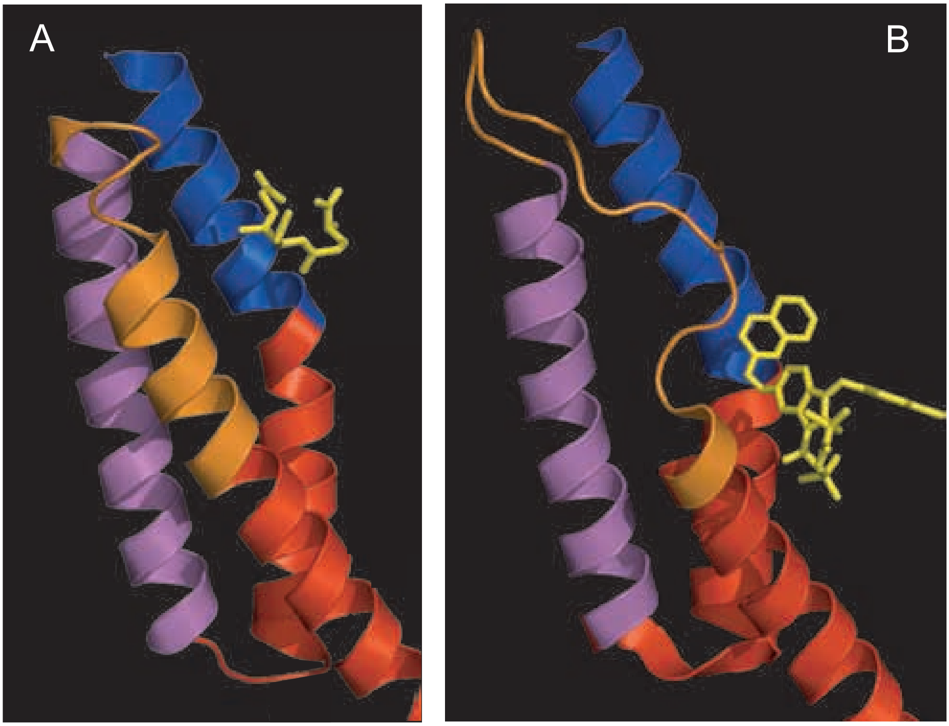


Figure 4.4. Structural comparison of inhibited FLAP and LC$_4$S. The purple colored helices indicate the regions corresponding to the diagnostic pattern of FLAP and the blue helix indicates the region of the MGST1 and LC$_4$S diagnostic patterns (Paper **I**). The orange region is a conserved motif in FLAP which corresponding region is helical in LC$_4$S (**A**). When FLAP is bound to the inhibitor MK-591 this region is not helical (**B**).

part of the active sites has been confirmed upon the publication of the structures of MGST1 [126] and LC$_4$S [127]. FLAP is a protein without enzymatic activity, which had no overlap between our pattern and the site of inhibition in the structure [128]. However, if our FLAP-pattern is extended to include another conserved motif that is located immediately upchain of the one we presented earlier, it overlaps with a region that may be structurally affected by the inhibitor. Hence, the predictions of active sites we made before the high resolution structures were known still holds, and in addition, the extended FLAP-pattern may provide valuable information on the physical interaction between FLAP and 5-lipoxygenase that is yet to be characterized.

Paper II TRIM, Ro52 and the RBL region

The TRIM protein family is defined by the N-terminal domain organization, which consists of a RING domain followed by one or two B-box domains and a coiled-coil region (Figure 4.5) [130,131]. The B-box is always of type 2 if only one copy is present and in the case of two B-boxes, the first one is always of type 1, followed by a B-box of type 2. Due to the modular structure the family is also denoted RING/B-box/coiled-coil (RBCC). The TRIM proteins are involved in various cellular processes including apoptosis, cell cycle regulation and viral response and the conserved N-terminal domain architecture has been suggested to be implicated in E3-ligase ubiquitination [131]. The region between the RING and B-box, denoted RING - B-box linker (RBL), is disease associated in at least two members of the family; TRIM5α [132] and Ro52. The latter of these is investigated in Paper **II** and the Ro52-RBL is known to have a Sjögren syndrome disease-associated autoantigenic epitope [133]. The RBL region is also known to be important for the stability of the RING-B-box region and features a completely conserved aspargine. The aim of Paper **II** was to characterize the RBL region and to determine whether the RBL region is more tightly associated to the RING or the B-box. By calculating the hydrophobic moment of the RBL region and performing PCA we determined three distinct subtypes of TRIM proteins, denoted A, B and C.

The RING and B-box domains bind Zn^{2+} and each type of domain is characterized by a unique ligand-binding pattern. The residues that bind the ligands are either cysteines or histidines. The RING domain has a well-defined C3HC4 pattern, while the B-box motifs are modeled slightly different in Pfam [99–101], Paper **II** and SwissProt [6, 7] depending on

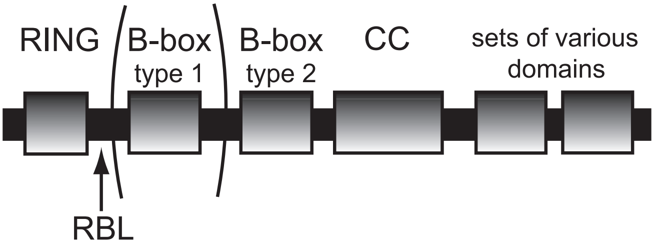


Figure 4.5. Domain organization of TRIM proteins. TRIM proteins contain a RING, one or two B-box domains and a Coiled-coil region (CC). If a protein only has one B-box it is of type 2 and if it has two B-box domains the first one is of type 1 and the second is of type 2. The region of interest in Paper **II** is the linker between the RING and B-box domains, denoted RING - B-box linker (RBL).

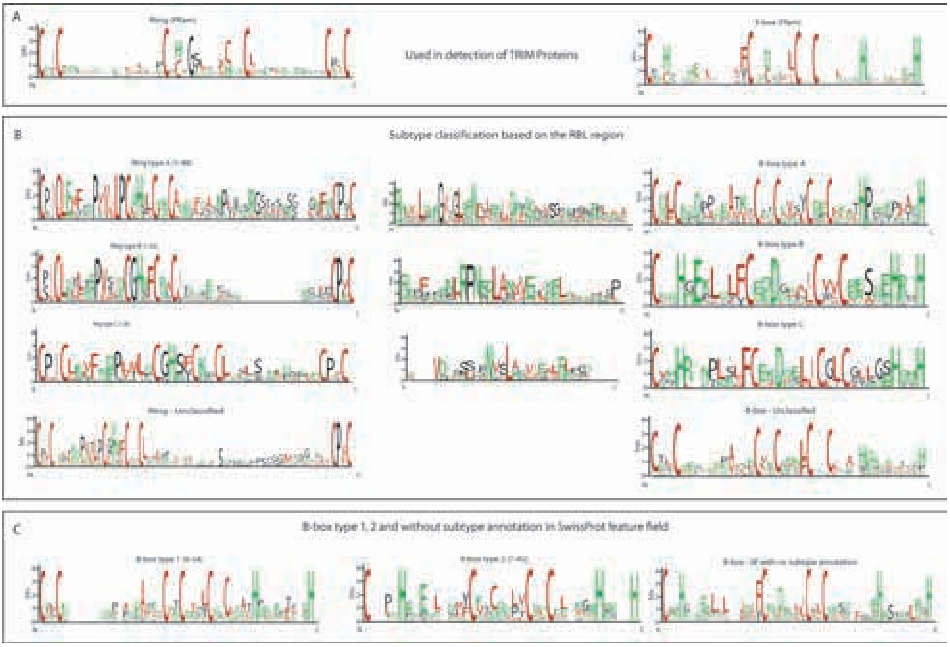


Figure 4.6. Subtype sequence logos of RING, RBL and B-box. (A) Pfam models of RING and B-box used in Paper **II** to find the boundaries of the RBL region. **(B)** The subtypes discovered in Paper **II**, which were based solely on the RBL region. **(C)** the subtypes of the B-box domain according to sequence feature annotations in SwissProt [6, 7].

which subfamily it is trained on (Figure 4.6). The B-box model in Pfam (Figure 4.6A) is trained on both type 1 (C6H2) and type 2 (CHC3H2). However, the Pfam model is biased towards the latter because B-box type 2 is always present in TRIM proteins and it is therefore more frequent in comparison to type 1. The RBL type A, defined in Paper **II**, occurs in TRIM proteins having a B-box type 1, while subtypes B and C correspond to proteins that only have a type 2 B-box.

A few differences of the non-Zn^{2+} binding sites in the RING and B-box are observed. In type A proteins, a completely conserved amino acid triplet motif Leu-Pro-Cys is observed at the site of the third cysteine in the RING domain. Furthermore, there is a completely conserved glycine in the RBL region and the otherwise hydrophilic B-box domain features a well conserved hydrophobic alanine between the second and third cysteine. Type B proteins also contain completely conserved triplet motifs; Cys-Gly-His that includes the third and fourth ligand-binding sites of the RING and Arg-Pro-Asn in the RBL region. Furthermore, glutamate seems to be

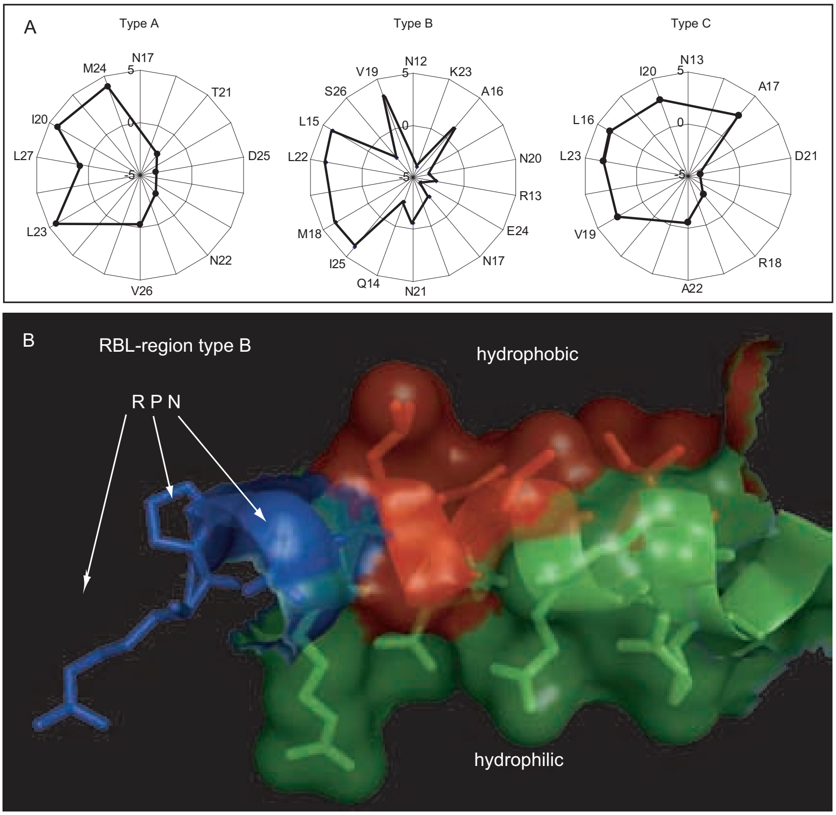


Figure 4.7. The amphipathic helix of the RBL region. (A) Helical wheel models based on the Kyte and Doolittle scale [73]. The asparagine (N), which is conserved in all RBL subtypes, is used as index reference. **(B)** Structural model of the RBL region in type B proteins. The conserved triplet arginine, proline, asparagine N-terminally of the helix is colored blue, the hydrophobic region is red and the hydrophilic is green.

preferred two positions downchain of the first histidine of the B-box type B. RING type C resembles that of type B, but their RBL regions are very different. The unclassified proteins in Paper **II** seem to be of type A/1, with respect to the conserved ligand-binding pattern of the B-box domain.

Although the differences in the RBL region of type A, B and C are difficult to quantify, they seem to be of crucial importance, especially between type B and C, which flanking regions are more similar than the corresponding RBL region. In paper **II**, we discovered that an amphipathic α-helix is present in all subtypes of the RBL region. The sequence motifs of the three subtypes are different, however all possess similar amphipathic char-

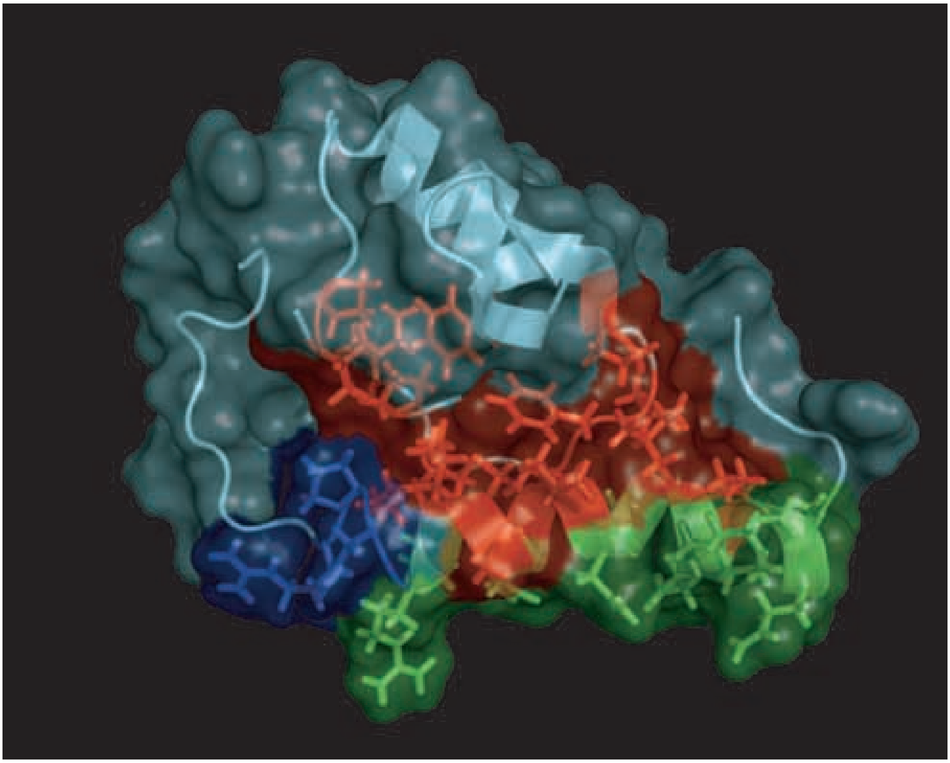


Figure 4.8. Model of RING domain and RBL region of Ro52. The completely conserved triplet (Arg-Pro-Asn) of the RBL-B is exposed on the surface (shown in blue, bottom left) and the amphipathic α-helix of the RBL is located at the bottom right, showing the hydrophobic core in red and the hydrophilic exterior in green.

acteristics (Figure 4.7A). In Figure 4.7B, a structural model of RBL-B is shown with the side chains colored according to their hydrophobicity.

Taxonomic studies using the GenomLKPG sequence database (Paper **VI**) showed that type A, B and C proteins are found in mammals, opossums, birds and frogs; type A is also found in fish and sea urchins. Type B proteins, including Ro52, are more frequent in mammals than type A, while type C is predominantly distributed among primates. This suggests that subtype A is the oldest form of TRIM proteins and subtype C is the youngest. Interestingly, the older type 1/A B-box is not present in all TRIMs in contrast to the younger type B/C/2 B-box. Hence, if the dual B-box motif is a result of domain duplication it seems likely that the type A/1 B-box has been lost later on in proteins that contain only the type B/C/2 B-box. Taking both the sequence and the taxonomical analysis in consideration, it seems likely that the difference in type A proteins compared to the others is due to the B-box type A/1 domain.

The TRIM subtypes were also investigated for bias in functional annotation using the Ontology Annotation Treebrowser (OAT), a tool described in Paper **III**. Nine of the 25 functionally annotated human, rat and mouse type A proteins were significantly associated to the cytoskeleton (p-value < 0.01). Type B proteins had a significant enrichment of annotations to nucleic acid binding activity (p-value <0.01) for 7 of 33 annotated proteins, where Ro52 is part of that group although it is known that it does not bind DNA [134]. There was no overlap between the type A and type B proteins for these two annotation terms. Although most members are not associated to these annotations, it could indicate a difference in functionality between the two types. Unclassified proteins show some bias towards DNA-binding and regulation of transcription, but for type C proteins no enrichment of annotations was observed other than metal binding, a feature common to all TRIM proteins.

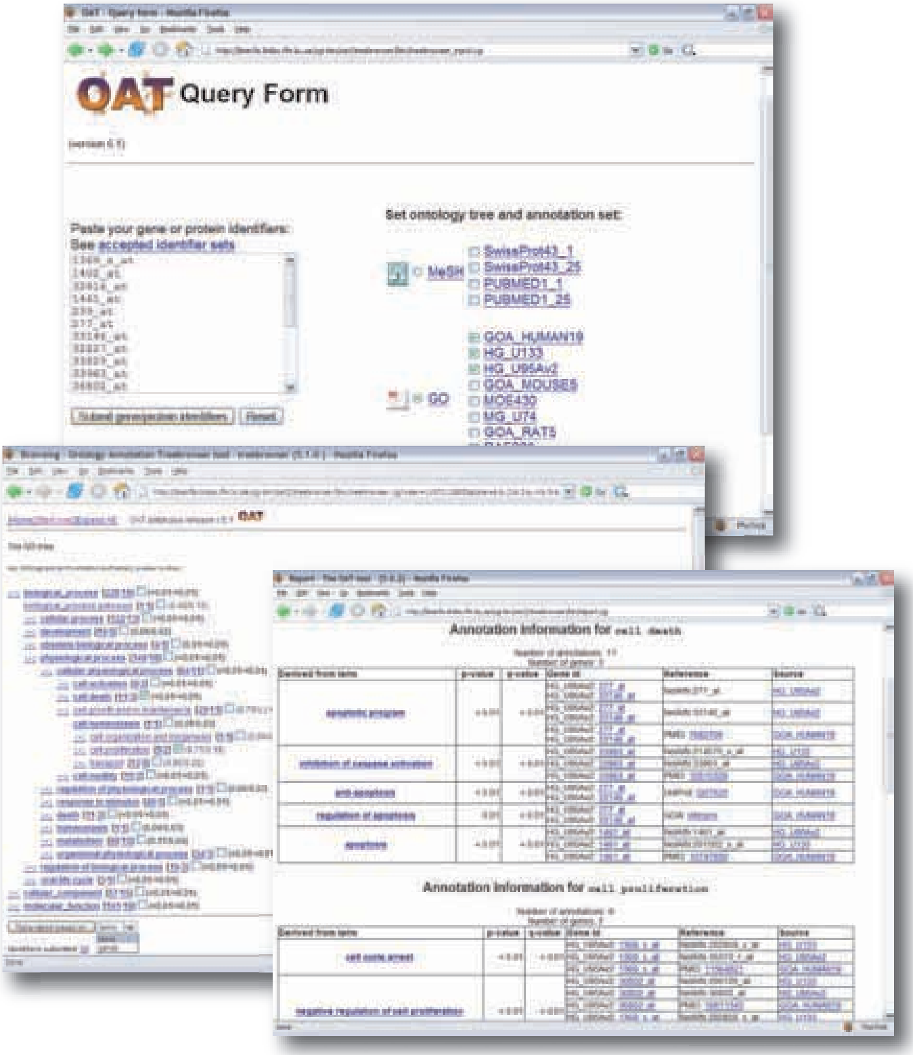
From the results of the bioinformatic and biophysical characterization of the N-terminal domains in Paper **II**, it was concluded that the RBL region of TRIM type B proteins was more closely connected to the RING in comparison to the B-box. A simulated docking of the RING and RBL region is shown in Figure 4.8, where the hydrophobic side chains of the RING and amphipathic helix of the RBL are directed towards the core of the model, which supports the increased stability that is observed for the RING and RBL when expressed together in comparison to when they are expressed separately.

Paper III OAT: Ontology Annotation Treebrowser

Life science research does not longer only apply to single-gene problems, and we have seen a shift towards the systems biology approach that implies that many thousands of biomolecules are considered at once (section 1.3.5). The explosion of sequence-related data (section 1.2.1) and development of methods for high-throughput analysis (e.g. microarrays and proteomics) have resulted in a new challenge within the life science discipline. Due to the complexity of biological processes, systemic analyses frequently result in long list of genes or proteins that are considered important. Putting the problem of accurately isolating the important biomolecules aside, the information associated with the resulting list is still massive and it is a challenge to understand what the genes in the list do and to conclude their common properties. There is evidently a need of systemizing and condensing the huge amount of knowledge and this issue is addressed by a number of tools, including our own development Ontology Annotation Treebrowser (OAT) (Paper **III**). A list of these tools are found at the web site of GO [135], all using GO [23,25] in one aspect or another. Furthermore, numerous tools provide analysis of gene lists and the association to scientific literature or MeSH [17] (discussed in Paper **III**). These two ontologies are described in more detail in section 1.3.1.

The purpose with OAT was to develop a tool to analyze annotations associated with a list of genes or proteins in an exploratory way. The simplistic user interface and the web server environment enable easy access to all the various resources available online. The use of OAT is described in three steps (Figure 4.9); i) submit a list of gene or protein identifiers and choose a list of annotation sets, ii) browse the annotations in the ontology tree and choose the interesting branches and the level of details and iii) summarize the annotations and gene identifiers as a table of web links for further investigation. At each level of detail, the enrichment of annotations is illustrated by the number of gene identifiers and annotations that are found in the branch. Using binomial distribution, we perform a statistical test to check whether the number of annotations is large enough to be significant. As discussed in section 3.4.1, the multiple comparison problems of using classical hypothesis testing comes into play and we addressed this by calculating the q-value, which is the False Discovery Rate (FDR)-analog of the p-value.

The major novelty of OAT was the incorporation of both MeSH and GO in the same tool, which allowed us to compare the two ontologies from



OAT Query Form

Paste your gene or protein identifiers:
See accepted identifier sets

Set ontology tree and annotation set:

Annotation information for cell death

Annotation information for cell proliferation

Figure 4.9. OAT user example. The top window shows the query form where the user enters the whitespace-separated gene list and selects among ontology and annotation sets. The lower left window shows the browsable tree. The first number within brackets is the number of annotations in the branch; the second is the number of genes. On the right hand side, enclosed by parentheses, a *p*-value and a *q*-value are shown for having that many or more annotations. The lower right window shows the report page listing the ontology term, *p*-value, *q*-value, gene identifiers grouped by the internal representation of OAT identifier, and origin of annotation (database or article and name of annotation set).

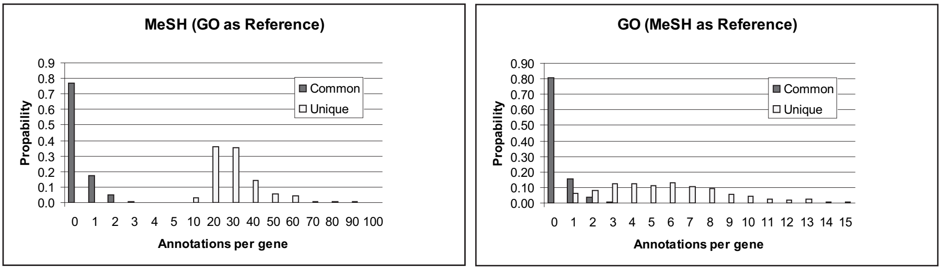


Figure 4.10. Complementarity of MeSH and GO. The bar diagrams show the probability mass function of common and unique annotations for MeSH and GO. The probability mass functions for unique annotations are clearly shifted to the right compared with annotations in common. Consequently, more unique annotations than common ones are expected. Data shown is for 721 randomly collected genes with annotations to MeSH (left) or GO (right). The assessment of annotations was made by mapping the terms to UMLS® [136].

a gene list/annotation perspective. Using GO annotations, a gene or a protein can be studied in terms of its function or process it is involved in, while MeSH has a wider scope and associates the genes and proteins to biomedical literature references that usually provides much more in-depth knowledge than the GO annotation. One may suspect that the annotations of the two ontologies are overlapping as GO annotations might be derived from knowledge published in scientific literature and vice versa, but in Paper **III** we prove them to be highly complementary. Figure 4.10 shows the distribution of common and unique terms between 721 randomly collected gene identifiers.

The MeSH annotations is derived in-house by mining PubMed [16, 20] entries to obtain the connection between a gene or protein identifier and the MeSH term. The problem with creating annotations in this way is that a scientific article might not necessarily describe only one gene. If multiple genes are described in the same article, it cannot be stated which gene is associated with a specific MeSH term. Indirectly, the MeSH term that is associated with one of the genes may also be valid for another gene in the same article. At least, by the argument that if they are mentioned in the same article there should also be an observed biomedical association between them. However, these types of annotations may become inexact and can also be due to negative association (e.g. ”in this paper we show that gene A is not co-regulated with gene B”). Nevertheless, it is important to separate indirect annotations from direct annotations. We therefore separated MeSH annotations that are derived from articles dealing with only one gene from those that deal with multiple genes. Articles dealing with more than five genes were discarded in order to avoid unspecific annotations. Thus, we arrived at one annotation set built on single gene articles and one set built on articles of 2–5 genes.

The peer-reviewed scientific articles in PubMed represent the summary of current knowledge and we would not like to exclude this exhaustive resource of knowledge. However, two issues must be pointed out; i) old articles can be outdated and erroneous, as in the case of Ro52 that has been described as DNA binding, an association that later on turned out to be wrong [134], and ii) the use of MeSH is intended to describe what an article is "about", which is a weaker association criterion than GO's *is a* and *part of* relationships. In the current release of OAT there is no information about how old the article is that the annotation is based on. Each annotation can be checked manually by accessing the PubMed entry, but there is currently no systematic way to exclude older articles. It would be possible to collect this information when generating the MeSH annotations. However little is known about how incorrect older articles might be. Although, it might be tractable to have some kind of weighting scheme that corresponds to the likeliness of an annotation to be correct. Currently, the reliability of MeSH annotations is only quantified by the number of times a connection between a gene and a descriptive MeSH term has been made. Such a reasoning seems scientifically sound, assuming that parallel observations of several individual research groups make the association more reliable than if it has been observed only once. In GO, the annotations are categorized based on the type of evidence ranging from *inferred from electronic annotation* (IEA), which is considered not very reliable, to *traceable author statement* (TAS), which is one of the most strongest types of evidence. The evidence codes are currently not utilized in OAT. However, as these provide valuable information they should be implemented in future releases.

Paper IV Characterization of oligopeptide patterns in large protein sets

Sequence analysis is frequently used either to infer function (by homology) or to determine sites in the sequence that are of importance for structure or function. In both cases, the sequence is analyzed with respect to a protein family. Such biased analyses are of major importance in bioinformatics but are not suitable for drawing general conclusions on the occurrence of patterns in full sequence databases. In Paper **IV**, we investigated all proteins regardless of their family membership instead of using a protein family centered approach. Thus, aiming at characterizing oligopeptide patterns in an unbiased and unsupervised manner.

Oligopeptide patterns have been analyzed in a range of studies including taxonomical, functional and structural investigations [137–141] and especially in Prosite [65–67]. Hence, it is beyond doubt that short oligopeptide patterns carry information. Consequently, many patterns are either expected to be over- or under-represented. In Paper **IV**, we characterized pentapeptide patterns of four categories and focused on the 100 most extreme cases in each category. The categories were:

positively selected patterns (POPs) which are the most abundant peptide patterns in observed data and are found not at all or only occasionally in randomized data. They are expected to contain favored structural or functional motifs which might be associated with large protein families. They are expected in low numbers in view of their amino acid compositions but are in fact over-represented and must therefore result from positive selective pressure.

negatively selected peptides (NEPs) are those with extremely low abundance in available protein data but with high frequencies in randomized data. NEPs are expected to result from negative selective pressure and can be explained as structurally disfavored building blocks.

over-represented peptides (ORPs) are the most frequent kingdom-specific peptide patterns. ORPs are unique to a kingdom and might be used as diagnostic patterns. They will cause bias in databases that do not have equal portions of proteins from the three kingdoms.

under-represented peptides (URPs) are those with extremely low abundance in a particular kingdom. URPs can be parts of epitopes

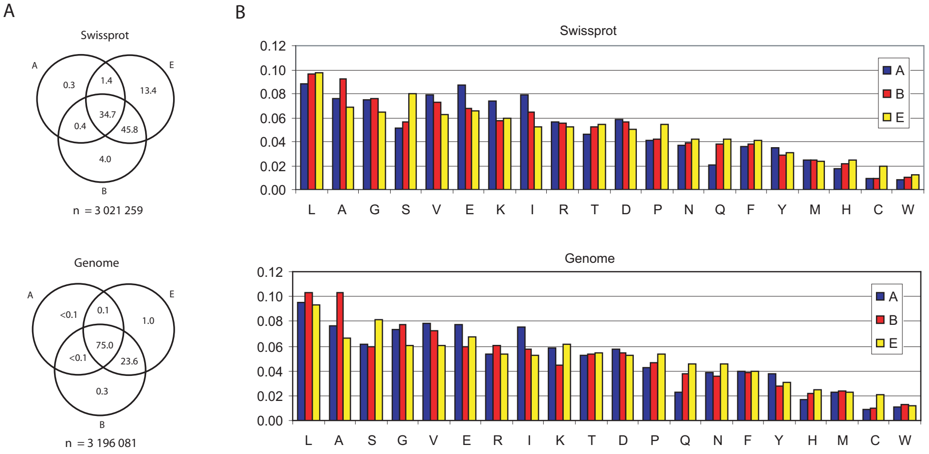


Figure 4.11. Common pentapeptide patterns and relative residue composition. (A) The Venn diagrams show the percentage of peptide patterns common to the kingdoms in the original sequence sets (A=archaea, B=bacteria and E=eukaryota). Only few peptide patterns are unique to a kingdom in the genome data set. As many as 98.7% of the peptide patterns are common to two or more kingdoms in the genome set, while the corresponding number for SwissProt is 82.3%. **(B)** The bar diagrams show the overall relative amino acid compositions for each kingdom in the two data sets, ordered by their average frequencies in the respective data set.

that are inappropriate to the kingdom or avoided for other reasons and, as for the ORPs, this will lead to bias in protein databases.

The pentapeptide abundance was analyzed in the data sets of SwissProt [7] and GenomeLKPG (Paper **VI**). On the global level we observed that only a few percent of the patterns are unique to a kingdom (Figure 4.11A). In the SwissProt and genome sets, 35% and 75% respectively, of the patterns were common for all kingdoms. Hence, the number of ORPs and URPs is expected to be low. Nevertheless, we observed an extensive overlap between bacterial ORPs and eukaryotic URPs, where many of the proteins have translation machinery and biosynthesis as common themes. Interestingly, several of these patterns (and the respective proteins in which they occur) are related to immune response activity and are suggested as therapeutic targets (see Paper **IV** for details).

Several of the POPs were motifs in large protein families, especially cytochromes in eukaryotes and bacteria. Many were also frequently annotated with Zn^{2+} or metal binding. One may speculate that zinc ligation, which frequently occurs in eukaryotes [142], is an important contributor

Peptide		Memory requirement		
length (n)	Patterns	Counter[a]	Data structure[b]	Smart encoding[c]
3	8 000	32 kB	88 kB	16 kB
4	160 000	640 kB	1.92 MB	320 kB
5	3 200 000	12.8 MB	41.6 MB	6.4 MB
6	64 000 000	256 MB	896 MB	128 MB
7	1 280 000 000	5.12 GB	19.2 GB	2.56 GB
8	25 600 000 000	102.4 GB	409.6 GB	51.2 GB

[a] long integer (4 Bytes)

[b] A dictionary type of patterns and occurrences:
(string (n bytes)+address (4 Bytes) + counter (4 Bytes))x nbr of patterns.

[c] 16 bit allocation of all counters.

Table 4.2. Memory requirements. The *counter* column gives the memory size for storing occurrence of peptide patterns as integers of type long. The *Data structure* column gives the size using a high-level dictionary data structure with peptide pattern as key (string) and occurrences of patterns as value (integer). The *Smart encoding* column gives the size using a low-level data structure by allocating memory addresses based on the peptide pattern itself and a 16-bit encoded integer for the number of occurrences.

to the elevated levels of the rare cysteine residue type in eukaryotes (Figure 4.11B). An additional observation of major importance was the discovery that many of the over-represented patterns were exclusively found in species-specific multicopy retrotransposons. This bias was specific for the translations in the genome data set. Nevertheless, functional assignment is lacking for many POPs, as well as for other categories.

A practical issue that needs to be considered when collecting the data for the categories above is the memory requirement, which increases exponentially with the length of the peptide pattern. A traditional 32-bit computer system, with a theoretical maximum of 4x10^9 memory addresses, cannot store longer oligopeptide patterns than six, assuming we need a long integer (4 bytes) to store the number of observations of a pattern (Table 4.2). If we want an additional high-level dictionary data type to lookup the pattern, we need even more memory. However, it is possible to store much more data in the memory of a 64-bit system or to parallelize the search of patterns on a computer cluster. It is also possible to use low-level data structures by allocating memory addresses based on the peptide pattern itself and to store the integer in a sufficiently large binary structure. As an example, an integer of max 65 535 can be stored in only 2 bytes, while a long integer, a memory address and a sequence pattern of length 5 needs 13 bytes. Although, the low-level storage approach uses approximately 6 times less memory, it will only lead to a linear improvement, in contrast to the exponential increase of number of patterns used in studies of longer oligopeptides. It is probably possible to analyze longer patterns by utilizing computer cluster with shared memory, but the increase of possible oligopeptide patterns at those lengths outnumbers the performance of high-end computer solutions. Fortunately, we show that from an informational content perspective occurrences of oligopeptide patterns are optimally studied at the pentapeptide level, of which only few patterns are never observed (Paper **IV**). It is possible to also analyze hexapeptides but at the cost of longer computational time. However, analyzing hexapeptides would lead to a partitioning of the occurrences of heptapeptides into 20 subcategories of the additional sixth residue. Hence, we would analyze the same information but at a more detailed level. Furthermore, for a majority of the longer oligopeptides, the number of occurrences will not be sufficient in order to make statistically reliable interpretations. This fact needs to be considered also on the pentapeptide level. We addressed this by an initial filtering step in order to avoid including patterns that do not exist in real proteins but which are discovered occasionally due to possible sequencing errors.

In summary, three major classes of pentapetides were observed: (i) patterns widespread in a kingdom such as those originating from respiratory chain-associated proteins and translation machinery; (ii) proteins with structurally and/or functionally favored patterns, which have not yet been ascribed a role; (iii) multicopy species-specific retrotransposons, only found in the genome set. The three categories will affect the accuracy of sequence pattern algorithms that rely mainly on amino acid residue usage. Interestingly, one of the most important findings of this study is the type (ii) category patterns that are in majority. These are not annotated in the sequence feature tables. However, these frequently occurring POPs seem to be of major importance since they are not expected at all. It would be very interesting to characterize these patterns further. Methods presented in Paper **IV** may be used to discover targets for antibiotics, as we identify numerous examples of kingdom-specific antigens among our peptide classes. The methods may also be useful for detecting coding regions of genes, an approach we proved successful in Paper **V**.

Paper V The *i*-score method

It is now possible to determine the complete genomic sequence of an organism at a relatively low cost [14]. The primary challenge is not longer to obtain the complete genomic sequence. The difficulties we are faced with are how to accurately detect the genes and the protein coding regions within the genome. Finding the region of a gene is now fairly straightforward with an accuracy of 90% at the nucleotide level [94]. This is done by mapping promotors and poly-A signals, EST evidence and homologous protein sequences onto the DNA assembly. It is much harder to determine the exact gene structure, for example regarding exon/intron boundaries and splice variants [143]. In the human ENCODE genome annotation assessment project (EGASP), it was estimated that the exact protein coding sequence was accurately predicted only for roughly 50% of the genes [94]. In addition, as little as 2.3% of predicted exons without annotation could be validated experimentally, indicating a high false positive rate for novel protein predictions. In short, the following was concluded:

> Unfortunately, there are very few processes in place to remove erroneous sequences and annotations from the public databases, so it will still take some time to get a better picture of exact gene structures. It has to be noted that the human genome and its annotation for protein coding genes are still works in progress. [94]

In line with the challenge presented here, we developed the *i*-score method, aiming at discriminating between a correct full length protein sequence and a sequence containing translated introns. The principle idea of detection is directly related to the work in Paper **IV**, where over- and under-represented oligopeptides were characterized. The working hypothesis is that, due to the selective pressure, true exons are more likely to contain over-represented peptide patterns and less likely to contain under-represented peptide patterns. Consequently, the opposite should be true for translated introns. The *i*-score method represents a region, denoted *window*, in the amino acid sequence using the ratios of over-/under-represented tripeptide patterns in SwissProt. The more positions that are considered simultaneously (i.e. increasing the *window* length), the more likely it is to encounter extreme peptides that contain the information we need to accurately classify the regions as intronic or exonic. This is illustrated in Table 4.3 by assessing the 5-fold cross validation accuracy for various length of the window. However, the negative effect of choosing longer windows is that the averaging effect results in a less exact determination of the exon/intron boundary.

Window	20	30	40	50	70	80	100
Accuracy (%, 5-CV)	77	82	83	85	89.5	90.3	92

Table 4.3. Accuracy as a function of window size. The accuracy of detecting translated intron sequences are retrieved using 5 fold cross-validation (5-CV). The numbers are for windowed data and we chose a size of 80 (90% accuracy) to build a model to detect regions of translated introns in full length protein sequences.

Aiming at a recall above 90%, we found that a window of 80 residues was a good compromise between the sensitivity of detecting translated introns and correctly assigning the extent of the unspliced intron.

We used SVM (section 3.2.4) to discriminate between intron and exon windows. Using a sliding window approach, we defined the *i*-score as the averaged estimated probability that the region is a translated intron. The choice of using SVM techniques to solve the two class separation problem is usually based on the attractive nonlinear properties of the kernel function. However, the *i*-score method uses a linear kernel. In a sense, our approach applies a nonlinear data manipulation already at time of preparation of the input vector. The elements of the input vector represent the positions in the sequence. Each element contains information on how frequently the residue (at that position) is observed together with the residues immediately up- and down-chain of it. Hence, the nonlinear mapping into the input vector can be seen as a mapping from the 20-dimensional amino acid residue space to the 8000-dimensional space of tripeptide patterns. We tested several nonlinear kernels but obtained no improvement in comparison to the linear kernel. This is inline with the discussion in section 3.2.4, where both the number of data points and number of features is large.

In Figure 4.12, six typical prediction curves are shown for proteins that are contaminated with a translated intron. The curves are easily interpreted, where an *i*-score near one represents a region likely to contain intronic material and an *i*-score close to zero represents exonic material. The performance of the *i*-score method is illustrated with an ROC curve. From the ROC curve we can decide an appropriate *i*-score cutoff depending on the size and expected distribution of translated introns and true exons (see section 3.4.2 for discussion). With an impressive accuracy of 89% and an MCC of 0.784 on balanced data, it will preform well also in high-throughput procedures. This is proved on both simulated and real data sets. Using the *i*-score method, it was possible to detect several doubtful protein predictions in the human genome. Interestingly, several of the translations predicted to be of poor quality were confirmed to contain translated intron

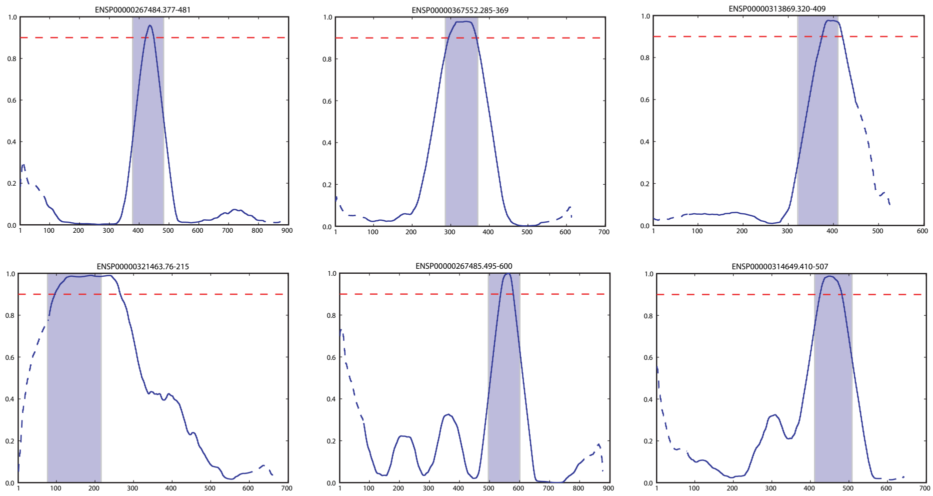


Figure 4.12. Predictions on simulated intron inserts. Typical prediction curves for six proteins with translated introns. The x-axis represents the residue position in the sequence. A high *i*-score value (blue line) indicates that the surrounding region is a translated intron. The dashed edges are regions with low confidence and the dashed red line marks an *i*-score of 0.90. The violet box indicates the inserted translated intron.

upon the ongoing manual curation of the human genome [144, 145].

In the post-genomic era, the focus has changed from finding the genes to determining the correct transcripts. The majority of the proteins from the human genome project are not yet verified by human curators. In Paper **V**, we presented an approach that can detect regions of proteins that are likely to represent translations of erroneously included introns. We proved that it performs well on both simulated and real data. We also proved that it is useful for high-throughput approaches, since it is possible to control the false positive rate. Our novel method can be used by manual annotators to provide suggestions on which part of a translation that is likely to include intronic material. It can also be incorporated into any gene prediction algorithm. Furthermore, we recommend using our method prior to building protein family models (e.g. Pfam/HMM [72, 99–101] or Prosite [65–67]) in order to remove sequences containing translated introns that would otherwise introduce noise or errors in the model.

Paper VI GenomeLKPG

In absence of fossil samples, molecular phylogenetics is widely applied in reconstruction of the evolutionary history of a protein family [146]. However, despite recent advances phylogenetic methods often suffer from incorrect branchings and branch lengths [147]. It is often more appropriate to only investigate the taxonomical occurrences among different phyla [61]. In such taxonomical studies, shortcomings of evolutionary models become less evident as conclusions rely only on known events (e.g. the origin of eukaryotes, Cambrian explosion, speciation into mammals). A tree of life based on fossil evidence has been made available [61] and a schematic view is shown in Figure 4.13.

When analyzing the spread of individual protein families among kingdoms or divisions, the bias towards well characterized proteins (e.g. UniProtKB/SwissProt [6, 7] and Pfam [99–101]) should be avoided. The solution is to use the proteomes of completely sequenced organisms, which will give a balanced picture of the occurrences of proteins. However, working with multiple complete proteome sets usually involves extensive manual pre- or post-processing of data as no individual database currently provides a comprehensive set of complete proteomes in a systematic way. Subsections of completed proteomes can be obtained from RefSeq [84], which includes only a minor set of high-quality proteomes, Ensembl [86] that is limited to eukaryotes, and the Comprehensive Microbial Resource (CMR) [148], which includes only microbial proteomes. Taxonomical grouping of proteins from a range of sequence databases is found in Integr8 [149]. However, several of the resources which Integr8 is based on (e.g. EMBL nucleotide database [5] and UniProtKB [6, 7]) have a bias towards frequently analyzed genes and protein families, which we are trying to avoid.

A full list of completed, published and ongoing genome projects is available via the Genomes OnLine Database (GOLD) [14, 15]. GOLD includes meta data, such as the institute responsible for the project, funding, numbers of chromosomes and ORFs, publication details, links to species specific database etc. Unfortunately, there is no way to systematically retrieve the protein sequences from GOLD. The increase in number of completely sequenced genomes (almost 700, November 2007 [15]) makes it a very labor-intensive task to download all proteomes manually. To accomplish this, several servers need to be accessed. Furthermore, filename conventions are not standardized between the servers, nor is the format of the individual fasta files. To address this situation, we developed the GenomeLKPG sequence database, which contains all publicly available proteins of completed genomes. The proteins are encoded by the NCBI

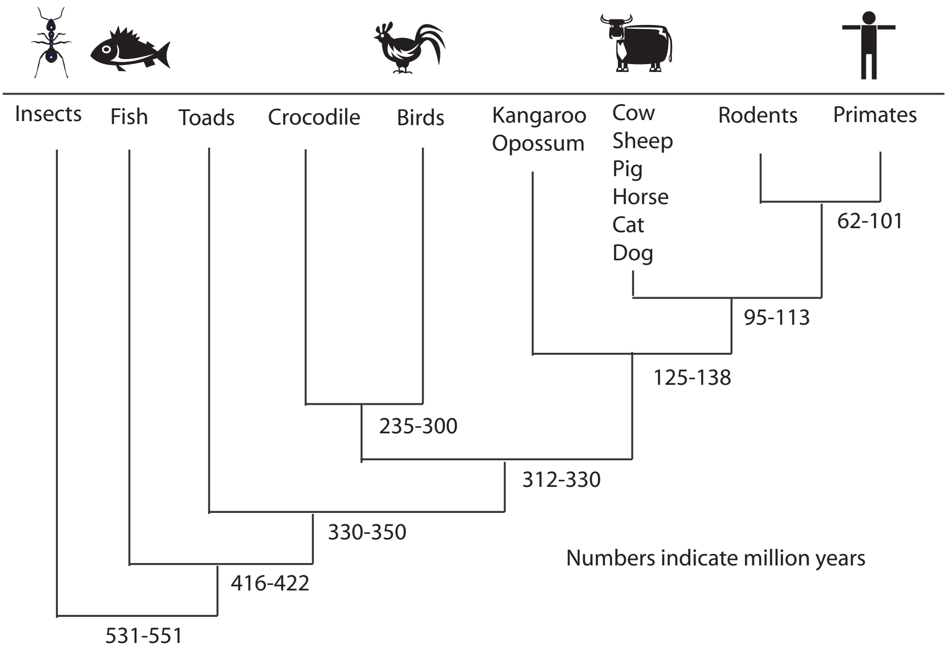


Figure 4.13. Tree of Life. This tree of life is based on paleontological evidence [61] and can be used to date the origin of a protein family. The distances are given in intervals. Note however that fossil records are recommended only for minimum constraint on the timing of lineage divergence events. Full data tables and interactive view of the tree and its evidence is found at [150].

taxonomy database [151,152], which enables easy analysis of the taxonomical occurrences of proteins among kingdoms and divisions in an unbiased manner. The database is compiled in a semiautomatic manner using FTP crawlers and an in-house version of the NCBI taxonomy database.

The genomeLKPG database has been used in several studies. In an enzymatic and bioinformatic characterization of the MAPEG superfamily (Paper **I**) the database was used to find novel members not yet published in GenBank. This led to a doubling of the number of members and with the new data it was possible by taxonomical analysis to date some of the families back to the Cambrian explosion.

In another study, oligopeptide patterns in the genomeLKPG and Swiss-Prot databases were analyzed and we found an inherent bias of certain patterns in naturally occurring proteins that could not be explained solely by the residue distribution in single proteins, kingdoms or databases (Paper **IV**). Three predominant categories of patterns were determined: (i) patterns widespread in a kingdom such as those originating from respiratory chain-associated proteins and translation machinery; (ii) proteins with structurally and/or functionally favored patterns, which have not yet been ascribed this role; (iii) multicopy species-specific retrotransposons, only found in the genome set. Especially the latter category would not have been detected without the genomeLKPG database. The three categories might affect the accuracy of sequence pattern algorithms that rely mainly on amino acid residue usage.

The genomeLKPG database was also utilized in a study of the linker region between the RING and B-Box domain of TRIM proteins (Paper **II**). The study revealed three distinct subtypes (A, B and C) of the linker region based on differences in their amphipathic alpha-helix. The three subtypes had three very distinct taxonomical patterns, where type A was found in many species ranging from mammals, opossums, birds, frogs down to sea urchins. Type B and C proteins were found in the same set of phyla except fish and sea urchins. Hence, subtype A seems to be the oldest form of TRIM proteins.

Furthermore, genomeLKPG was useful by revealing new member proteins not yet included in UniProtKB in an analysis of ancient sequence motifs in the H^+-PPase family [153]. GenomeLKPG has also contributed significantly to two ongoing studies of the medium chain dehydrogenases/reductases (MDR) and BRICHOS protein families, respectively (manuscripts in preparation).

Chapter 5

Concluding remarks

5.1 Summary

It has been an exiting five years of research using a variety of bioinformatic methods applied on important biological problems. The common theme of this thesis can be ascribed patterns in protein sequences. The patterns are ranging from short oligopeptide motifs (Paper **IV**), via intermediate motif lengths of active site and diagnostic pattern recognition (Paper **I**), even longer stretches containing translated introns (Paper **V**) and up to full length domains (Paper **II**). In two of the projects the hydrophobicity analyses have been of central importance; in the MAPEG study, the four hydrophobic transmembrane helices are a typical signature of the superfamily and in the TRIM study, an amphipathic helix was ascribed a typical feature of the RING - B-box linker (RBL) region.

Besides the biophysical properties, the abundance of patterns also was an important parameter; In the MAPEG study, the abundance of representative members from fish and bacteria provided new knowledge about the origin of the family. In the TRIM study, the abundance of the RING and B-box domains and the RBL region was analyzed and we could determine that the RBL region is unique to the TRIM proteins. In the oligopeptide study, the abundance of pentapeptide patterns was used to detect positively and negatively selected patterns, and in the *i*-score method this type of abundance was used to detect proteins containing translated introns. Abundance was also the common theme of the OAT tool, which quantified the common annotation terms of a gene list and by applying statistics we could detect enrichments of descriptive keywords. In some sense, the GenomeLKPG sequence database (Paper **VI**) also had a focus on abundance by enabling a balanced data set of proteins that can be used to analyze the abundance of patterns and proteins in various species and kingdoms.

However, the following findings are in my opinion the most important contributions to the life science community.

MAPEG The diagnostic patterns of the LC_4S and MGST1 members that later were confirmed to be part of active sites. Possibly the FLAP pattern will be an important factor when analyzing the yet uncharacterized physical interaction between FLAP and 5-lipoxygenase.

TRIM The conserved amino acid residues and the subtype identification of the RBL-region may be of crucial importance for the interaction and stability of the RING and B-box domains. Hence, these clues will be central in obtaining the structure of the N-terminal domains of the disease-associated Ro52 protein.

OAT The complementarity of MeSH and GO, which represent the current knowledge from literature and annotation projects respectively, are of importance in order to interpret gene lists.

Oligopeptides The identification of the three categories (large protein families, retrotransposons in genome data sets, and abundant patterns not yet assigned functional roles) will be of major importance in pattern detection algorithms.

The *i*-score method The impressive performance of the method will make it to an valuable tool in the post-processing of genome-level protein predictions. It will help in finding regions in proteins that need to be reannotated.

GenomeLKPG provides the first comprehensive and systematic sequence database for analysis of proteins of completed genomes and it has already been proven useful in various studies.

5.2 Future studies

The work performed in this thesis is part of a bigger picture. Here I will outline some aspects of the research that will be important in the future.

MAPEG Especially the FLAP and 5-lipoxygenase interaction would be interesting to study. Another interesting aspect is to determine the role of MAPEG proteins in bacteria.

TRIM It would be interesting to make further structural models of the B-box type 1 and 2 together with the RING and the three subtypes of the RBL region. It is also possible that additional structural clues are obtained by the current analysis on the interaction of E3-ligase and TRIM proteins using NMR spectroscopy.

OAT It would be nice to merge the MeSH and GO hierarchical structures in order to make simultaneous analyses. It would also be interesting to investigate weighting schemes for reliable MeSH annotations. Current work is focused on designing a system that facilitates automated updates on a regular basis and allows for easy addition of further ontologies.

Oligopeptides The most interesting investigation would be to see if these patterns could be explained from a structural perspective. Such an analysis has been initiated and promising results have already been obtained.

The *i*-score method This is one of the most important efforts in this thesis. The next step would be to accommodate the standards of reporting sequence features (e.g. GFF [154]), integrate other signals such as splice sites and *supporting evidence* of Ensembl. It would also be interesting to make 'hard' tests together with manual curation initiatives.

GenomeLKPG A further improvement would be to make the algorithm fully automated and to provide a dynamic download user interface in order to enable subsections of interest to be retrieved. It would also be useful to develop a more application 'friendly' version of the NCBI taxonomy (e.g. MySQL implementation), which would enable easier use, update and integration.

Appendix A

Acronyms

A	adenine
API	application programming interface
BLAST	Basic Local Alignment and Search Tool
BLOSUM	BLOcks of amino acid SUbstitution Matrix
C	cytosine
CDS	protein coding sequence
DNA	deoxyribonucleic acid
EBI	European Bioinformatics Institute
EGASP	the human ENCODE genome annotation assessment project
EMBL	European Molecular Biology Laboratory
EMBOSS	European Molecular Biology Open Software Suite
ER	endoplasmic reticulum
EST	Expressed Sequence Tag
FDR	False Discovery Rate
FLAP	5-lipoxygenase-activating protein
FPR	False Positive Rate
G	guanine
GO	Gene Ontology
GOA	Gene Ontology Annotation
GOC	the Gene Ontology Consortium
GOLD	the Genomes OnLine Database

GST	glutathione S-transferase
HMM	Hidden Markov Model
HSP	high-scoring segment pair
IPI	International Protein Index
LC$_4$S	leukotrine C_4 synthase
NCBI	National Center for Biotechnology Information
NC-IUBMB	Nomenclature Committee of the International Union of Biochemistry and Molecular Biology
NEP	negatively selected peptide
NIH	National Institute of Health
NMR	nuclear magnetic resonance
MAPEG	membrane associated proteins in eicosanoid and glutathione metabolism
MCC	Matthews' correlation coefficient
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings
MGST	microsomal glutathione transferase
mRNA	messenger RNA
MSA	multiple sequence alignment
MVDA	multivariate data analysis
OAT	Ontology Annotation Treebrowser
OBO	Open Biological Ontologies
ORP	over-represented peptide
PAM	Point Accepted Mutation
PCA	principal components analysis
PDB	Protein Data Bank
PGES	prostaglandin E synthase
POP	positively selected pattern
PSI	Position-Specific Iterated
RBCC	RING/B-box/coiled-coil
RBL	RING - B-box linker
RefSeq	NCBI's reference sequence database

RNA	ribonucleic acid
ROC	receiver operating characteristics
SAM	Sequence Alignment and Modeling System
SIB	Swiss Institute of Bioinformatics
SQL	Structured Query Language
SRS	Sequence Retrieval System
SVM	Support Vector Machine
T	thymine
TPR	True Positive Rate
TrEMBL	Translated EMBL
TRIM	Tripartite motif
UMLS	Unified Medical Language System
UniProtKB	UniProt Knowledgebase
URP	under-represented peptide
XML	eXtensible Markup Language

References

[1] Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227**(5258):561–3.

[2] **NIH Working definition of bioinformtaics and computational biology** [http://www.bisti.nih.gov/CompuBioDef.pdf].

[3] **Wikipedia - The free encyclopedia** [http://wikipedia.org/].

[4] **Wikipedia description of Bioinformatics** [http://en.wikipedia.org/wiki/Bioinformatics].

[5] Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MPG, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R: **EMBL Nucleotide Sequence Database in 2006.** *Nucleic Acids Res* 2007, **35**(Database issue):D16–20.

[6] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res* 2004, **32**(Database issue):D115–9.

[7] The UniProt Consortium: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35**(Database issue):D193–7.

[8] Etzold T, Argos P: **SRS–an indexing and retrieval tool for flat file data libraries.** *Comput Appl Biosci* 1993, **9**:49–57.

[9] Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks.** *Methods Enzymol* 1996, **266**:114–28.

[10] Zdobnov EM, Lopez R, Apweiler R, Etzold T: **The EBI SRS server–recent developments.** *Bioinformatics* 2002, **18**(2):368–73.

[11] Zdobnov EM, Lopez R, Apweiler R, Etzold T: **The EBI SRS server-new features.** *Bioinformatics* 2002, **18**(8):1149–50.

[12] Berman H, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nat Struct Biol* 2003, **10**(12):980.

[13] **PDB webpage**[`http://www.pdb.org`].

[14] Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC: **The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide.** *Nucleic Acids Res* 2006, **34**(Database issue):D332–4.

[15] **GOLD, Genomes OnLine Database** [`http://www.genomesonline.org/`].

[16] **PubMed**[`http://www.pubmed.org/`].

[17] Lipscomb CE: **Medical Subject Headings (MeSH).** *Bull Med Libr Assoc* 2000, **88**(3):265–6.

[18] Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**(6):276–7.

[19] Olson SA: **EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite.** *Brief Bioinform* 2002, **3**:87–91.

[20] Schuler GD, Epstein JA, Ohkawa H, Kans JA: **Entrez: molecular biology database and retrieval system.** *Methods Enzymol* 1996, **266**:141–62.

[21] **MEDLINE**[`http://medline.cos.com/`].

[22] Pearson H: **Biology's name game**. *Nature News* 2001.

[23] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–9.

[24] **Open Biomedical Ontologies** [`http://obo.sourceforge.net/main.html`].

[25] The Gene Ontology Consortium: **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res* 2006, **34**(Database issue):D322–6.

[26] Willingham AT, Gingeras TR: **TUF love for "junk" DNA.** *Cell* 2006, **125**(7):1215–20.

[27] Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet* 2006, **15 Spec No 1**:R17–29.

[28] Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8**(6):413–23.

[29] Brenner SE: **A tour of structural genomics.** *Nat Rev Genet* 2001, **2**(10):801–9.

[30] Chandonia JM, Brenner SE: **The impact of structural genomics: expectations and outcomes.** *Science* 2006, **311**(5759):347–51.

[31] Gileadi O, Knapp S, Lee WH, Marsden BD, Muller S, Niesen FH, Kavanagh KL, Ball LJ, von Delft F, Doyle DA, Oppermann UCT, Sundstrom M: **The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins.** *J Struct Funct Genomics* 2007, **8**(2-3):107–19.

[32] Ananiadou S, Kell DB, ichi Tsujii J: **Text mining and its potential applications in systems biology.** *Trends Biotechnol* 2006, **24**(12):571–9.

[33] Cohen AM, Hersh WR: **A survey of current work in biomedical text mining.** *Brief Bioinform* 2005, **6**:57–71.

[34] Sauer U, Heinemann M, Zamboni N: **Genetics. Getting closer to the whole picture.** *Science* 2007, **316**(5824):550–1.

[35] Henikoff S, Henikoff JG: **Performance evaluation of amino acid substitution matrices.** *Proteins* 1993, **17**:49–61.

[36] Dayhoff MO, Schwartz RM, Orcutt BC: **A model for evolutionary change in proteins.** In *Atlas of Protein Sequence and Structure*,

Volume 5, Supplement 3. Edited by Dayhoff MO, Washington DC: National Biomedical Research Foundation 1978:345–52.

[37] Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**(3):443–53.

[38] Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195–7.

[39] Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.** *Proc Natl Acad Sci U S A* 1998, **95**(11):6073–8.

[40] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–402.

[41] Moore GE: **Cramming More Components onto Integrated Circuits.** *Electronics* 1965, **April 19**:114–117.

[42] Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227**(4693):1435–41.

[43] Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**(8):2444–8.

[44] Pearson WR: **Rapid and sensitive sequence comparison with FASTP and FASTA.** *Methods Enzymol* 1990, **183**:63–98.

[45] Pagni M, Jongeneel CV: **Making sense of score statistics for sequence alignments.** *Brief Bioinform* 2001, **2**:51–67.

[46] Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673–80.

[47] Feng DF, Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees.** *J Mol Evol* 1987, **25**(4):351–60.

[48] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–25.

[49] Nuin PAS, Wang Z, Tillier ERM: **The accuracy of several multiple sequence alignment programs for proteins.** *BMC Bioinformatics* 2006, **7**:471.

[50] Katoh K, ichi Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33**(2):511–8.

[51] Morgenstern B: **DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment.** *Bioinformatics* 1999, **15**(3):211–8.

[52] Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205–17.

[53] Lee C, Grasso C, Sharlow MF: **Multiple sequence alignment using partial order graphs.** *Bioinformatics* 2002, **18**(3):452–64.

[54] Katoh K, Misawa K, ichi Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30**(14):3059–66.

[55] Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.

[56] Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment.** *Genome Res* 2005, **15**(2):330–40.

[57] Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B: **DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment.** *BMC Bioinformatics* 2005, **6**:66.

[58] Lassmann T, Sonnhammer ELL: **Kalign–an accurate and fast multiple sequence alignment algorithm.** *BMC Bioinformatics* 2005, **6**:298.

[59] Pang A, Smith AD, Nuin PAS, Tillier ERM: **SIMPROT: using an empirically determined indel distribution in simulations of protein evolution.** *BMC Bioinformatics* 2005, **6**:236.

[60] Thompson JD, Koehl P, Ripp R, Poch O: **BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark.** *Proteins* 2005, **61**:127–36.

[61] Benton MJ, Donoghue PCJ: **Paleontological evidence to date the tree of life.** *Mol Biol Evol* 2007, **24**:26–53.

[62] Knoll AH: **The early evolution of eukaryotes: a geological perspective.** *Science* 1992, **256**(5057):622–7.

[63] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25**(24):4876–82.

[64] Felsentein J: **Confidence limits on phylogenies: An approach using the bootstrap.** *Evolution* 1985, **39**(4):783–91.

[65] Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors.** *Brief Bioinform* 2002, **3**(3):265–74.

[66] Hulo N, Sigrist CJA, Saux VL, Langendijk-Genevaux PS, Bordoli L, Gattiker A, Castro ED, Bucher P, Bairoch A: **Recent improvements to the PROSITE database.** *Nucleic Acids Res* 2004, **32**(Database issue):D134–7.

[67] Hulo N, Bairoch A, Bulliard V, Cerutti L, Castro ED, Langendijk-Genevaux PS, Pagni M, Sigrist CJA: **The PROSITE database.** *Nucleic Acids Res* 2006, **34**(Database issue):D227–30.

[68] Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284**(4):1201–10.

[69] Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**(5):1501–31.

[70] Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press 1998.

[71] **SAM**[http://www.soe.ucsc.edu/compbio/sam.html].

[72] **HMMER**[http://hmmer.janelia.org/].

[73] Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157**:105–32.

[74] Jackson JE: *A User's guide to Principle Components*. John Whiley 1991.

[75] Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2**(2):171–8.

[76] Gogos A, Jantz D, Senturker S, Richardson D, Dizdaroglu M, Clarke ND: **Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: an experimental test using DNA glycosylase homologs.** *Proteins* 2000, **40**:98–105.

[77] Yang ZR: **Biological applications of support vector machines.** *Brief Bioinform* 2004, **5**(4):328–38.

[78] Schölkopf B, Smola AJ: *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts: The MIT Press 2002.

[79] Hsu CW, Chang CC, Lin CJ: **A Practical Guide to Support Vector Classification**. Guide, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan 2007.

[80] Joachims T: *Advances in Kernel Methods - Support Vector Learning.*, Cambridge, Massachusetts: The MIT Press 1999 chap. Making large-Scale SVM Learning Practical.

[81] **SVMlight**[http://svmlight.joachims.org/].

[82] Chang CC, Lin CJ: **LIBSVM: a library for support vector machines** 2001. [Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm].

[83] **LIBSVM**[http://www.csie.ntu.edu.tw/~cjlin/libsvm/].

[84] Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**(Database issue):D61–5.

[85] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2007, **35**(Database issue):D21–5.

[86] Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35**(Database issue):D610–7.

[87] **The Ensembl website**[`http://www.ensembl.org/`].

[88] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh

T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291**(5507):1304–51.

[89] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock

GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860–921.

[90] **Ensembl species by October 2007**[`http://oct2007.archive.ensembl.org/sitemap.html`].

[91] Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SMJ, Stabenau A, Storey R, Clamp M: **The Ensembl analysis pipeline.** *Genome Res* 2004, **14**(5):934–41.

[92] Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res* 2004, **14**(5):942–50.

[93] Mathe C, Sagot MF, Schiex T, Rouze P: **Current methods of gene prediction, their strengths and weaknesses.** *Nucleic Acids Res* 2002, **30**(19):4103–17.

[94] Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG: **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biol* 2006, **7 Suppl 1**:S2.1–31.

[95] Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser.** *Genome Res* 2004, **14**(5):951–5.

[96] **The MySQL website**[`http://www.mysql.org/`].

[97] Durinck S, Moreau Y, Kasprzyk A, Davis S, Moor BD, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21**(16):3439–40.

[98] Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data.** *Genome Res* 2004, **14**:160–9.

[99] Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, **26**:320–2.

[100] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32**(Database issue):D138–41.

[101] Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**(Database issue):D247–51.

[102] Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C: **PRINTS and its automatic supplement, prePRINTS.** *Nucleic Acids Res* 2003, **31**:400–2.

[103] Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D: **The ProDom database of protein domain families: more emphasis on 3D.** *Nucleic Acids Res* 2005, **33**(Database issue):D212–5.

[104] Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Res* 2006, **34**(Database issue):D257–60.

[105] Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O: **TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes.** *Nucleic Acids Res* 2007, **35**(Database issue):D260–4.

[106] Wu CH, Nikolskaya A, Huang H, Yeh LSL, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvear J, Dinkov G, Barker WC: **PIRSF: family classification system at the Protein Information Resource.** *Nucleic Acids Res* 2004, **32**(Database issue):D112–4.

[107] Wilson D, Madera M, Vogel C, Chothia C, Gough J: **The SUPERFAMILY database in 2007: families and functions.** *Nucleic Acids Res* 2007, **35**(Database issue):D308–13.

[108] Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA: **Gene3D: modelling protein structure, function and evolution.** *Nucleic Acids Res* 2006, **34**(Database issue):D281–4.

[109] Mi H, Guo N, Kejariwal A, Thomas PD: **PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways.** *Nucleic Acids Res* 2007, **35**(Database issue):D247–52.

[110] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJA, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C: **New developments in the InterPro database.** *Nucleic Acids Res* 2007, **35**(Database issue):D224–8.

[111] Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**(9):847–8.

[112] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W116–20.

[113] Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments.** *Proteomics* 2004, **4**(7):1985–8.

[114] Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, Apweiler R: **The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro.** *Genome Res* 2003, **13**(4):662–72.

[115] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32**(Database issue):D262–6.

[116] Perneger TV: **What's wrong with Bonferroni adjustments.** *BMJ* 1998, **316**(7139):1236–8.

[117] Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100**(16):9440–5.

[118] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16**(5):412–24.

[119] Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405**(2):442–51.

[120] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L: **The use of receiver operating characteristic curves in biomedical informatics.** *J Biomed Inform* 2005, **38**(5):404–15.

[121] Fawcett T: **ROC Graphs: Notes and Practical Considerations for Researchers**. Tech. rep., HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto, CA 94304 2004.

[122] Jakobsson PJ, Morgenstern R, Mancini J, Ford-Hutchinson A, Persson B: **Common structural features of MAPEG – a widespread superfamily of membrane associated proteins with highly divergent functions in eicosanoid and glutathione metabolism.** *Protein Sci* 1999, **8**(3):689–92.

[123] Holm PJ, Morgenstern R, Hebert H: **The 3-D structure of microsomal glutathione transferase 1 at 6 A resolution as determined by electron crystallography of p22(1)2(1) crystals.** *Biochim Biophys Acta* 2002, **1594**(2):276–85.

[124] Kallberg Y, Persson B: **KIND-a non-redundant protein database.** *Bioinformatics* 1999, **15**(3):260–1.

[125] Vuilleumier S: **Bacterial glutathione S-transferases: what are they good for?** *J Bacteriol* 1997, **179**(5):1431–41.

[126] Holm PJ, Bhakat P, Jegerschold C, Gyobu N, Mitsuoka K, Fujiyoshi Y, Morgenstern R, Hebert H: **Structural basis for detoxification and oxidative stress protection in membranes.** *J Mol Biol* 2006, **360**(5):934–45.

[127] Ago H, Kanaoka Y, Irikura D, Lam BK, Shimamura T, Austen KF, Miyano M: **Crystal structure of a human membrane protein involved in cysteinyl leukotriene biosynthesis.** *Nature* 2007, **448**(7153):609–12.

[128] Ferguson AD, McKeever BM, Xu S, Wisniewski D, Miller DK, Yamin TT, Spencer RH, Chu L, Ujjainwalla F, Cunningham BR, Evans JF, Becker JW: **Crystal structure of inhibitor-bound human 5-lipoxygenase-activating protein.** *Science* 2007, **317**(5837):510–2.

[129] Plante H, Picard S, Mancini J, Borgeat P: **5-Lipoxygenase-activating protein homodimer in human neutrophils: evidence for a role in leukotriene biosynthesis.** *Biochem J* 2006, **393**(Pt 1):211–8.

[130] Reymond A, Meroni G, Fantozzi A, Merla G, Cairo S, Luzi L, Riganelli D, Zanaria E, Messali S, Cainarca S, Guffanti A, Minucci S, Pelicci PG, Ballabio A: **The tripartite motif family identifies cell compartments.** *EMBO J* 2001, **20**(9):2140–51.

[131] Meroni G, Diez-Roux G: **TRIM/RBCC, a novel class of 'single protein RING finger' E3 ubiquitin ligases.** *Bioessays* 2005, **27**(11):1147–57.

[132] Li X, Li Y, Stremlau M, Yuan W, Song B, Perron M, Sodroski J: **Functional replacement of the RING, B-box 2, and coiled-coil domains of tripartite motif 5alpha (TRIM5alpha) by heterologous TRIM domains.** *J Virol* 2006, **80**(13):6198–206.

[133] Hennig J, Ottosson L, Andresen C, Horvath L, Kuchroo VK, Broo K, Wahren-Herlenius M, Sunnerhagen M: **Structural organization and Zn2+-dependent subdomain interactions involving autoantigenic epitopes in the Ring-B-box-coiled-coil (RBCC) region of Ro52.** *J Biol Chem* 2005, **280**(39):33250–61.

[134] Ottosson L, Hennig J, Espinosa A, Brauner S, Wahren-Herlenius M, Sunnerhagen M: **Structural, functional and immunologic characterization of folded subdomains in the Ro52 protein targeted in Sjogren's syndrome.** *Mol Immunol* 2006, **43**(6):588–98.

[135] **The Gene Ontology website**[`http://geneontology.org`].

[136] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD: **The UMLS Metathesaurus: representing different views of biomedical concepts.** *Bull Med Libr Assoc* 1993, **81**(2):217–22.

[137] Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS: **Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla.** *Proteins* 2004, **54**:20–40.

[138] Tekaia F, Yeramian E: **Evolution of proteomes: fundamental signatures and global trends in amino acid compositions.** *BMC Genomics* 2006, **7**:307.

[139] Otaki JM, Ienaka S, Gotoh T, Yamamoto H: **Availability of short amino acid sequences in proteins.** *Protein Sci* 2005, **14**(3):617–25.

[140] Figureau A, Soto MA, Toha J: **A pentapeptide-based method for protein secondary structure prediction.** *Protein Eng* 2003, **16**(2):103–7.

[141] Yang AS, yong Wang L: **Local structure prediction with local structure-based sequence profiles.** *Bioinformatics* 2003, **19**(10):1267–74.

[142] Gamsjaeger R, Liew CK, Loughlin FE, Crossley M, Mackay JP: **Sticky fingers: zinc-fingers as protein-recognition motifs.** *Trends Biochem Sci* 2007, **32**(2):63–70.

[143] Larsson TP, Murray CG, Hill T, Fredriksson R, Schioth HB: **Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery.** *FEBS Lett* 2005, **579**(3):690–8.

[144] Ashurst JL, Chen CK, Gilbert JGR, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, Wilming L, Hubbard T: **The Vertebrate Genome Annotation (Vega) database.** *Nucleic Acids Res* 2005, **33**(Database issue):D459–65.

[145] Loveland J: **VEGA, the genome browser with a difference.** *Brief Bioinform* 2005, **6**(2):189–93.

[146] Benton MJ: **Finding the tree of life: matching phylogenetic trees to the fossil record through the 20th century.** *Proc Biol Sci* 2001, **268**(1481):2123–30.

[147] Roger AJ, Hug LA: **The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361**(1470):1039–54.

[148] Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The Comprehensive Microbial Resource.** *Nucleic Acids Res* 2001, **29**:123–5.

[149] Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, Gattiker A, Kulikova T, Faruque N, Duggan K, Mclaren P, Reimholz B, Duret L, Penel S, Reuter I, Apweiler R: **Integr8 and Genome Reviews: integrated views of complete genomes and proteomes.** *Nucleic Acids Res* 2005, **33**(Database issue):D297–302.

[150] **The Fossil Record**`[http://www.fossilrecord.net/]`.

[151] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2000, **28**:10–4.

[152] **NCBI taxonomy**
`[http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy]`.

[153] Hedlund J, Cantoni R, Baltscheffsky M, Baltscheffsky H, Persson B: **Analysis of ancient sequence motifs in the H-PPase family.** *FEBS J* 2006, **273**(22):5183–93.

[154] **GFF: General Feature Format** [http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml].