Computational Analyses of Biological Sequences - Applications to Antibody-based Proteomics and Gene Family Characterization

Mats Lindskog
Cover: Currently, the sequences of more than 300 different genomes are available. Computer-based methods are used for analyses and visualization of these biological sequences.
To my Mother and Father
Abstract

Following the completion of the human genome sequence, post-genomic efforts have shifted the focus towards the analysis of the encoded proteome. Several different systematic proteomics approaches have emerged, for instance, antibody-based proteomics initiatives, where antibodies are used to functionally explore the human proteome. One such effort is HPR (the Swedish Human Proteome Resource), where affinity-purified polyclonal antibodies are generated and subsequently used for protein expression and localization studies in normal and diseased tissues. The antibodies are directed towards protein fragments, PrESTs (Protein Epitope Signature Tags), which are selected based on criteria favourable in subsequent laboratory procedures.

This thesis describes the development of novel software (Bishop) to facilitate the selection of proper protein fragments, as well as ensuring a high-throughput processing of selected target proteins. The majority of proteins were successfully processed by this approach, however, the design strategy resulted in a number of fall-outs. These proteins comprised alternative splice variants, as well as proteins exhibiting high sequence similarities to other human proteins. Alternative strategies were developed for processing of these proteins. The strategy for handling of alternative splice variants included the development of additional software and was validated by comparing the immunohistochemical staining patterns obtained with antibodies generated towards the same target protein. Processing of high sequence similarity proteins was enabled by assembling human proteins into clusters according to their pairwise sequence identities. Each cluster was represented by a single PrEST located in the region of the highest sequence similarity among all cluster members, thereby representing the entire cluster. This strategy was validated by identification of all proteins within a cluster using antibodies directed to such cluster specific PrESTs using Western blot analysis. In addition, the PrEST design success rates for more than 4,000 genes were evaluated.

Several genomes other than human have been finished, currently more than 300 genomes are fully sequenced. Following the release of the tree model organism black cottonwood (Populus trichocarpa), a bioinformatic analysis identified unknown cellulose synthases (CesAs), and revealed a total of 18 CesA family members. These genes are thought to have arisen from several rounds of genome duplication. This number is significantly higher than previous studies performed in other plant genomes, which comprise only ten CesA family members in those genomes. Moreover, identification of corresponding orthologous ESTs belonging to the closely related hybrid aspen (P. tremula x tremuloides) for two pairs of CesAs suggest that they are actively transcribed. This indicates that a number of paralogs have preserved their functionalities following extensive genome duplication events in the tree’s evolutionary history.

Key words: antigen, antibodies, antibody-based proteomics, biocomputing, bioinformatics, Populus, PrEST, protein fragment, software, visualization
List of Publications
List of Publications

This thesis is based on the following publications, which in the text are referred to by their Roman numerals:


In addition, previously unpublished data is presented.

Related publications:


B Pettersson, E., Lindskog, M., Lundeberg, J. and Ahmadian, A. Tri-nucleotide Threading for Parallel Amplification of Minute Amounts of Genomic DNA, Submitted

†Publication I and IV are reproduced with kind permission of Biotechniques and Springer Science and Business Media.
# Table of Contents

**Abstract** | V  
---|---
**List of Publications** | VII  
**INTRODUCTION** | 1  
**1. The Human Genome** | 3  
  1.1 Towards a Complete Sequence of the Human Genome | 3  
  1.2 Genetic Evolution  
  
  1.2.1 Adaptive Evolution | 4  
  1.2.2 Duplication Events | 4  
  1.2.3 Other Evolutionary Factors | 5  
  1.2.4 Genetic Variation | 6  
  1.2.5 Defining Homology | 6  
  1.3 Genome Size, Structure and Content | 7  
  1.4 Genes | 8  
**2. The Proteome and Proteins** | 11  
  2.1 From Genome to Proteome | 11  
  2.2 Proteome Diversity | 12  
  2.3 Proteins | 14  
**3. Proteomics from a Bioinformatic Perspective** | 15  
  3.1 Background | 15  
  3.2 Protein Expression and Purification | 16  
  3.3 Two Dimensional Gel Electrophoresis | 17  
  3.4 Mass Spectrometry | 17  
  3.5 Structural Proteomics | 18  
  3.6 Antibodies in Proteomics  
  
  3.6.1 Antibodies and Other Affinity Ligands | 18  
  3.6.2 Antigens and Epitopes | 19  
  3.6.3 Protein Analysis on Microarrays | 20  
  3.6.4 Immunohistochemistry | 21  
  3.6.5 In Vitro Detection | 22  
  3.7 Prediction of Subcellular Localization | 23  
  3.8 Post Translational Modifications | 23  
  3.9 Protein Interactions and Complexes | 24  
**4. Biological information resources** | 25  
  4.1 Genomic Data | 26  
  4.2 Proteome and Protein Data  
  
  4.2.1 Protein Sequence Data and Protein Information | 26  
  4.2.2 Protein Structure Data | 27  
  4.2.3 Protein Domains and Families | 27  
  4.2.4 Protein Interactions | 28  
**5. Data Searches, Sequence Comparisons, and Biocomputing** | 29  
  5.1 Comparative Genomics and Proteomics | 29  
  5.2 Sequence Alignments  
  
  5.2.1 Pairwise Sequence Alignments | 30  
  5.2.2 Practical Aspects of Pairwise Alignments | 32  
  5.2.3 Multiple Sequence Alignments | 33  
  5.3 Prediction of Signal Sequences and Transmembrane Regions | 33  
  5.4 Biocomputing – a Practical Approach | 34
# Table of Contents

**PRESENT INVESTIGATION**

6. Objectives

7. Design of Protein Epitope Signature Tags (PrESTs) for Antibody-based Proteomics
   7.1 Background
   7.2 Design of Protein Fragments for Recombinant Protein Production (publication I)
   7.3 A Systematic Analysis of Designed PrESTs (publication II)
   7.4 Specific PrEST Design Strategies
      7.4.1 Antibody-based Proteomics on High Sequence Similarity Proteins (publication III)
      7.4.2 Design of Isoform Specific Protein Fragments (unpublished data)

8. Identification and Characterization of Gene Family Members in a Fully Sequenced Tree Genome
   8.1 *Populus*, a Tree Model System
   8.2 Cellulose Synthases
   8.3 Identification of Previously Unknown CesA Family Members (publication IV)

Concluding Remarks

List of Abbreviations

Acknowledgements

Populärvetenskaplig Sammanfattning (In Swedish)

References

APPENDICES

Publication I: Selection of Protein Epitopes for Antibody Production

Publication II: Design of Protein Epitope Signature Tags for Antibody-based Proteomics

Publication III: Processing of High Sequence Similarity Proteins in Antibody-based Proteomics

Publication IV: The Genome Sequence of Black Cottonwood (*Populus Trichocarpa*) Reveals 18 Conserved Cellulose Synthase (CesA) Genes
Introduction
Introduction
2

The Human Genome
1. The Human Genome

1.1 Towards a Complete Sequence of the Human Genome

Within each cell of all living organisms on earth lies their genetic blueprint, which is written in the form of DNA (DeoxyriboNucleic Acid). This blueprint is what gives each organism its unique characteristic features. Up until the late 1800s, little was known about the properties and attributes of DNA. Yet, in 1866 Gregor Mendel set us on the path to unravel its secrets (see figure 1.1). He discovered that many traits of an organism depend upon two distinct factors, one from the male and one from the female parent. These factors were later termed genes, carriers of the genetic information. In 1871 Friedrich Miescher successfully isolated nuclein (nuclein contains nucleic acids), which he proposed to be the hereditary material. This idea was refined in 1944 when Oswald Avery described DNA as the carrier of hereditary information. Nine years later James Watson and Francis Crick solved the now famous double stranded helical structure of DNA (Watson and Crick, 1953). Furthermore, Watson and Crick postulated the central dogma of molecular biology, which describes the process in which DNA is transcribed to mRNA (messenger RiboNucleic Acid), which is subsequently translated into protein (see chapter 2.1).

The advent (Maxam and Gilbert, 1977; Sanger et al., 1977), and automation (Smith et al., 1986) of DNA sequencing provided the foundation for deciphering the information in whole genomes. The Human Genome Project (HGP) was initiated in 1990, and its objectives were to sequence the human genome, together with four model organism genomes, by the year 2005. These species included a bacterium (Escherichia coli), a yeast (Saccharomyces cerevisiae), a nematode (Caenorhabditis elegans) and a fruitfly (Drosophila melanogaster). The Institute for Genome Research (TIGR) finished sequencing of the first bacterial genome (Haemophilus influenzae) in 1995 (Fleischmann et al., 1995), and in 1998, Celera Genomics introduced a private genome project initiative. A first glimpse of the human genome was provided in 1999 with the sequencing of chromosome 22 (Dunham et al., 1999).
A first draft of the human genome sequence was released in 2001, concurrently by the HGP (Lander et al., 2001) and Celera Genomics (Venter et al., 2001). However, this primary draft included a large number of sequence gaps. In the spring of 2003, 50 years after the DNA structure first was solved, a high-quality sequence map was released, covering 99% of the euchromatic (gene-rich) regions and with the majority of the previously identified gaps bridged (Stein, 2004). Finally, the largest internationally coordinated genomic research effort ever had fulfilled its goals by releasing a publicly available human genome sequence that will provide data for various avenues of research.

1.2 Genetic Evolution
The recent completions of genome sequences from a large number of different organisms (see chapter 5.1) have allowed the exploration of the evolution of genomic entities, such as genes and proteins. Genomes are not static, but dynamic, complex units that continuously change, adapt, and evolve, and are affected by several evolutionary factors.

1.2.1 Adaptive Evolution
"All living things have much in common, in their chemical composition, their germinal vesicles, their cellular structure, and their laws of growth reproduction. Therefore I should infer that probably all the organic beings which have ever lived on earth have descended from some one primordial form."
Charles Darwin, The Origin of Species by Means of Natural Selection.

The theory of natural selection stated by Charles Darwin was highly controversial in 1859. However, evidence supporting natural selection or adaptive evolution, has accumulated over time, and is now widely accepted by the scientific community. Adaptive evolution is the process by which genes (or organisms) adapt to their surrounding environment, resulting in evolutionary changes if the adaptations are benificial and functional. Genetic regions subjected to adaptive evolution are most likely functionally relevant. The identification of such regions might therefore result in predictions concerning their function (Swanson, 2003). For instance, adaptive evolution has promoted diversity in the antigen recognition site in the major histability complex locus (Hughes and Nei, 1988).

1.2.2 Duplication Events
Genetic duplications are crucial for organismal evolution, due to the large increase of genetic material following a duplication event (reviewed by Taylor and Raes, 2004). Moreover, it has been stated that two consecutive rounds of complete genome duplication have provided the genetic material necessary for the evolution of the vertebrate genome (Ohno, 1970; Sidow, 1996).

Duplication events might arise at different levels, duplications of entire genomes (polyploidization), of chromosomal segments (segmental duplications), of individual genes, or parts of genes. Nonduplicated genes must maintain their function and accordingly their evolution is generally constrained by selective pressure (Kent et al., 2003). In contrast, following a duplication event, one or both duplicates may experience relaxed functional constraints due to redundancy. This is in some cases followed by an increased evolutilonal rate, which can provide novel gene functions and regulations (Kondrashov et al., 2002). Genes may be altered by creating new function (neofunctionalization) and/or partitioning of the ancestral gene function (subfunctionalization) or become silent (nonfunctionalization or pseudogenization; see figure 1.2; Ohno, 1970; Force et al., 1999). However, the
majority of gene duplicates evolve towards silencing rather than preservation (Lynch and Conery, 2000) and silenced genes are generally termed pseudogenes (see chapter 1.4).

Rates of duplications varies among species and genes, shorter genes appear to be subjected to duplication events more frequently than longer genes (Taylor and Raes, 2004). According to this context, gene duplication is a relatively common occurrence (0.01 estimated duplications/gene/million years), and the average half-life of duplicated genes is relatively short (around 4.0 million years; Lynch and Conery, 2000; Gu et al., 2002; Lynch and Conery, 2003; Taylor and Raes, 2004).

Figure 1.2 A schematic view of the possible fates of duplicated genes. A: The classical model of duplicate gene fates, which consists of non- and neofunctionalization in coding regions (Ohno, 1970). B: Recent theories propose a more complex chain of events, which includes subfunctionalization as well as alterations in regulatory regions. The probabilities of the events are proportional to the thickness of the arrows (adapted from Moore and Purugganan, 2005).

1.2.3 Other Evolutionary Factors

Genome evolution is influenced by deletions and transpositions of genome segments or genes. In addition, retrotranspositions are of importance and comprise about 17% of the human genome (Lander et al., 2001). Retrotransposition is transposition by means of reverse transcription of RNA and subsequent insertion into the genome.

Horizontal gene transfer, where genes are transferred between species has also influenced genome evolution. For instance, about 18% of the genes present in E. coli appear to arise from horizontal gene transfer after the divergence from the salmonella lineage one billion years ago (Lawrence and Ochman, 1998).
1.2.4 Genetic Variation
Individual genomes within the same species differ as a result of chromosomal alterations and point mutations, such as single base pair substitutions, insertions and deletions. Variations that occur as single nucleotide bases as well as stretches of longer DNA segments, are invaluable markers in mapping of disease-related genes, in diagnostics and human population studies, as well as in evolutionary genetics (Ahmadian and Lundeberg, 2002).

The majority of genetic variations consists of single nucleotide polymorphisms (SNPs), which account for around 85% of all genomic variations (Little, 1999). A SNP is a genetic variation between individuals in a given population, which is limited to a single base pair and occurring at a frequency exceeding one percent (Ahmadian and Lundeberg, 2002). SNPs are distributed non-randomly throughout the human genome, occurring in average at every 1,250th base pair (Lander et al., 2001). Furthermore, SNPs located in the protein coding regions and regulatory segments of genes may give rise to phenotypic differences by altering protein structure and/or function, as well as protein expression levels, which subsequently might lead to disease. The majority of SNPs, however, are situated in noncoding regions of the human genome and are commonly used as convenient markers in population genetics and evolutionary studies (Syvanen, 2001).

Identification and amplification of relevant SNPs was facilitated by the introduction of the polymerase chain reaction (PCR; Mullis and Faloona, 1987; Saiki et al., 1989). For genome-wide studies, amplifications of SNPs need to be performed in parallel with large numbers of SNPs (Hardenbol et al., 2003; Fan et al., 2003; related publication B).

1.2.5 Defining Homology
Evolutionary relationships among genes might be complex and difficult to define, hence a set of descriptive terms have been coined (reviewed by Fitch, 2000). Homology is defined as the relationship of two characters (e.g. genes) sharing a common evolutionary origin (Fitch, 1970). Note that homology is not a measurement of similarity, but a strict statement describing a divergent relation between sequences.

There are a number of different subtypes of homology. Orthology is defined as the relationship of two homologous genes cognated by speciation (Fitch, 1970). Paralogy is the relationship of two genes arising from a duplication event (Fitch, 1970). However, paralogy falls short in specifying if the genes described are members of the same species. For instance, if duplication is subsequently followed by speciation, two genes belonging to different species can be defined as paralogs. As a consequence, two subtypes of paralogy have been proposed to bring clarity to this issue: paralogs within the same species and paralogs residing in different species; inparalogs and outparalogs, respectively (Sonnhammer and Koonin, 2002). Xenology, the third subtype of homology, is defined as the relationship of two homologous genes involving a horizontal (interspecial) gene transfer (see chapter 1.2.3) for at least one of the genes (see figure 1.3; Gray and Fitch, 1983). Conversely to homology, analogy is defined as the relationship of two genes that have descended from unrelated ancestors by convergence (Fitch, 1970). Unfortunately, these definitions of homology have been, and still are, inconsistently used throughout the research community (Fitch, 2000; Jensen, 2001; Koonin, 2001; Petsko, 2001).
1.3 Genome Size, Structure, and Content

Following the completion of the human genome (see chapter 1.1), in-depth sequence analyses have to take into consideration the specific properties of the genome. Genetic information is coded in DNA, which consists of four different building blocks, nucleotides, which are designated A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). DNA predominantly exists in a double stranded configuration, and has a self-complementary property. This property facilitates the recovery of the complementary strand, which is essential during cell-division when a cell needs to replicate its DNA (see figure 2.1). Moreover, DNA is tightly bundled in the nucleus by several levels of organization. At the highest level, human DNA is split into units called chromosomes. The haploid human genome consists of around three billion bases which are distributed among 24 chromosomes (see table 1.1).

Only about 1.5% of the genetic material constitutes protein coding regions, so called exons. The remaining part constitutes of about 48% of unique intergenic DNA, for which no function has yet been identified. The majority of this unique DNA is probably an evolutionary artifact, lacking present function. Furthermore, 49% of the human genome consists of repetitive elements (REs), and a small part, about 0.5%, comprise nonfunctional pseudogenes (see chapter 1.4; Makalowski, 2000). There are more than 4.3 million REs in the human genome, and many
translated REs are found in proteins, thus having importance for protein function (Li 2001). Although the specific function of these repetitive elements is still unknown, they do interact with the genome and influence its evolution (Makalowski 2000).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Size (Mbp)</th>
<th>% of genome</th>
<th>Number of genes</th>
<th>% genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>245.5</td>
<td>7.98</td>
<td>2,281</td>
<td>10.12</td>
</tr>
<tr>
<td>2</td>
<td>243.0</td>
<td>7.90</td>
<td>1,482</td>
<td>6.58</td>
</tr>
<tr>
<td>3</td>
<td>199.5</td>
<td>6.48</td>
<td>1,168</td>
<td>5.18</td>
</tr>
<tr>
<td>4</td>
<td>191.4</td>
<td>6.22</td>
<td>866</td>
<td>3.84</td>
</tr>
<tr>
<td>5</td>
<td>180.9</td>
<td>5.88</td>
<td>970</td>
<td>4.31</td>
</tr>
<tr>
<td>6</td>
<td>171.0</td>
<td>5.56</td>
<td>1,152</td>
<td>5.11</td>
</tr>
<tr>
<td>7</td>
<td>158.6</td>
<td>5.16</td>
<td>1,116</td>
<td>4.95</td>
</tr>
<tr>
<td>8</td>
<td>146.3</td>
<td>4.75</td>
<td>794</td>
<td>3.52</td>
</tr>
<tr>
<td>9</td>
<td>138.4</td>
<td>4.50</td>
<td>919</td>
<td>4.08</td>
</tr>
<tr>
<td>10</td>
<td>135.4</td>
<td>4.40</td>
<td>862</td>
<td>3.83</td>
</tr>
<tr>
<td>11</td>
<td>134.5</td>
<td>4.37</td>
<td>1,426</td>
<td>6.33</td>
</tr>
<tr>
<td>12</td>
<td>132.4</td>
<td>4.30</td>
<td>1,104</td>
<td>4.90</td>
</tr>
<tr>
<td>13</td>
<td>114.1</td>
<td>3.71</td>
<td>399</td>
<td>1.77</td>
</tr>
<tr>
<td>14</td>
<td>106.4</td>
<td>3.46</td>
<td>733</td>
<td>3.25</td>
</tr>
<tr>
<td>15</td>
<td>100.3</td>
<td>3.26</td>
<td>766</td>
<td>3.40</td>
</tr>
<tr>
<td>16</td>
<td>88.8</td>
<td>2.89</td>
<td>957</td>
<td>4.25</td>
</tr>
<tr>
<td>17</td>
<td>78.8</td>
<td>2.56</td>
<td>1,257</td>
<td>5.58</td>
</tr>
<tr>
<td>18</td>
<td>76.1</td>
<td>2.47</td>
<td>322</td>
<td>1.43</td>
</tr>
<tr>
<td>19</td>
<td>63.8</td>
<td>2.07</td>
<td>1,468</td>
<td>6.52</td>
</tr>
<tr>
<td>20</td>
<td>62.4</td>
<td>2.03</td>
<td>631</td>
<td>2.80</td>
</tr>
<tr>
<td>21</td>
<td>46.9</td>
<td>1.53</td>
<td>271</td>
<td>1.20</td>
</tr>
<tr>
<td>22</td>
<td>49.5</td>
<td>1.61</td>
<td>552</td>
<td>2.45</td>
</tr>
<tr>
<td>X</td>
<td>154.8</td>
<td>5.03</td>
<td>931</td>
<td>4.13</td>
</tr>
<tr>
<td>Y</td>
<td>57.7</td>
<td>1.88</td>
<td>104</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 1.1 Chromosomal data and gene content of the human genome (based on data from Ensembl v. 33.35d).

1.4 Genes
A gene is a stretch of DNA coding for a protein, and might be located on either DNA strand. In eukaryotic organisms, genes appear split into separated segments where regions of protein coding exons are interrupted by stretches of non-coding intervening sequences (introns). In humans, the exons have an average length of 150 nucleotides and the substantially longer introns are generally spanning around 3.5 kilo bases (kb). Yet, intron-lengths as large as 500 kb have been observed (Rowen et al., 2002). In a process entitled splicing, the introns are removed from the transcribed RNA and the protein coding elements are assembled into mature mRNAs which are subsequently translated into proteins. If alternative splicing occurs, the protein coding exons are united differently. This results in various protein isoforms with distinct functions, and is an important source of protein diversity (see chapter 2.2).

Through duplication of DNA segments (see chapter 1.2.2), a single gene can give rise to multiple copies of duplicated genes, which retain parts of their functionality. These genes might be referred to as a gene family (see chapter 8.2). The exact definition of a gene family differs, ranging from at least 35% sequence identity
between corresponding protein sequences (Orengo et al., 2002), to merely sharing a common protein fold. However, members in a gene family are clearly evolutionary related, often having similar or related functions.

Genes with close similarity to their respective paralogs, but unable to code for a functional protein, are termed pseudogenes (reviewed by Mighell et al., 2000). Pseudogenization is the most common fate for duplicated genes (see chapter 1.2.2). According to the theory of positive selection, one or several of duplicates are unconstrained by selective pressure, therefore randomly accumulating mutations. These mutations might subsequently silence gene function via disruptions of the original reading frame (e.g. frame shifts or introduction of a stop codon). Pseudogenes exist as two different types, processed and nonprocessed. Processed pseudogenes are formed through retrotransposition of mature RNAs, and these genes lack upstream promoters, as a result of random integration into the genome. Nonprocessed pseudogenes most commonly arise after a duplication event, followed by nonfunctionalization (see chapter 1.2.2). Pseudogenes are thought to be responsible for misincorporations in gene collections, and as much as 20% of predicted genes are thought to consist of nonfunctional pseudogenes (Lander et al., 2001; Mounsey et al., 2002). However, the total number of pseudogenes and their genomic locations are in many cases still unknown (Torrents et al., 2003).
2. The Proteome and Proteins

2.1 From Genome to Proteome

The majority of the genes in the human genome code for functional proteins, and the path from gene to protein includes many turns before the final gene product is reached. The process where DNA is transcribed into RNA and thereafter translated into protein, is termed the central dogma of molecular biology (see figure 2.1). Furthermore, the dogma includes DNA replication, in which DNA is copied during cell division. Following the transcription of DNA into mRNA, the mRNA migrates from the nucleus to the cytoplasm and encounters ribosomes. The ribosomes provide a framework for holding mRNA in position during protein synthesis, a process known as translation. After translation, proteins might be subjected to a wide variety of modifications which alter the protein properties by adding a modifying group to one or several amino acids, or by proteolytic cleavage (Mann and Jensen, 2003). There are hundreds of different post translational modifications (PTMs). For instance, cleavage of signal sequences, glycosylations, and phosphorylations (Garavelli, 2004). The PTMs might determine localization, activity, interaction and turnover of the proteins (Mann and Jensen, 2003). Moreover, mRNA might be subjected to pre-translational editing, resulting in amino acid changes not inferable from the DNA sequence.

Figure 2.1 The central dogma of molecular biology, which comprises transcription, translation, and replication. The dogma is hereby displayed in combination with the process of reverse transcription as well as with factors contributing to proteome diversity such as RNA editing and post-translational modifications. Exons are colored in green and introns are grey, however, introns are usually much longer than exons (see chapter 1.4). cDNA (complementary DNA) contains only exons and are therefore colored in green. mRNA might be subjected to pre-translational editing and these editions are here shown in orange. Moreover, proteins can be modified after the translation (e.g. phosphorylations) and these modifications are shown in red.
In addition, transcription can be reversed *in vitro* by using the enzyme reverse transcriptase, in a process where mRNA functions as a template for conversion into cDNA (complementary DNA). cDNA contains only exons and is often used in studies when coding gene regions are of interest.

### 2.2 Proteome Diversity

In 1994, Marc Wilkins first coined the term proteome as the set of proteins encoded by the genome (Wilkins et al., 1996). The proteome consists of a vast array of proteins, among which diversity is generated by events such as alternative splicing, somatic DNA rearrangements, and PTMs. At least one-third of the human genes are subjected to alternative splicing and there are several different types of alternative splicing events (see figure 2.2). Different selections of splice sites, such as 5' or 3' splicing, give rise to alternative isoforms. In addition, although rarely occurring, entire introns might be retained and transcribed (Garcia-Blanco et al., 2004). Alternative splicing often results in additional protein domains, rather than alternating domains (Mironov et al., 1999).

![Figure 2.2](image.png) 

**Figure 2.2** Different types of alternative splicing events. Constitutive exons are shown in light grey color, alternative exons are dark grey, and the intron is dark grey with stripes. Exons can have multiple 5'- or 3'-splice sites (A, B). Fully contained exons that are alternatively used are called cassette exons. Single cassette exons can reside between two constitutive exons and the alternative exon is either skipped or included (C). Multiple cassette exons can be located between two constitutive exons, and the splicing machinery must choose between them (D). Finally, a lack of splicing might result in intron retention (E; adapted from Graveley, 2001).
Including all protein variants, the human proteome is estimated to contain millions of different molecular species (O’Donovan et al., 2001; Bauer and Kuster, 2003; Uhlen and Ponten, 2005). However, the total number is clearly dependant on how the proteome is defined. The proteome might be divided into five sub-proteomic classifications, which all contribute to the total number of protein molecules (see table 2.1; Uhlen and Ponten, 2005). One class constitutes a nonredundant protein set, in which every gene locus contributes with a single representative protein. The current human gene count is 22,218 (Ensembl v33.35; Hubbard et al., 2005), and although this number will change, the final number in this class is most probably between 20,000 and 25,000. Another class is comprised of a large number of proteins, which result from alternative splicing events. An estimation of this category might give between 50,000 and 100,000 protein isoforms. Somatic rearrangements involved in the immune response give rise to a third class. For instance, the estimated number of different IgG molecule rearrangements might be as many as ten million. Many proteins undergo PTMs (see chapter 2.1), and this category might account for approximately 100,000 different protein molecules. However, this is a difficult number to estimate since many modifications are stepwise processes. Finally, there are is a myriad of protein variants among which differences are due to coding SNPs (see chapter 1.2.4).

<table>
<thead>
<tr>
<th>Protein class</th>
<th>Description</th>
<th>Number of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonredundant proteins</td>
<td>A protein representative from every gene locus</td>
<td>20,000-25,000</td>
</tr>
<tr>
<td>Protein isoforms</td>
<td>Variants obtained by alternative splicing</td>
<td>50,000-100,000</td>
</tr>
<tr>
<td>Combinatorial variants</td>
<td>Proteins generated by somatic DNA rearrangements</td>
<td>&gt;10,000,000</td>
</tr>
<tr>
<td>Protein species</td>
<td>Proteins that differ as a result of PTMs</td>
<td>&gt;100,000</td>
</tr>
<tr>
<td>Protein alleles</td>
<td>Differences in genetic variations resulting</td>
<td>75,000-150,000</td>
</tr>
</tbody>
</table>

Table 2.1 Estimated number of protein molecules in the human proteome, as a result of different classifications (adapted from Uhlen and Ponten, 2005).

This extensive diversity of the proteome might explain why species that are highly similar in their DNA differ substantially at the phenotypic level. Indeed, recent studies imply that human and chimpanzee genomes differ only by 1.2% in intergenic regions and 0.6% percent in protein coding regions (Ebersberger et al., 2002; Sakate et al., 2003). Since the proteome is not fully characterized, proteome comparisons might only be guessed.
2.3 Proteins

Proteins display a wide range of different functions (e.g. building blocks, enzymes, cellular, signal mediators, and receptors). They consist of 20 different building blocks, called amino acids, which differ in size, shape, hydrophobicity, and charge. Protein structure is of complex nature and contains four different levels of structural organization. The order in which the amino acids are linked together is the primary structure. The secondary structure of proteins refers to the elementary structural patterns present in most known proteins, where the two main structures are alpha helices and beta sheets. The tertiary structure is the complete three-dimensional folded structure of the protein chain (see figure 3.2). The final structural level, the quaternary structure, is only present if the protein contains more than one chain, and refers to the chain organization and interconnections. For example, hemoglobin contains a group of four amino acid chains.

Figure 2.3 The protein structure of the enzyme cellobiose dehydrogenase, hereby shown on a cellulose surface. Secondary structural elements are colored; alpha helices are red and beta sheets are blue.
3. Proteomics from a Bioinformatic Perspective

3.1 Background
Proteins are key functional and structural molecules. Therefore, the systematic and extensive molecular characterization of the proteome, proteomics, is necessary for understanding biological processes and systems (reviewed by de Hoog and Mann, 2004). Proteomics is a broad field which encompasses identification, quantification, and characterization of proteins expressed in various cells and tissues. Moreover, it includes methods for structural determination and characterization of PTMs, the use of antibodies as protein capture agents as well as bioinformatic approaches. A large-scale single initiative similar to the HGP (see chapter 1.1) is not possible due to the complex and dynamic nature of proteins. In proteomics, many different approaches in parallel are necessary (Lesley, 2001) and often used in combination to complement each other (see figure 3.1). As most drug targets are proteins, it is clear that proteomics will have a great impact on drug discovery and clinical diagnostics (Tyers and Mann, 2003).

Genomic and proteomic initiatives are producing an avalanche of data. Public databases (see chapter 4) are continuously expanding and sequence repositories grow at exponential rates (Benson et al., 2005). As a result of the enormous amounts of data, computer systems have become essential to researchers. Bioinformatics is frequently described as the organization and interpretation of biological data using computational systems and methods (reviewed by Yu et al., 2004). It is positioned at the interface of biology, computer science, statistics, and mathematics (Benton, 1996; Luscombe et al., 2001). The bioinformatics discipline aims at organizing and structuralizing biological research data in a human-readable format. Moreover, it focuses on development of software, tools, and resources facilitating data analysis, as well as the usage of the latter for visualization and interpretation of data in a relevant fashion. Bioinformatic methods and systems such as sequence analysis and comparative proteomics (see chapter 5), might provide valuable aid and provide insights to different areas of research. Proteomics approaches are currently highly dependent on computer-assisted systems and bioinformatic methods (see figure 3.1).
An overview of proteomic methods and their relations to each other. Moreover, a number of approaches can be complemented with corresponding *in silico* prediction methods. 'Antibody-based proteomics' includes methods such as ELISA (Enzyme-Linked ImmunoSorbent Assay), Western blot analysis (see chapter 3.7.2), and immunoprecipitation, whereas 'Structural proteomics' comprises NMR (Nuclear Magnetic Resonance) and x-ray crystallography. The dashed arrows from 'Mass spectrometry' and 'Two-dimensional gel electrophoresis' to 'In silico predictions' symbolize database screening against known data (see chapter 3.3).

### 3.2 Protein Expression and Purification

A challenge when analyzing proteins from a proteome-wide point of view is the expression and purification of human proteins in large numbers (Lesley, 2001). There are both prokaryotic and eukaryotic protein expression hosts, where the most frequently used bacterial host is *E. coli*. This bacterium has a rapid growth and generates high yields as well as having thoroughly characterized genomics. Furthermore, *E. coli* exists as a multitude of different mutant strains and is compatible with a large number of cloning vectors (Lu et al., 1996; Baneyx, 1999). However, the approach of expressing eukaryotic proteins in prokaryotic cell systems commonly encounters by problems in proper folding and satisfactory expression levels (Lesley, 2001). Therefore, some researchers use eukaryotic expression hosts such as the predominantly used *S. cerevisiae*, which is available in various different strains (Gellissen et al., 1992; Cregg et al., 2000).
Inclusion bodies are frequently formed in *E. coli*, and these aggregates can often be avoided by production of recombinant fusion proteins (Lu et al., 1996). Fusion tags are also used to facilitate purification and handling of recombinant proteins. For instance, the histidine-tag enables purification on immobilized metal ion affinity chromatography (IMAC) columns (Lesley, 2001). Moreover, fusion tags such as thioredoxin (Lu et al., 1996) provide a means to significantly raise expression levels.

A common procedure to analyse unknown human proteins begins with the identification of gene function by querying publicly available datasets (see chapter 4). This is followed by amplification of the gene from total RNA by reverse transcriptase polymerase chain reaction (RT-PCR) to generate cDNA (see chapter 2.1). Amplified gene products are subsequently cloned into a suitable vector (Liu, 1999; Hartley et al., 2000; Graslund et al., 2002). The cells expressing the proteins are subsequently cultivated, the proteins harvested and finally purified.

### 3.3 Two Dimensional Gel Electrophoresis

The widely used two-dimensional gel electrophoresis (2DE) technology was initially described 30 years ago (O’Farrell, 1975) and allows for separation of proteins in a poly-acrylamide gel according to isoelectric point and molecular weight. As many as 10,000 proteins can simultaneously be analyzed on a single gel (Jungblut et al., 1996). Following separation, the proteins are visualized by staining with either Coomassie blue, if the protein is in high abundance, or silver staining, which is more sensitive (Shevchenko et al., 1996). Separation and detection of PTMs of proteins is possible in 2D gel electrophoresis, however, 2DE fails to analyse certain types of proteins. For instance, proteins expressed in low copy number are not detected by the staining (Santoni et al., 2000). 2DE is frequently used with mass spectometry (MS) and bioinformatics and includes protein solubilization of tissue sample, followed by 2DE separation of the proteins, and computer analysis of protein spot patterns. Subsequently, the isolated protein spots are subjected to proteolytic digestion and characterization by MS, and the generated information is screened against databases for identification of the proteins.

### 3.4 Mass Spectrometry

MS is a method that has been further developed due to the recent availability of genome sequence databases (see chapter 4) and technical breakthroughs (Aebersold and Mann, 2003) and has become an important technique for protein identification. In MS, individual molecules are analyzed by ionization and measurement of their trajectory response to electric and/or magnetic fields. Analyses by MS are capable of revealing two specific protein features, the peptide-mass fingerprint and product-ion spectras. Peptide-mass fingerprints are determined by MS and product-ion spectras by tandem MS (MS/MS) or MS\(^n\). The two techniques are commonly used in combination, due to their capabilities of generating complementary information of the target protein (Beranova-Giorgianni, 2003). MS is used in conjunction with several other applications (see figure 3.1). For instance, MS and 2DE (see chapter 3.3) in combination with antibodies (see chapter 3.6.1) recognizing phosphorylated proteins may facilitate the identification and characterization of PTMs (Mann and Jensen, 2003).

The interpretation of large sets of acquired MS data is a cumbersome task and underlines the need for MS spectra analysis software. Several different analysis algorithms have been developed but need further refinements in order to finally supersede the need of human intervention (Blueggel et al., 2004).
3.5 Structural Proteomics

The functions of proteins are related to their structure and therefore three-dimensional (3D) protein structures are of great importance. The field of structural proteomics (also known as structural genomics) aims at mapping protein space, and determining 3D structures for all possible protein folds (Vitkup et al., 2001). Following sequencing of previously uncharacterized genes, 10-40% of corresponding protein sequences might be deduced a structure by comparisons to sequences of known structures (Rost et al., 2003). Protein structure is more evolutionary conserved than amino acid sequence (Blundell and Mizuguchi, 2000), and often provides more functional information than the sequence itself. This structural conservation is the result of retained protein interaction sites consisting of residues situated in close proximity in space, but might be distant in sequence location (Sanchez and Sali, 1998).

Structures for proteins that are easily expressed, crystallized and diffracted are solved by x-ray crystallographic methods (Terwilliger and Berendzen, 1999). Alternatively, proteins that fail to crystallize might be structurally determined by nuclear magnetic resonance (NMR; Montelione and Anderson, 1999).

Various computer-based methods for prediction of protein structure are in use (reviewed by Rost et al., 2003). Sequence-based large-scale structure predictions generally consist of sequence similarity searches against sequences with known 3D-structure followed by model building from detected structural templates, and finally a validation of the resulting models (Skolnick et al., 2000). Searching for proteins sharing structural similarities, but lacking obvious sequence similarities, are performed with fold recognition threading methods (Jones, 1999). Threading by fold recognition compares amino acid sequences to known 3D structures, and subsequently evaluates the level of fitting to these structures (Rost et al., 1997). Furthermore, fold recognition allows for identification of analogous (see chapter 1.2.5) pairs of proteins with no significant sequence similarities (Jones, 1999).

3.6 Antibodies in Proteomics

Antibodies have been used in protein studies for decades, owing to their strong affinity and high specificity in target binding. Currently, antibodies are used in several different applications such as immunolocalization studies (see chapter 3.6.4; reviewed by Uhlen and Ponten, 2005), in vitro detection (see chapter 3.6.5) and analyses on arrays (see chapter 3.6.3). Furthermore, antibodies have been modified by in vitro selection techniques (see chapter 3.6.1) to enhance their properties and increase their specificities and affinities.

3.6.1 Antibodies and Other Affinity Ligands

Advances in purification methods and selection techniques (e.g. phage display) have fueled the emergence of several different kinds of antibodies. As of today, four major types of antibodies are commonly used: polyclonal antibodies (pAbs), monoclonal antibodies (mAbs), mono-specific antibodies (msAbs), and other classes of affinity reagents.

pAbs derived from immunization of suitable animals (e.g. rabbits), were the primary antibody source until 30 years ago. The introduction of antigens (see chapter 3.6.2) into a host animal induces an immune response resulting in antigen-specific antibodies. This pool of antibodies binds to multiple epitope sites on the antigen. This multi-epitope binding contributes to a tolerance towards minor structural changes in the antigen, such as partial denaturation or PTMs
This makes pAbs suitable for multi-platform efforts wherein proteins often exist in different forms. However, pAbs are not renewable and may exhibit high cross-reactivities (Nilsson et al., 2005).

The introduction and development of the mAb technology (Kohler and Milstein, 1975), allowed the use of antibodies as reproducible chemical entities. This was made possible as a result of the development of immortal hybridoma cell lines continuously producing the selected mAbs (Borrebaeck, 2000). Currently, mAbs are the key immunoreagents for medical diagnostic applications (Borrebaeck, 2000). However, mAbs have shortcomings in large-scale efforts, as a result of the need for time-consuming screening (Uhlen and Ponten, 2005). In contrast to pAbs, mAbs recognize single epitopes and are therefore less versatile for multi-platform functional and analytical applications which often involve proteins in denatured forms. For instance, proteins are often exposed to formalin in tissue fixation for immunolocalization studies (see chapter 3.6.4).

msAbs are enriched from polyclonal antisera by high stringency affinity-purification. Therefore mAbs have the same advantages of pAbs, but display low cross-reactive tendencies as a result of the purification (Uhlen and Ponten, 2005). Preliminary studies also indicate that reimmunizations of the target antigen in combination with the stringent immunoaffinity purification yield antibodies with very similar binding characteristics (related publication A). The msAbs might therefore be considered as a renewable resource, and they have been used in studies involving extensive tissue profiling (Agaton et al., 2003; Kampf et al., 2004).

More recently, techniques have been developed that allow antibodies and other affinity reagents to be developed without the involvement of laboratory animals. These affinity ligands and antibodies are developed with in vitro selection techniques, such as phage display (Bradbury and Marks, 2004) or ribosomal display (Lipovsek and Pluckthun, 2004). They include antibody fragments, such as scFvs (single chain fragment) and Fabs (Better et al., 1988; Liu and Marks, 2000) and other protein-based reagents developed by recombinant protein techniques (reviewed by Hey et al., 2005) such as affibodies (Nord et al., 1997).

### 3.6.2 Antigens and Epitopes

The generation of antibodies requires selection of target antigens and three different types of protein antigens are generally produced: full-length proteins, synthetically produced peptides, and recombinant protein fragments. Full-length protein antigens are used for correct folding of the native protein, or when studying PTMs (Miroux and Walker, 1996). The use of a correctly folded full-length protein might however be questionable in some applications. For instance, if the protein is immunized into animals, it is most likely slightly denaturated upon mixing with different adjuvants (Uhlen and Ponten, 2005). Synthetically generated peptides are the shortest of the antigen-types, usually spanning less than 40 residues, and therefore the peptides are unlikely to be correctly folded. However, antibodies towards synthetic peptides have successfully been generated (Meloen et al., 2003). Recombinant protein fragments have the great advantage of being easily scaled-up for whole-proteomics studies (Agaton et al., 2003). These protein fragments are shorter than full-length proteins, thereby facilitating the cloning by usage of relatively short PCR fragments and the fusion to affinity-tags (see chapter 3.2) allows for high-throughput purification.
Antibodies recognize different antigenic determinants, epitopes, which are either linear or discontinuous. Linear epitopes consist of a continuous stretch of amino acids, whereas an epitope with residues in close proximity in space, but distantly located in sequence, is discontinuous (Blythe and Flower, 2005). The majority of epitopes are discontinuous (Barlow et al., 1986) however, most in silico methods focus on predicting linear epitopes. The majority of algorithms are using amino acid based scales to predict epitopes, e.g. to find local hydrophilicity points (Hopp and Woods, 1981). In conclusion, the most recent and comprehensive study shows no significant correlation between predicted and previously experimentally mapped epitopes. Furthermore, current epitope predictions perform only slightly better than by random chance (Blythe and Flower, 2005).

3.6.3 Protein Analysis on Microarrays
Following the footsteps of DNA microarrays (Schena et al., 1995), the development of protein-based array technologies has increased at a rapid rate. In contrast to traditional experiments where antibodies are used in very limited numbers, simultaneous analyses of large numbers of protein samples are now possible by using antibodies as binding agents (Bradbury et al., 2004; Nilsson et al., 2005). Protein arrays are highly specific and experiments have successfully identified a single protein among 10,000 others (MacBeath and Schreiber, 2000). Microarrays consisting of protein-based capture ligands as probes, or proteins themselves being fixed to the surface of a silicon chip, are valuable protein analysis tools.

There are three standard types of protein-detecting arrays (see figure 3.2). Antibody arrays where antibodies are fixed to a solid support and subsequently incubated with pre-labelled antigens (Haab, 2001; Haab et al., 2001; Schweitzer and Kingsmore, 2002), prior to the detection of antibody-antigen bindings. On sandwich arrays antibodies are fixed on the slide and captured protein targets are subsequently detected using an additional labeled antibody (Schweitzer and Kingsmore, 2002). Finally, on reverse phase protein arrays, proteins or protein mixtures are immobilized and detected upon binding to labeled antibodies. These experiments might be performed with primary or secondary labeled antibodies (Paweletz et al., 2001). It is important to keep in mind that the labeling of proteins might result in chemical modifications, including epitope damage and affinity changes (Barry and Soloviev, 2004). A major hurdle is difficulties in detecting all PTMs and isoforms of a protein (Lopez and Pluskal, 2003).
3.6.4 Immunohistochemistry

Immunohistochemistry (IHC) is the process of using antibodies for detection of antigens in tissues, and might provide information of protein expression in tissues at the subcellular level (reviewed by Warford et al., 2004). Determination of the subcellular localization of a protein is usually performed by fusion of the corresponding gene to a reporter or by epitope-tagging (Kumar et al., 2002), where small tags are fused to the target protein.

Prior to analyses on tissue samples, the tissues are fixed, and thereafter embedded in paraffin or alternatively flash frozen. The availability of frozen tissues is limited and therefore not as widely used as paraffin-embedded tissues. Tissue-fixation is most commonly performed directly following collection by submersion in formalin and thereafter insertion into paraffin blocks. Alternatively, the fixation of tissues is performed in ethanol (Ahram et al., 2003). A number of methods for simultaneous analysis of multiple tissue samples have been implemented. The primary step towards high-throughput histology methods was the tissue sausage method (Battifora, 1986), which allowed for the analysis of over 100 different tumour tissues. However, the tissues were not oriented to allow linkage to their clinical origin and the tissue sausage failed to represent the entire number of tissues in a section (Warford et al., 2004). The tissue microarray (TMA; see figure 3.3) technology surmounted these limitations and dramatically increased the number of simultaneously analysed tissue samples (Kononen et al., 1998). Following incubation with antibodies and staining, the capture of TMAs is performed by conventional flatbed scanners. Subsequent storage of high-resolution images requires large amount of storage space.

Figure 3.2 Different detection methods on protein/antibody arrays: direct antibody arrays, sandwich arrays, and reverse phase protein arrays. In addition, reversed phase arrays might be performed with a secondary labeled antibody, as well as with complex protein samples hybridized to the chip surface.
When using antibodies in TMA experiments, there is a risk for cross-reactivities. Antibody recognition of related or unrelated epitopes in tissues, where the target protein is nonexisting, may result in false positives (Warford et al., 2004). On the contrary, false negatives might arise as a result of difficulties of the antibody to identify the target, destruction of the epitope as a result of tissue-fixation, or because the epitope might be unreachable due to cross-linking or protein-protein interactions. Moreover, the detection system might fail to identify the target epitope due to expression levels below detection levels (Warford et al., 2004). Also, special care should be taken to select tissue cores being representative of the pathology or histology of standard sections (Simon et al., 2003).

Figure 3.3 A view of a TMA (Tissue MicroArray) containing immunohistochemically stained sections of normal human tissues. Each spot has a diameter of 1 mm and consists of a different tissue sample. In this example, the protein expression pattern is displayed in a spot representing part of a normal appendix. The higher magnification view shows that glandular cells in the mucosa are positive. Surface epithelial cells show only weak immunoreactivity, whereas the glandular cells lining the colonic crypts show strong cytoplasmic positivity. The inflammatory cells surrounding the crypts are essentially negative. Brown color represents positive staining. The blue color (hematoxylin) is used as non-specific counter-staining to expose tissue structure and cell morphology.

3.6.5 In Vitro Detection
Tissue localization is also possible to perform using Western blot analysis, where tissue lysates are separated by gel electrophoresis. Western blot is a commonly employed technique for detection of protein antigens in complex mixtures (Dunn, 1999). Following electrophoresis, the proteins (or other biological samples) are transferred to a membrane, usually composed of polyvinylidene difluoride (PVDF) or nitrocellulose, for subsequent detection. Prevention of non-specific binding of antibodies to the surface is performed by incubation of the membrane in appropriate blocking solution. The detection in Western blot analysis is commonly performed by with direct or indirect methods. In the direct method, the primary antibody, which binds the antigen, is labeled by an enzyme or fluorescent dye. The indirect method involves a secondary labeled antibody, which binds to the primary antibody bound to the antigen. Labeling of the secondary antibody can be done by the use of various compounds, including probes, biotin and enzyme conjugates. Western blots also give valuable information about protein size and antibody specificity.
3.7 Prediction of Subcellular Localization

The majority of eukaryotic genes are transcribed in the nucleus, and subsequent translation occurs in the cytosol. Many proteins undergo further sorting into various subcellular components by signals within the amino acid sequence (Schatz and Dobberstein, 1996; Mattaj and Englmeier, 1998). Prediction of these signal sequences is commonly achieved by different methods. By establishing homology to a protein with known localization, the destiny of unknown proteins might be solved. Proteins may also be directed to different cellular compartments by the aid of carrier proteins. These proteins identify specific sequence motifs, which target the protein to a certain part of the cell. \textit{In silico} based methods can successfully predict signal peptide cleavage sites with greater than 80\% accuracy (see chapter 5.3; Emanuelsson et al., 2001). However, methods predicting N-terminal signals are biased as a result of errors in predicting start codons in genome projects (Reinhardt and Hubbard, 1998). \textit{Ab initio} methods predict the subcellular localization from the amino acid composition alone, but these methods are less successive compared to the sequence motif methods (Rost et al., 2003). Finally, localization can be determined by identifying interactions with other proteins that have known subcellular localization.

3.8 Post Translational Modifications

Analysis of individual PTMs (see chapter 2.1) is most commonly performed by comparing experimentally generated data to previously identified amino acid sequences. The primary step is to identify the target protein, which can be done by MS techniques or antibody recognition using Western blot analysis (Mann and Jensen, 2003).

Modifications of many proteins are essentially determined by four different approaches, 2DE, affinity-based enrichment of modified proteins, PTM identification in complex mixtures, and affinity-based methods and derivatization. In 2DE (see chapter 3.3), the modified proteins can be visualized on gels, by for instance, anti-phosphoamino acid antibodies, and subsequently analyzed in MS (Soskic et al., 1999). Alternatively, proteins with PTMs can be run twice on 2D gels with an enzymatic removal of the modifying group between runs. Spots only present in the first run indicate modifications of the proteins (Yamagata et al., 2002). When the mass of the modified protein is not sufficient to establish the modification, MS/MS (see chapter 3.4) fragmentation is used to identify and localize the PTMs (Mann and Jensen, 2003).

Bioinformatic methods to predict PTMs are promising and might be closely integrated with experimental approaches. As proteomics approaches identify larger numbers of modification-sites \textit{in vivo}, prediction algorithms will benefit and their performance will increase (Mann and Jensen, 2003). \textit{In silico} prediction of PTMs comprises the identification of local consensus sequence motifs of signal sequence cleavage sites (Nielsen et al., 1997) and more complicated amino acid correlation patterns such as secondary structure and surface accessibility (Hansen et al., 1998; Blom et al., 1999).
3.9 Protein Interactions and Complexes

Proteins can interact with a myriad of other molecules to regulate and support them in large interaction networks (Chen and Xu, 2003; Bader and Hogue, 2000). Deciphering protein interactions might shed light on molecular mechanisms and biological processes (Ito et al., 2000) and provide clues to the function of uncharacterized proteins. Multi-protein complexes are regulating various cellular processes, of which some are associated with disease. Screening of these protein interaction partners might lead to the identification of possible drug targets. Since the development of the two-hybrid assay 15 years ago (Fields and Song, 1989), this powerful screening method has been extensively used in interaction studies. In this assay, the ‘bait’ and ‘prey’ proteins are fused to DNA binding- and transcriptional factor activation domains, respectively. Both fused proteins are subsequently co-expressed in yeast, and if proper interaction occurs, the two parts are brought in close proximity resulting in the expression of a reporter molecule used for detection (Uetz et al., 2000).

*In silico* predicted protein-protein interactions are not relevant in the absence of experimental validation. However, the two different approaches might be used in combination, complementing and validating each other (Chen and Xu, 2003). There are several interaction prediction methods, such as genetic neighborhood conservation, gene fusion and phylogenetic profiling.

Prokaryotic gene clusters and operons (contiguous sets of genes that are co-regulated and co-transcribed) provide the basis for the genetic neighborhood conservation method. Prokaryotic gene clusters typically consist of genes related by function (Overbeek et al., 1999). Valuable information on which proteins form complexes can be gained by identifying and comparing operons and gene clusters from different organisms. The major drawback is that this method only applies to bacteria.

A method called the gene fusion method predicts protein-protein interactions on the basis of homologous fused proteins. For example, protein A and B are separated proteins in organism Y, and expressed as a fused protein in organism Z. The fusion (and therefore probable interaction in organism Z) implies that protein A and B also interact in organism Y (Marcotte et al., 1999). One limitation of this method is that many interactions have evolved through other mechanisms, and therefore might be missed.

The phylogenetic profiling method is based on the assumption that interacting proteins functioning together also have a similar evolution. By constructing a phylogenetic profile for each protein that constitutes the proteins evolutionary history, linkage between proteins with similar profiles might be established. The method thereby allows for prediction of functions of unknown proteins (Pellegrini et al., 1999).
4. Biological Information Resources

Databases of biological information are an important resource behind biotechnology research (see table 4.1). Proteomic approaches depend on complete, reliable and recently updated databases (Apweiler et al., 2004a). Some researchers have shifted their focus from small-scale studies towards analyses involving large data sets. This includes comparative studies between sets of sequences or between organisms. The enormous quantities of gene and protein information need to be stored in a relevant way, allowing for integration and cross-references between different data sources. Besides functioning as storage facilities, the data sources have to provide access, interpretation and visualization of the data. The number of databases is steadily increasing and currently there are 719 publicly available molecular biology databases; an increase with 31% from previous year (Galperin, 2005).

<table>
<thead>
<tr>
<th>Database Name</th>
<th>URL</th>
<th>Database Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl</td>
<td><a href="http://www.ensembl.org">http://www.ensembl.org</a></td>
<td>Information around large automatically annotated genomes</td>
</tr>
<tr>
<td>The Vertebrate genome annotation (VEGA)</td>
<td><a href="http://vega.sanger.ac.uk/">http://vega.sanger.ac.uk/</a></td>
<td>Manually annotated vertabrate genome sequences</td>
</tr>
<tr>
<td>Swissprot</td>
<td><a href="http://www.ebi.ac.uk/swissprot">http://www.ebi.ac.uk/swissprot</a></td>
<td>Extensively annotated nonredundant protein sequences</td>
</tr>
<tr>
<td>TrEMBL</td>
<td><a href="http://www.ebi.ac.uk/trembl">http://www.ebi.ac.uk/trembl</a></td>
<td>Computer-annotated protein sequences</td>
</tr>
<tr>
<td>Uniprot</td>
<td><a href="http://www.expasy.uniprot.org/">http://www.expasy.uniprot.org/</a></td>
<td>Centralized repository for protein information</td>
</tr>
<tr>
<td>Protein data bank</td>
<td><a href="http://www.rcsb.org/pdb/">http://www.rcsb.org/pdb/</a></td>
<td>Three dimensional protein structures</td>
</tr>
<tr>
<td>Pfam</td>
<td><a href="http://www.sanger.ac.uk/Software/Pfam/">http://www.sanger.ac.uk/Software/Pfam/</a></td>
<td>Protein families and domains</td>
</tr>
<tr>
<td>Interpro</td>
<td><a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a></td>
<td>Unites domains, protein families, and from protein signature databases</td>
</tr>
<tr>
<td>Database of interacting proteins (DIP)</td>
<td><a href="http://dip.doe-mbi.ucla.edu/">http://dip.doe-mbi.ucla.edu/</a></td>
<td>Experimentally and literature search</td>
</tr>
<tr>
<td>Biomolecular interaction network database (BIND)</td>
<td><a href="http://binddb.org">http://binddb.org</a></td>
<td>generated protein-protein interaction data</td>
</tr>
<tr>
<td>The Predictome database</td>
<td><a href="http://predictome.bu.edu/">http://predictome.bu.edu/</a></td>
<td>Protein-protein, protein-DNA, protein-RNA, and protein-ligand interaction data</td>
</tr>
</tbody>
</table>

Table 4.1 A summary of a number of databases for biological information.

The need of integrating different biological information resources, in combination with the severe problem of defining protein function, emphasizes the importance of a standardized vocabulary. The first attempt was an enzymatic activity classification using four digits; EC numbers (Nomenclature committee, 1992). Every EC number is associated with a name for the respective enzyme, and the numbers represent a progressively finer classification of the enzyme, in a hierarchic fashion.

Currently, the gene ontology (GO) terms are the major set of terms used for describing protein function used. The GO consortium has defined a set of structured classifications in three levels: i) molecular function (e.g. enzyme), ii) biological process (e.g. cell growth and maintenance), and iii) cellular component (e.g. nuclear membrane; G.O. Consortium, 2001). The GO terminology is, however, not entirely complete, although the most comprehensive set of definitions currently available.
4.1 Genomic Data

There are three main systems that provide organized information around large automatically annotated genomes: Ensembl (Birney et al., 2004), the NCBI (National Center for Biotechnology Information) genome resource (Wheeler et al., 2004) and the USCS (University of California Santa Cruz) genome browser (Karolchik et al., 2004). All three resources are interlinked and updated regularly.

Ensembl provides several tools for users who need to manipulate parts of the genomic data. For instance, the EnsMart system (Kasprzyk et al., 2004), which is a web-based data mining and retrieval system, allows for extraction of relevant data in several output formats. Moreover, a series of downloadable standard data sets are available (i.e. a protein FASTA file of all proteins). Over the last year, a total of seven new genomes have been added to Ensembl, which is presently comprised of a total of 16 different genomes such as human, mouse, rat, dog, and chicken (see table 5.1).

The vertebrate genome annotation (VEGA) database is a resource for browsing manually annotated finished vertebrate genome sequences. It currently includes human, mouse, dog, and zebrafish. The entire set of annotated genes are supported by transcriptional evidence from protein sequences, cDNA, or expressed sequence tags (ESTs; Ashurst et al., 2005). At present, 14 of the human chromosomes, covering almost half of the genome, are annotated. Manual annotation is more precise in recognizing splice variants and pseudogenes as compared to automated annotation, as in Ensembl (Ashurst et al., 2005).

4.2 Proteome and Protein Data

4.2.1 Protein Sequence Data and Protein Information

SWISS-PROT (Bairoch and Apweiler, 2000) is maintained by the Swiss institute of bioinformatics (SIB) and the European bioinformatics institute (EBI). It provides annotated data with minimum redundancy, and is integrated with other databases. The entries in SWISS-PROT are extensively annotated and analysed by biologists to ensure high quality. There are two classes of data in SWISS-PROT, the core data and the annotation. The mandatory core information associated with each SWISS-PROT entry consists of the protein name, the amino acid sequence, citation information, and taxonomic data. The annotation includes description regarding protein function, PTMs, secondary structure elements, similarities to other proteins, domains and sites, splice isoforms, etc. Redundancies are kept at a minimum by merging together separate entries belonging to different literature reports. Furthermore, SWISS-PROT provides cross-references to many external databases (O’Donovan et al., 2002).

TrEMBL is comprised of computer-annotated sequences derived from translation of all coding sequences in GenBank (Benson et al., 2005), the DNA databank of Japan (DDBJ; Gasteiger et al., 2001), and European molecular biology laboratory (EMBL; Kanz et al., 2005) nucleotide sequence databases. The data include large amounts of information which is not yet included in the SWISS-PROT dataset. The TrEMBL data generation is performed in three steps: i) translation of the coding sequences previously mentioned, ii) removal of redundancy by merging of multiple entries (O’Donovan et al., 1999), and iii) automated information enhancement through annotation transfer from well-annotated proteins in SWISS-PROT to uncharacterized proteins in TrEMBL, which belong to well-defined groups, by using InterPro (Mulder et al., 2005).
The Universal protein resource (UniProt; Apweiler et al., 2004b) has been formed by uniting SWISS-PROT, TrEMBL, and protein information resource (PIR). It is a centralized repository for protein information and UniProt consists of three database components. The UniProt archive (UniParc), the most comprehensive publicly available protein sequence collection provides nonredundant protein sequences from all publicly accessible protein data bases. UNIPROT is the UniProt knowledgebase, which contains a section of fully manually annotated sequences and a section awaiting annotation. Finally, the UniProt NREF (UniRef) database provides nonredundant sequence collections, grouped by sequence identity. Three different clustering limits (NREF100, NREF90, NREF50) provide a full sequence coverage, while showing only sequences corresponding to the specific limitations (Bairoch et al., 2005).

4.2.2 Protein Structure Data
PDB (the Protein Data Bank; Bernstein et al., 1977) was initiated as early as 1971 and is the single repository for three dimensional protein and nucleic acid structural data. PDB archives, annotates, and distributes sets of atomic coordinates. In addition to the structural data itself, PDB also stores the measurements from which the data is derived (i.e. x-ray structure determinations). Protein structure (see chapter 3.5) is associated with the functions of proteins and therefore of importance.

4.2.3 Protein Domains and Families
Pfam (Protein families database) is a large repository of protein families and domains, including 7,973 families (version 18). Pfam families are built around multiple sequence alignments (see chapter 5.2.3) and are divided into two main classes, Pfam-A and -B. Pfam-A families are based on manually curated multiple alignments, whereas Pfam-B are derived from an automatic clustering of SWISS-PROT and TrEMBL and are therefore less reliable. Pfam families cover around 75% of all proteins present in SWISS-PROT and TrEMBL (Bateman et al., 2004), and have been widely used when annotating the human genome sequence (Lander et al., 2001).

Another resource is InterPro, which unites domains, protein families, and functional site records from protein signature databases. Each InterPro entry includes annotation, functional annotation in GO terms and literature references. At present, eight major databases are included, e.g. Pfam (Bateman et al., 2004), SMART (Simple Modular Architecture Research Tool; Letunic et al., 2004), and PRINTS (Protein fingerprint database; Attwood et al., 2003). InterPro’s family, domain, and functional site definitions will primarily be used in functional classification and annotation of unknown sequences (Apweiler et al., 2001).
**4.2.4 Protein Interactions**

The Database of interacting proteins (DIP) contains experimentally generated data and results from literature searches, regarding protein-protein interactions (Marcotte et al., 2001). It mostly includes data from human, yeast, and *Helicobacter pylori*, and allows for comprehensive visualization and extraction of interaction and network data (Xenarios et al., 2000).

BIND (the Biomolecular Interaction Network Database) stores protein-protein, protein-DNA, protein-RNA, and protein-ligand interaction data (Bader and Hogue, 2000). In addition, it includes information on molecular complexes and pathways. BIND describes various conditions and data regarding the interaction of interest such as experimental conditions, cellular localization, sequence data, kinetics, thermodynamics, etc.

*In silico* predicted interactions are stored at the Predictome database. The putative interactions are based on predictions using phylogenetic profiling, domain fusion, and chromosomal proximity (see chapter 3.9), involving proteins from 44 different species. This database also provides comparison of different prediction methods and their correlation with known pathway data (Mellor et al., 2002).
5. Data Searches, Sequence Comparisons, and Biocomputing

5.1 Comparative Genomics and Proteomics
Comparative genomics and proteomics are the analyses and comparisons of genomes and proteomes from different organisms in order to investigate the evolution of species and functions of genes and proteins. Bioinformatic methods may be applied to all organisms, as a result of the uniform genetic code. Due to the availability of completely sequenced genomes from various organisms (see table 5.1), comparative genomics and proteomics have rapidly emerged. Currently, a total of 303 (www.genomesonline.org) different genomes are finished. By examining features, such as genetic locations and conserved regions, correlations between genes can be established, providing valuable information. In closely related species (e.g. human and chimpanzee) differences might result in candidate genes for adaptive evolution (see chapter 1.2.1). Conversely, in distantly related species (e.g. human and mouse) the basic procedure involves screening for functional conservation (Ruvolo, 2004). Comparative genomics often involve the use of pairwise or multiple sequence alignments (see chapter 5.2.1, 5.2.3).

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of genes</th>
<th>Genome size (Mb)</th>
<th>Original reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (Homo sapiens)</td>
<td>22,218†</td>
<td>3,272</td>
<td>Lander et al., 2001, Venter et al., 2001</td>
</tr>
<tr>
<td>Chimpanzee (Pan troglodytes)</td>
<td>22,475†</td>
<td>2,733</td>
<td>The Chimpanzee sequencing consortium, 2005</td>
</tr>
<tr>
<td>Dog (Canis familiaris)</td>
<td>18,201†</td>
<td>2,385</td>
<td>Kirkness et al., 2003</td>
</tr>
<tr>
<td>Chicken (Gallus gallus)</td>
<td>17,709†</td>
<td>1,054</td>
<td>Wallis et al., 2004</td>
</tr>
<tr>
<td>Mouse (Mus musculus)</td>
<td>25,613†</td>
<td>2,267</td>
<td>Waterston et al., 2002</td>
</tr>
<tr>
<td>Rat (Rattus norvegicus)</td>
<td>21,952†</td>
<td>2,718</td>
<td>Gibbs et al., 2002</td>
</tr>
<tr>
<td>Fish (Danio rerio)</td>
<td>22,877†</td>
<td>1,688</td>
<td><a href="http://www.sanger.ac.uk/projects/d_reio/">http://www.sanger.ac.uk/projects/d_reio/</a> ¶</td>
</tr>
<tr>
<td>Fly (Drosophila melanogaster)</td>
<td>14,359†</td>
<td>132</td>
<td>Adams et al., 2000</td>
</tr>
<tr>
<td>Worm (Caenorhabditis elegans)</td>
<td>19,745†</td>
<td>100</td>
<td>The C. elegans sequencing consortium, 1998</td>
</tr>
<tr>
<td>Protozoan (Trypanosoma cruzi)</td>
<td>22,570†</td>
<td>60</td>
<td>El-sayed et al., 2005</td>
</tr>
<tr>
<td>Yeast (Saccharomyces cerevisiae)</td>
<td>6,680</td>
<td>12</td>
<td>Goffeau et al., 1996</td>
</tr>
<tr>
<td>Tree (Populus trichocarpa)</td>
<td>58,036†</td>
<td>473</td>
<td><a href="http://genome.jgi-psf.org/poptr1">http://genome.jgi-psf.org/poptr1</a> ¶</td>
</tr>
<tr>
<td>Plant (Arabidopsis thaliana)</td>
<td>31,270†</td>
<td>115</td>
<td>The Arabidopsis genome initiative, 2000</td>
</tr>
</tbody>
</table>

Table 5.1 The complete sequenced genomes of 12 different organisms. †Number of genes according to Ensembl, October 2005. ¶ No reference yet available.

5.2 Sequence Alignments
Initial clues to function and/or structure of newly sequenced proteins are commonly derived from amino acid sequence similarities to known proteins. As a consequence, the focus of research might be directed towards unexpected relations and correlations (Lipman and Pearson, 1985; Altschul and Lipman, 1990). Furthermore, discoveries of new protein families have been made by using sequence searches alone (Pearson, 1990). Homologous (see chapter 1.2.5) proteins have their fold and/or active site of binding domains in common, as well as sharing significant sequence similarities.

Alignments, unambiguous mapping of residues between sequences, are mainly performed to establish inference of homology between two (see chapter 5.2.1) or multiple numbers (see chapter 5.2.3) of sequences. When performing sequence alignments, there is a trade-off between sensitivity (the ability of identifying sequences with distant evolutionary relationships) and selectivity (the avoidance
of non-related matches with spurious high similarity scores). The latter might appear as a result of biased amino acid composition. Various sequence alignment algorithms and scoring parameters are in use and the choice of algorithm depends upon the type of comparison problem needed to be solved.

5.2.1 Pairwise Sequence Alignments

Pairwise alignments are performed either as local or global alignments. Local algorithms seek subsequence alignments to find the strongest region of similarity between two sequences (thereby disregarding dissimilarities outside that region). Local alignments are preferred when identifying shared conserved regions or domains, as well as shuffled domains, as a result of complex evolutionary pedigree (e.g. internal gene duplications). Such alignments are often used when screening DNA or protein databases as well as in exon-identification, where mRNA is compared to genomic DNA.

Global alignment algorithms compare the two aligned sequences from their respective beginnings to ends. These algorithms are the most suitable choice when the similarities span the full length of the sequences, or when homology has already been established. Scores generated by global alignment are calculated with or without gap-penalties at the ends of the sequences.

Alignment algorithms can be divided into two general classes. Rigorous comparison algorithms calculate the optimal similarity score, and heuristic algorithms are less computationally intense and are therefore faster. Several programs exist that are based on optimal algorithms. The first pair-wise sequence similarity algorithm was the optimal and global Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), which is available as the implementation Needle (www.ebi.ac.uk/emboss). In contrast, The application Water performs local, optimal alignments using the Smith-Waterman algorithm (Smith and Waterman, 1981; www.ebi.ac.uk/emboss).

The heuristic algorithms do not necessarily find the optimal score for every single sequence included in a search database. The two most commonly used heuristic comparison algorithms are BLAST (Basic Local Alignment Search Tool; Altschul et al., 1990) and FASTA (Pearson, 1990), which stands for Fast All (fast searches for ‘all’, meaning both protein and nucleotide comparisons). They are typically 5-50 times faster compared to the Smith-Waterman algorithm, since they examine only a portion of the potential alignments between sequences. However, in many cases FASTA and BLAST generate similar results compared to the more rigorous algorithms.

BLASTP (P for protein) is the most widely used rapid protein sequence comparison algorithm. It combines high sensitivity and selectivity, and seldom generates high spurious alignment scores for unrelated sequences. However, as many as 25% of all amino acids in protein sequences are situated in regions with severe biased amino acid composition (e.g. low complexity regions). Therefore, BLAST programs filter low complexity regions from proteins by utilizing the SEG algorithm (Wootton, 1994). The algorithms assess scores to the generated alignments which reflect a minimum of insertions, deletions and substitutions. Accordingly, highly similar sequences generate high scoring and distant sequences low scores.

Protein sequence alignment methods determine similarity by using substitution matrices, which assess scores for every possible mutation of one amino acid into another. Matrices for generating similarity scores have three main properties and can therefore differ in the following aspects: i) method used to construct the
matrix; ii) the scoring scale, and iii) the information content. Thus, mutations that are evolutionary unlikely to occur are assigned large negative scores, whereas positive scores are assessed to conservative changes (see figure 5.1; Pearson, 2000).

There are two types of commonly used matrices, PAM (Point Accepted Mutation) and BLOSUM (BLocks Of SUbstition Matrices). PAM is a molecular evolutionary model, which is constructed from observed residue replacements in closely related proteins (Dayhoff et al., 1978). An average change in 1% of all residue position corresponds to one PAM. Therefore homologous proteins might be recognized even though they are separated with more than 100 PAMs. This is due to the possibility of multiple mutations on some residues and absence of mutations on others. As evolutionary rates differ among proteins, no clear correlation between evolutionary time and PAM distance can be established.

The BLOSUMs are directly constructed from multiple sequence alignments (see chapter 5.2.3) of distantly related proteins, and not derived from closely related proteins (Henikoff and Henikoff, 1992). The PAM and BLOSUM matrices differ not only in the way they are constructed, but also in how they are used. PAMs are numbered based on the number of point accepted mutations (e.g. PAM120), whereas BLOSUMs are numbered according to the shared similarities among the multiple alignments used to construct them (e.g. BLOSUM80). PAM matrices with high numbers and BLOSUM matrices with low numbers are constructed for comparison of distantly related protein sequences. In conclusion, the choice of matrix is dependant on the diversity of the relation needed to establish. In general, BLOSUM62 (see figure 5.1) performs better than PAM matrices with BLASTP and FASTA (Henikoff and Henikoff, 1993).

Figure 5.1 A substitutorial matrix, BLOSUM62, which is used in sequence alignments. The matrix is derived from multiple sequences alignments of sequences with 62% sequence identity. Mutations, which are evolutionary unlikely to occur are assessed negative scores, whereas positive scores are assigned to conservative changes.
5.2.2 Practical Aspects of Pairwise Alignments

When compared to protein sequence comparisons, DNA sequence comparisons are significantly less informative due to several reasons. DNA sequences, which do not encode structural RNAs or proteins, have high divergence rates. As a consequence, there are difficulties in establishing reliable sequence homologies between sequences with a common ancestor older than 200 million years. In contrast, homology detection among protein sequences that diverged one billion years ago is possible (Pearson, 2000). Protein sequences harbor biochemical information within the amino acids themselves, which DNA sequences do not. For instance, there are large differences in chemical properties between isoleucine and glycine, and high similarities between lysine and arginine (i.e. size, hydrophobicity and charge; see chapter 2.3). Finally, the degeneracy of the genetic code results in many silent third-base codon changes, which do not change the encoded protein (Pearson, 1990).

For local alignment searches, the correlation between sequence similarity score, alignment length, and statistical significance, is complex. For instance, a database screening might generate two distinct hits, one 15-25 amino acid short domain with 85% sequence identity to the query sequence and one longer domain consisting of 50-80 residues sharing 30% identity with the query. Both hits generate identical similarity scores. Yet, the latter is more prone to be of biological relevance (Pearson and Miller, 1992).

Sequence-based searches are not suitable for identifying distant evolutionary relationships. Difficulties arise as the sequence similarities diminish, and around 20-30%, referred to as the twilight zone, relations are hard to establish (see figure 5.2; Doolittle et al., 1986). However, remote relations can be detected using methods such as fold recognition (see chapter 3.5; Jones, 1997), or iterated BLAST searches (see chapter 5.2.3; Altschul et al., 1997).

Figure 5.2 Limits of sequence similarity searches. As the evolutionary distance between sequences increases, the possibility of significantly establishing a relation decreases. In the so-called twilight zone (displayed as a dashed rectangle), relations are hard to establish with sequence-based searches (adapted from Pearson, 2000).
5.2.3 Multiple Sequence Alignments

Even though pairwise comparisons are essential when analyzing sequences, analyses of protein families sharing conserved characteristics, requires alignment and comparison of groups of sequences. This is done in multiple sequence alignments, which form the foundation for methods such as phylogenetic analysis, fold recognition (see chapter 3.5), and secondary structure prediction (Bateman et al., 2002).

The majority of multiple alignment algorithms are based on a progressive approach, which refers to the initial grouping of more similar sequences followed by the latter incorporation of more distantly related sequences (Boeckmann et al., 2003). The most widely used multiple sequence alignment method is ClustalW (Thompson et al., 1994), which exploits the evolutionary relationship between similar sequences. The method starts by computing an all-to-all pairwise comparison following the branching order of a family tree, in which more related sequences are aligning first and more distant thereafter. Consequently, ClustalW groups all the sequences based on similarities and then produces the final multiple alignment. ClustalW also generates the information needed to construct a phylogenetic tree.

The position-specific iterated BLAST (PSI-BLAST) is a hybrid algorithm, including elements from pairwise as well as multiple sequence alignment methods. The initial database search is followed by construction of position-specific sequence profiles, which, upon iteration, might be refined thereby raising the search sensitivity (Altschul et al., 1997). Besides the many advantages of PSI-BLAST, false incorporations of unrelated sequences and non-masked low complexity regions might bias the profiles and incorporate additional false positives in the following iterations (Holm, 1998).

5.3 Prediction of Signal Sequences and Transmembrane Regions

The identification of sequence motifs (see chapter 3.7), which direct proteins in the cell have been the goal for programs predicting signal peptides. SignalP takes into account that the residues at positions -3 and -1 relative the signal peptide cleavage site have to be neutral and small for cleavage to occur (von Heijne, 1983; von Heijne, 1985). SignalP can predict cleavage sites with an accuracy of 78% for eukaryotes and 89% for prokaryotes (Nielsen et al., 1997). Similar to signal peptides are signal anchors, which have signal sequences not cleaved by signal peptidases (von Heijne, 1988). A more recent version of SignalP has been developed, that distinguishes between cleaved signal peptides and uncleaved signal anchors with an accuracy of about 75% (Nielsen et al., 1999). Similar to SignalP is TargetP, which predicts subcellular localization of proteins (Emanuelsson et al., 2000). By using N-terminus protein sequence data, TargetP discriminates between secretory, mitochondrial, and chloroplast proteins. It performs better than other available similar tools, with an overall success-rate of 85% for plant sequences and 90% for nonplant sequences.

Membrane proteins are an important class of proteins, for which it is difficult to obtain atomic-resolution information about 3D structure. Accurate methods for predicting the locations of transmembrane helices are available. For instance, TMHMM (Trans Membrane Hidden Markov Model), a commonly used tool, predicts transmembrane helices correctly for about 95% of all proteins in a test data set (Krogh et al., 2001).
5.4 Biocomputing – a Practical Approach

The development and construction of biocomputational software applications is highly dependent on the type of problems that need to be solved. Different programming languages and database solutions are available, with respective strengths and weaknesses. The two most commonly used programming languages in biocomputing are Perl (perl.org) and Java (java.sun.com). They both have their own open-source projects which provide frameworks for handling biological data (BioJava (biojava.org) and BioPerl (biojava.org)). These frameworks include modules, classes and interfaces for file parsing, manipulation of biological sequences, and database client and server support. Perl is ideal for smaller scripts when less code is needed to be written and for construction of scripts without using an object-oriented approach. Java is more favourable when constructing larger software solutions resulting from more extensive visualization packages and cross-platform capabilities.

When designing and constructing novel software, every new problem is unique and therefore a general solution is not available. However, the system has to include general features such as proper storage and management of data, which involves a relevant and stable database-system and data-recovery facilities (back-ups). Moreover, the system needs to present and visualize the data in ways that facilitate the interpretation by the end-user. The system functionalities must be clearly defined and continuous interaction between software developers and end-users might be necessary.

In various large-scale, high-throughput approaches, the system needs to handle a large amount of data at high speed. Therefore, well-established prediction methods in combination with experimentally-gained knowledge might be useful (related publication A, B). The overall high production rates might result in a system, which focuses on high-throughput processing over extreme sensitivity. Yet, the approach has to be sensitive enough not to miss any relevant data.
Present Investigation
6. Objectives

This thesis describes the use of biocomputational approaches to analyse and characterize publicly available biological sequences in order to gain information about gene and protein functions (see figure 6.1). The work can be divided in three major research topics.

Firstly, novel software has been developed to facilitate selection of protein fragments subsequently used in antibody-based proteomics (publication I). Furthermore, the suggested design strategy, as well as experimental results, was evaluated in a separate systematic study (publication II). Secondly, additional software functionalities were developed to enable antibody recognition of splice variants (unpublished data, see chapter 7.4.2), and proteins with high sequence similarities, possibly evolved through duplication events (publication III). The previously mentioned strategies facilitated the move from a one-by-one manual gene analysis to processing genes in a high-throughput manner with a minimum of human interference. Finally, the data mining of the genomic sequence of a tree model organism, poplar (Populus trichocarpa), resulted in the identification of previously unknown genes belonging to the cellulose synthase (CesA) family (publication IV). These genes are thought to have evolved through duplication events at the gene or genome level.

The studies described include the use of genomic and proteomic data from human (publication I-III, unpublished data) as well as from P. trichocarpa (publication IV). Furthermore, sequence analysis such as homology searches (publication I-IV), PCR primer design (publication I, II), multiple alignments and phylogenetics (publication IV), have been performed.
Figure 6.1 A schematic overview of the publications (and unpublished data) presented in this thesis. 
**Publication III, IV** explore proteins that are thought to have evolved through duplication events (see chapter 1.2.2). **Publication I-III** as well as the **unpublished data** comprise studies involving the center-positioned PrEST (Protein Epitope Signature Tag; see chapter 7.2).
7. Design of Protein Epitope Signature Tags (PrESTs) for Antibody-based Proteomics

7.1 Background

On the 15th of August 2003, the Swedish human proteome resource (HPR) was initiated. It is an antibody-based proteomics approach, which aims for the generation of affinity-purified polyclonal antibodies towards recombinant protein fragments (see figure 7.1; related publication A). The antibodies are subsequently used in TMA studies (see chapter 3.6.4) on normal and diseased tissues (Kampf et al., 2004), as well as for in vitro detection (see chapter 3.6.5) and analysis on protein microarrays (see chapter 3.6.3). Protein profiling is also performed on cell-lines, which might provide a valuable complement to tissue localization studies. These cell-lines have the advantage of being thoroughly controlled during cultivation. Some cell-lines consist of cells from rare types of cancers, which are difficult to get tissue samples of. Well-characterized cell-lines might also provide negative and positive controls in the IHC assays (Andersson et al., manuscript in preparation).

The HPR approach is based on a pilot study of the human chromosome 21, which comprised areas such as protein expression and purification, antibody staining, as well as biocomputing (Agaton et al., 2003). Antibodies are generated towards recombinant protein fragments, protein epitope signature tags (PrESTs), and are given mono-specific properties (see chapter 3.6.1) by stringent purification of polyclonal antisera using the PrESTs as affinity ligands.

An important part of the HPR strategy is the selection of recombinant protein fragments, which typically span around 100-150 residues. These PrEST fragments are designed by the use of in-house prediction software (Bishop; publication I) and are selected on the basis of favorable sequence properties for subsequent cloning, expression and purification.

There is a need for specific and quality controlled antibodies, displaying a minimum of cross-reactivity to other proteins than their intended targets. The HPR approach deals with this issue in multiple ways. The PrEST design strategy emphasizes on locating PrESTs in regions of low sequence similarity to other human proteins. This reduces the risk for the msAbs to cross-react with similar proteins. This is accomplished by the PrEST design software (publication I), which selects recombinant protein fragments based on a local alignment scanning procedure.
Figure 7.1 A schematic overview of the HPR project, which includes several computer-assisted modules. PrEST fragments are determined by the use of specific design strategies, with regards taken to the properties of the target proteins (publication I-III, unpublished data). Quality control by cycle sequencing (Blakesley, 1993) ensures the amplification of a correct PrEST fragment. Data belonging to the designed PrESTs, and to the generated antibodies are stored and displayed in a publicly available data repository (www.proteinatlas.org). In addition, the entire project is integrated with a laboratory information management system (LIMS), which ensures proper processing and tracking of all the data.

Several quality control steps are included in the HPR project. Sequence verification of clones in the cloning procedure is obtained by sequencing and comparisons with the in silico generated PrEST. The antibodies are also used on reverse phase protein fragment microarrays (see chapter 3.6.3), on which the absence of cross-reactivity indicates the presence of mono-specific antibodies. In addition, the antibodies are verified by Western blot analysis (see chapter 3.6.5), and these results are compared to the TMA data. Reimmunizations of respective PrESTs and then comparison of different protein expression patterns further provide a means of validating the results. Also, two msAbs directed towards different regions of the same protein and designed in appropriate software (publication I), should provide correlating staining patterns (see table 7.2). The generated antibodies have been included in comparative studies together with well-characterized, commercial, and widely used mAbs. The similar results provided by this study corroborate the high specificity of the HPR-generated msAbs (Nilsson et al., 2005).
Following TMA studies of antibodies, corresponding data and data regarding the PrEST towards which the antibody has been directed, are stored in a publicly available protein atlas (www.proteinatlas.org; related publication A). This repository includes data such as high-resolution TMA images, target gene data, and validation of the msAbs. A research project of this magnitude encompasses several laboratory assays and generates vast amounts of data. A laboratory information management system (LIMS) facilitates tracking and handling of the complete laboratory procedures and of the sequence-data used for PrEST selection (see chapter 7.4). The LIMS is integrated with the publicly available data repository, as well as with current PrEST design software (publication I).

7.2 Design of Protein Fragments for Recombinant Protein Production (publication I)

A novel PrEST selection software has been developed (Bishop; see figure 7.2), predicting protein fragments based on several empirically determined requirements. Bishop allows for high-throughput selection of PrESTs with a minimum of end-user interference, providing a system which was integrated with several publicly available prediction tools (see below). In contrast, the previously used design strategy (Agaton et al., 2003), was performed in a more manual gene-by-gene fashion and required the involvement of several different programs. The selection requisites in Bishop facilitate the subsequent laboratory procedures and promote selection of the most suitable PrEST fragment for any given gene product.

Unwanted binding of the produced msAbs needed to be kept at a minimum. This was accomplished by a local alignment scanning procedure based on BLASTP (Altschul et al., 1990), which ensured the selection of fragments exhibiting lowest possible sequence similarity to other human proteins. The procedure used a sliding window approach that traversed the target protein sequence, moving one amino acid per iteration. The size of the window could be individually determined by the end-user, but the default size was 125 amino acids.

Transmembrane regions cause problems in the expression and purification procedures in the subsequently used expression host E. coli (Miroux and Walker, 1996), as well as being poorly accessible to antibodies in tissue profiling. Transmembrane regions were therefore avoided by the use of the prediction program TMHMM (Krogh et al., 2001), which was executed by the main program.

Signal peptides (see chapter 3.7.3) are cleaved off from proteins during translocation and are therefore inappropriate targets for antibody recognition. Bishop allows for the exclusion of signal peptides by visualizing the predictions produced by SignalP (Nielsen et al., 1997).

Finally, avoidance of restriction enzyme motifs used in subsequent cloning is a necessity. Therefore, the corresponding cDNA sequence of the target protein was screened for restriction enzyme sites. The software allows for searching of any desired restriction site or basic sequence motif. Based on the four requirements mentioned above, two non-overlapping PrEST fragments, if possible, were automatically predicted for all analyzed target proteins.
Design of Protein Epitope Signature Tags (PrESTs) for Antibody-based Proteomics

### Table A

<table>
<thead>
<tr>
<th>PrEST</th>
<th>Description</th>
<th>Location (aa)</th>
<th>Score</th>
<th>E-value</th>
<th>% Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrEST A</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>PrEST B</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Table B

<table>
<thead>
<tr>
<th>PrEST Design</th>
<th>Candidate Epitopes</th>
<th>Score (E-value)</th>
<th>% Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrEST C</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>PrEST D</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Table C

<table>
<thead>
<tr>
<th>T9</th>
<th>KP</th>
<th>TP</th>
<th>KY</th>
<th>KQ</th>
<th>SY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>54</td>
<td>54</td>
<td>54</td>
<td>54</td>
<td>54</td>
</tr>
</tbody>
</table>

---

Note: The tables and diagrams represent data on the design and analysis of protein epitope signature tags (PrESTs) for antibody-based proteomics. The tables include information on candidate epitopes, scores, and percent identity, while the diagrams illustrate the design process and analysis metrics.
Figure 7.2 The five different tabs of the Bishop software (A-E): A: The sequence alignment tab, in which the target protein sequence is aligned to all human proteins. The user also has the possibility to exclude splice isoforms (see chapter 7.4.2, unpublished data) or high sequence identity proteins (see publication II) from the subsequent sequence similarity search (displayed in tab C). B: The PrEST data tab, where data concerning the predicted PrESTs are shown. For instance, PrEST locations, sequence identities and amino acid sequences. C: The sequence feature tab, which displays several transcript and protein features as horizontal bars. These are (from top to bottom): transcript sequence, protein sequence, and predicted PrEST sequences. In addition (although not displayed in the figure) Bishop allows for the exclusion of signal peptides, transmembrane regions, and restriction enzyme sites located within the PrEST sequences. D: The PCR primer design tab, in which primers are selected based on non-stringent criteria. E: The result tab, in which all previously generated data are displayed before committing the predicted PrEST, allowing for subsequent ordering of the designed PCR primers.
As previously mentioned, Bishop allows for selection of fragments with minimum sequence similarity to other proteins. However, in some cases there might be a trade-off between locating the PrEST in a target region, and identifying the region of lowest possible sequence similarity. Therefore, the region of lowest similarity might not always be the optimal choice. For instance, when designing PrESTs representing groups of proteins of high sequence similarity (publication III) or PrESTs located in common or unique parts of multiple splice variants (unpublished data, see chapter 7.4.2), other design criteria might be more relevant. Also, it is possible to perform a manual PrEST selection in Bishop (see figure 7.2, tab B). For instance, when designing PrESTs towards conserved active domains or extra-cellular parts of receptor proteins. Nevertheless, designing PrESTs in the region of lowest similarity should be the primary choice when processing genes in a high-throughput manner.

Following PrEST design, a primer design module was included in Bishop. An appropriate melting temperature (Tm), as well as a suitable length of the primer, can be set by the end-user. The primer design is based on non-stringent criteria such as ending the primer sequence with a C or a G and avoiding continuous stretches longer than three identical nucleotides. However, when compared to a more stringent primer design this non-stringent criteria showed only slightly lower success rates in the subsequent RT-PCR and sequencing modules (publication II). Following primer design, all data related to the target protein is transferred to the LIMS (see chapter 7.1), from where the primer ordering is performed. Bishop is fully integrated with the LIMS, performing data exchanges and overnight synchronization of information.

Due to the existence of discontinous and linear epitopes (see chapter 3.6.2), it would be beneficial to obtain antibodies recognizing both types of epitopes. That accomplished, the antibodies might be used in assays including proteins in a native fold, as well as studies comprising slightly denaturated proteins.

In order to explore the capabilities of the cloned and expressed PrESTs to form discontinous epitopes, designed protein fragments were evaluated by analyzing their 3D structures. This was performed by marking the PrEST regions on known available protein structures retrieved from the protein data bank (see figure 3 in appendix I; Bernstein et al., 1977). As antigens located at the N- or C-termini are commonly used for the expression of recombinant protein fragments for antibody production (Agaton et al., 2003), these regions were also included in the analysis. The C- and N-termini of the protein were marked in regions corresponding to the PrEST lengths. This limited study suggested that the PrESTs designed by the software package seem to have a relatively good potential to display some conformational epitopes. However, this is not a sufficient proof of the expressed PrESTs ability to form secondary structures resembling the native protein. On the other hand, the PrEST fragments seem to be large enough to provide surface exposure for a number of residues on the native protein. This is of great importance for antibody recognition of the target.

The protein fragments are, in the majority of cases, shorter than full-length proteins and, therefore, a native protein fold of the expressed PrESTs is not obtained (see figure 7.3). Also, the immunized PrESTs are most likely denaturated upon immunization due to the exposure to adjuvants. Still, these protein-fold modifications seem to resemble the denaturated forms of the proteins located in formalin-fixated tissues in TMAs, since the majority of the msAbs actually recognize their intended targets.
Figure 7.3 A comparison of a full-length protein in relation to a PrEST, which is generated by the expression of a recombinant protein fragment. In the majority of cases the PrEST is shorter than a full-length protein. This in combination with the use of adjuvants in the immunizations, results in a slightly modified fold of the PrEST fragments.

7.3 A Systematic Analysis of Designed PrESTs (publication II)
PrESTs designed in the previously described software (publication I) as well as in a similar software named ProteinWeaver (Affibody, Bromma, Sweden), were analysed. The success rates of the PrEST design strategy was investigated in this study, with respect to target protein properties as well as PCR primer design approaches. The analysis was performed on proteins encoded by genes from human chromosomes 14, 22 and X together with target proteins requested from several research collaborations (see chapter 7.4). The proteins submitted from collaborators were identified by the use of several different proteomic methods, for example, MS (see chapter 3.4) and methods involving protein-protein interactions (see chapter 3.9). This systematic study included a total of 6,412 PrEST fragments located on proteins encoded by 4,263 genes, corresponding to 16% of the genes currently included in the Ensembl dataset. Chromosomes 14, 22, and X were over-representated, due to the focus on these chromosomes (see table 7.1).

The design of PCR-primers for amplification of the PrESTs, was performed using a high stringency- and a low stringency approach (previously described in publication I). The high stringency algorithm included additional criteria, considering dimer- and hairpin formation. Comparison of results from the two primer design approaches revealed a slightly higher success rate for the high stringency algorithm when compared to the low stringency approach, showing 82% and 74%, respectively. The success rates were calculated from subsequent RT-PCR analysis and sequencing. However, the actual difference between the two approaches is probably lower, since the high stringency approach was recently added to the HPR project. Thus, continuous improvement and refinement of laboratory procedures might have contributed to the higher success rates of the stringent primer design approach.
Table 7.1 The current status of the number of designed PrESTs and sequence verified clones compared to the gene count on all human chromosomes (The gene count is according to Ensembl August 2005).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of Ensembl genes</th>
<th>Genes with PrEST</th>
<th>Number of PrEST clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2158</td>
<td>246</td>
<td>204</td>
</tr>
<tr>
<td>2</td>
<td>1443</td>
<td>163</td>
<td>135</td>
</tr>
<tr>
<td>3</td>
<td>1146</td>
<td>162</td>
<td>142</td>
</tr>
<tr>
<td>4</td>
<td>850</td>
<td>105</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>964</td>
<td>125</td>
<td>103</td>
</tr>
<tr>
<td>6</td>
<td>1146</td>
<td>141</td>
<td>117</td>
</tr>
<tr>
<td>7</td>
<td>1094</td>
<td>145</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>785</td>
<td>93</td>
<td>84</td>
</tr>
<tr>
<td>9</td>
<td>904</td>
<td>110</td>
<td>96</td>
</tr>
<tr>
<td>10</td>
<td>874</td>
<td>84</td>
<td>74</td>
</tr>
<tr>
<td>11</td>
<td>1397</td>
<td>142</td>
<td>123</td>
</tr>
<tr>
<td>12</td>
<td>1085</td>
<td>157</td>
<td>137</td>
</tr>
<tr>
<td>13</td>
<td>394</td>
<td>37</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>713</td>
<td>639</td>
<td>476</td>
</tr>
<tr>
<td>15</td>
<td>702</td>
<td>73</td>
<td>68</td>
</tr>
<tr>
<td>16</td>
<td>947</td>
<td>98</td>
<td>73</td>
</tr>
<tr>
<td>17</td>
<td>1219</td>
<td>145</td>
<td>123</td>
</tr>
<tr>
<td>18</td>
<td>316</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>19</td>
<td>1455</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td>622</td>
<td>440</td>
<td>440</td>
</tr>
<tr>
<td>21</td>
<td>265</td>
<td>718</td>
<td>718</td>
</tr>
<tr>
<td>22</td>
<td>535</td>
<td>129</td>
<td>129</td>
</tr>
<tr>
<td>X</td>
<td>957</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>Y</td>
<td>129</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to increase the efficiency of the RT-PCR analysis, and properly validate the generated antibodies (see chapter 7.1), the primary strategy was to design two PrESTs for each protein. The design strategy successfully selected two PrESTs on the same protein for approximately half of the investigated proteins. For shorter proteins lacking sufficient length, a single PrEST was designed. In cases where the protein was shorter than 100 amino acids, the entire protein sequence was selected as antigen. Proteins shorter than 100 residues constituted 5.3% of the total protein count in the Ensembl dataset (v26.35).

Failures in PrEST design were a result of interference by signal peptides, restriction enzyme sites or transmembrane regions. Additionaly, some proteins were too short for design. Together, these failures constituted approximately one percent of the total number of investigated proteins. Moreover, this study considered target proteins with high sequence similarities (above 80% pairwise sequence identity) to other human proteins as fall-out proteins, which in this case constituted around 13% of the total protein data set. Proteins encoded by genes located on human chromosomes 22 and X constituted a larger part of these fall-outs. The number of fall-out proteins from encoding genes from chromosome 14 was significantly lower due to few duplicated regions (Cheung et al., 2003). In contrast, chromosomes 22 and X are known to have high numbers of duplicated segments. However, these genes might be processed with a strategy solely developed for handling of such high sequence similarity proteins (publication III).
7.4 Specific PrEST Design Strategies

In the HPR, protein and cDNA sequences from Ensembl were the primary data sources for PrEST design. After retrieval, the sequences were initially grouped according to a basic classification scheme, resulting in three different protein categories. Firstly, the nonredundant set, comprised one representative protein from each unique gene locus (see chapter 2.2). For genes with splice variants, the longest isoform was selected. Secondly, a portion of the nonredundant set encompassed high sequence similarity proteins, which were processed using a specific design strategy (**publication III**). The final category contained additional spliced isoforms (see chapter 2.2) not covered by the first category. In this case, the PrEST design was directed towards specific splice variants and towards regions shared by several or all of the protein isoforms (**unpublished data**, see chapter 7.4.2).

Currently, the HPR is focusing on genes on human chromosomes 14, 22, and X, as well as specific requests from research collaborators. These genes include biomarkers for human cancers and stem cells, as well as genes from other species. For instance, evolutionary relevant chimpanzee genes, and *Populus* (see chapter 8.1) genes involved in plant growth and development.

7.4.1 Antibody-based Proteomics on High Sequence Similarity Proteins (**publication III**)

As previously mentioned (see chapter 7.4.1), three different protein classifications have been established. Accordingly, different PrEST design strategies have been developed for individual handling of the different categories. The second category consisted of high sequence similarity proteins (HSSPs), exhibiting more than 80% pairwise sequence identity to human proteins from different genes. This sequence identity limit also corresponded to the fall-out proteins previously described (**publication II**). A design strategy for processing HSSPs was developed, where the proteins initially were grouped in clusters based on a maximum of 80% sequence identity. The PrEST design is thereafter focused on locating PrESTs in regions which were common to all cluster members. As a proof of principle, antibodies generated towards such cluster specific PrESTs have been analyzed by Western blot analysis using whole-cell and tissue protein extracts. The ability of identifying proteins within a high sequence similarity cluster by the use of a single antibody was thus investigated. In addition, novel software was developed to estimate the minimum number of PrESTs needed for coverage of the HSSPs.

Initially, the HSSPs were identified using an all-to-all sequence similarity search, in which all human proteins predicted by Ensembl were included. Proteins from different genes sharing more than 80% sequence identities over the shortest of the aligned sequences were selected. This resulted in 3,250 identified HSSPs, corresponding to 14.6% of the non-redundant set of proteins. The subsequent PrEST design was facilitated using an enhanced version of Bishop (**publication I**). This allowed for the simultaneous visualization of all proteins within a cluster, and for location of PrESTs within aligned high sequence similarity regions. This version of Bishop contained an additional sequence similarity scanning procedure, in which the target proteins were aligned to all proteins within their respective cluster. The similarities to all human proteins could additionally be visualized. The optimal PrEST location within the aligned high sequence similarity region was the region of highest local sequence similarity among all proteins in a cluster. This local similarity was generated as previously described (**publication I**).
To explore the capacity of the generated antibodies in order to identify all members of the HSSP clusters, an in vitro detection by Western blot analysis was performed on antibodies towards cluster specific PrESTs. An example showed (see figure 3 in appendix II) that all cluster members could be identified in one of the cell lines. Although limited, this study implies the possibility to use this design strategy for generation of antibodies capable of recognizing all proteins in a high sequence similarity cluster, despite the fact that their sequences are not identical.

An estimation of the the minimum number of PrESTs sufficient to cover all the previously identified HSSPs was made possible by the development of novel software. This program performed homology searches using BLASTP (Altschul et al., 1990) which resulted in the identification of HSSPs. The proteins were subsequently assembled in clusters based on common hit sequences. The protein sequences in each cluster were then aligned and grouped into sub-clusters. Proteins with common regions of sufficient size (in this case 100 residues), where grouped in the same sub-cluster. Finally, in order to remove redundancies and to cover as many proteins as possible with a minimum of PrESTs, the entire set of sub-clusters was subjected to a sorting. In this sorting large sub-clusters were favored. This resulted in a total of 1,124 sub-clusters and in 186 single proteins remaining after the final sorting. This implies that 1,310 PrESTs are necessary for coverage of the complete set of HSSPs. In average, 2.73 proteins were covered by a region large enough to accommodate a PrEST fragment.

In conclusion, this resulted in a total of 20,281 PrESTs, sufficient for coverage of the total nonredundant set of proteins, corresponding to the entire set of human genes (22,221 according to Ensembl v26.35). To facilitate the study of multiple splice variants belonging to these genes, a specific approach was developed (see below). In addition, antibodies generated by this HSSP strategy could potentially be important for investigations of gene duplications and gene families.

7.4.2 Design of Isoform Specific Protein Fragments (unpublished data)

The final protein category, as defined by the HPR, was proteins with multiple splice variants, which were not included in the nonredundant set of proteins. According to Ensembl, about 32% of the human genes had multiple isoforms and the majority (30.3%) of these genes had between two and four observed variants. The primary design strategy (see chapter 7.4) focused on the nonredundant set by selecting a single representative protein for genes with several splice variants. However, designing PrESTs unique to a specific isoform or common to all of the splice variants might be desirable. A strategy for designing PrESTs on genes with several splice variants was therefore developed. The approach included novel software (ExonViewer), which was used in combination with Bishop (publication I). ExonViewer used exon boundaries from the Ensembl database and visualized them in a user-friendly interface (see figure 7.4). A Perl-based script for data retrieval and pre-processing of data was developed, whereas the main software was written in Java. Both parts of the system used the BioPerl and BioJava frameworks for handling of sequence data and for visualization of exons (see chapter 4.4). This allowed for in-depth studies of different splicing events. Designing PrESTs on unique splice variants, or regions common to several or all variants was also possible.
In this design strategy, the target genes were visualized in ExonViewer and regions common or unique to all or several of the isoforms, were selected as candidate PrEST design regions. These regions were at least 100 continuous amino acids in length. Yet, any desired length limit could be defined. Subsequently, these regions were exported to Bishop (publication I) for convenient design of protein fragments.
TMA results from four antibodies generated against two pairs of PrESTs were used as means to validate the strategy. Each pair was designed on the same target protein. The first pair of PrESTs was located on a region unique to one of the isoforms on the VWF gene (ENSG00000110799; see figure 7.4A). The second pair was placed on a region common to both isoforms of the TJP1 gene (ENSG00000104067; see figure 7.4B). A comparison of the protein expression patterns within the PrEST-pairs in 43 different tissues, revealed a clear correlation between PrESTs designed on the same protein (see table 7.2).

Table 7.2 A comparison of the tissue distribution in 43 different tissues of the four investigated antibodies. Red color indicates strong staining, orange color medium staining, and yellow color weak staining. Blue color displays no staining. Tissues with no available pictures were excluded from the study.

The different colors in the protein expression patterns indicated the presence of the proteins in the tissues. Red color indicated strong staining, orange medium staining, and yellow color indicated weak staining, whereas blue color indicated no staining. The majority of tissues showed similar staining patterns for PrESTs located on the same protein (antibody A-B and C-D in table 8.1). Similar staining in the included tissues of two PrESTs located on the same protein could be seen as a validation and confirmation of the design strategy (see chapter 7.1).

Yet, according to the applied three-level scheme, different staining intensities (weak, intermediate, and strong staining) might reveal a similar protein expression upon closer examination. Differences in staining patterns might be a result of variations in the tissues, due to that nonconsecutive sections of tissues have been used. The tissue-sampling has been performed on tissues from different individuals, and the tissues might therefore not be representative. Moreover, the annotation might have been performed by different annotators. The antibodies used in the TMAs...
might also have been diluted in different ways. Furthermore, the splice variants in the data source might not have been fully curated, i.e. unknown isoforms might have been present which interfered with the results.

The tissue distribution data used in this study is merely for validation of the strategy. This limited data indicates that the design strategy is promising. However, more antibodies need to be generated for a large-scale evaluation. Furthermore, PrESTs can be directed to unique regions of several isoforms thereby revealing clues about the different tissue distribution patterns.
8. Identification and Characterization of Gene Family Members in a Fully Sequenced Tree Genome

8.1 Populus, a Tree Model System

Trees belonging to the genus *Populus*, (i.e. poplars, including cottonwoods and aspens) have several attributes, which have contributed to their emergence as predominant model organisms for molecular tree biotechnology (reviewed by Brunner et al., 2004). *Populus* members are widely distributed throughout the Northern hemisphere. For instance, black cottonwood (*Populus trichocarpa*) grows in diverse habitats such as those in Alaska and Mexico (Brunner et al., 2004). This adaptivity and genetic variation make *Populus* trees ideal for studies of adaptive evolution (see chapter 1.2.1). Furthermore, their short generation time and easy transformation enable functional genomic studies.

The recent sequencing of the genome of *P. trichocarpa* (genome.jgi-psf.org/poptr1) provides means to further study and characterize the fairly small genome of this organism (see table 5.1). In comparison, loblolly pine (*Pinus taeda L.*) has a large and complex genome (about 20 times larger than *Populus*), and exhibits a long generation time. Pine is therefore an unsuitable model organism for tree genomics (Whetten et al., 2001).

8.2 Cellulose Synthases

The capability of forming wood is a specific feature distinguishing trees from plants. Cellulose, which is the primary component in wood, is synthesized by large membrane-bound protein complexes (Doblin et al., 2002). These complexes include catalytic subunits encoded by *CesA* gene family members. *CesA* enzymes are glycosyl transferases (GTs). GTs form glycosidic bonds by catalyzing the transfer of sugar moities from appropriate sugar donors to acceptor molecule, thus creating carbohydrates such as cellulose (Coutinho et al., 2003). Conserved residues present in all GTs, belonging to the same family as *CesA* enzymes, include three D residues as well as a QxxRW motif, proposed to be involved in substrate- and acceptor binding, as well as processivity (Saxena et al., 1995; Campbell et al., 1997; Garinot-Schneider et al., 2000). In addition, *CesA* family members harbor two zinc finger domains positioned near the N-terminus, as well as two hypervariable regions (HVR-1 and HVR-2) specific to plant genes (see figure 8.1, 8.2). *CesA* proteins are generally around 1,000 amino acids in length with eight predicted transmembrane regions.
8.3 Identification of Previously Unknown CesA Family Members
(publication IV)

In this study, the genome sequence of *P. trichocarpa* was mined for unidentified CesA family members. This was done by querying the genome with previously identified CesA EST- and full-length gene sequences from the closely related hybrid aspen (*P. tremula x tremuloides*; Djerbi et al., 2004). The screening was performed with BLAST and resulted in the identification of 18 putative CesA genes based on sequence similarities alone. Subsequently, the corresponding protein sequences were subjected to an all-to-all global sequence alignment search using the software Needle (www.ebi.ac.uk/emboss). This analysis revealed a clear sequence similarity pattern, which distinctly grouped the CesA genes into seven pairs, one single sequence, as well as a group of three sequences.

The 18 putative CesA gene sequences were further analyzed and conserved sequence characteristics were verified for confirmation of true CesA family membership (see chapter 8.2; figure 8.1). All CesA characteristics were identified in the 18 putative CesA sequences using tools such as ClustalW (see figure 9.2; Thompson et al., 1994), TMHMM (Krogh et al., 2001), as well as manual analysis. This extensive sequence analysis implies that the collected genes are indeed true CesA family members.
A thorough phylogenetic analysis was performed, which included the 18 identified CesA protein sequences together with CesA protein sequences and EST sequences from other plant genomes, i.e. hybrid aspen (P. tremula x tremuloides), quaking aspen (P. tremuloides), cotton (Gossypium hirsutum), rice (Oryza sativa), thale cress (A. thaliana) and maize (Zea mays). This analysis was restricted to conserved regions among CesA family members, thus the HVR-1 and -2 regions (see figure 8.1, 8.2) were excluded. According the phylogenetic tree generated by the Bayesian method

Figure 8.2 A multiple alignment showing the conserved sequence features common to cellulose synthase family members (these features are also presented in figure 8.1).
(see figure 2 in appendix IV), the analysis further implied that PtCesA2, 4, and 6, as well as PtCesA5, 7, and 8 (see figure 8.3) are most likely monophyletic (arise from a single common ancestral gene). Previous studies involving the PtrCesA4 and 5 (present in figure 8.3A) suggest their involvement in primary cell wall synthesis (Kalluri and Joshi, 2003; Kalluri and Joshi, 2004). Furthermore, PttCesA2 and 4 (present in figure 8.3B) are constitutively expressed in several different tissues (Djerbi et al., 2004). It is therefore tempting to speculate that PtCesA2, 4 and 6 are active in synthesis of primary cell walls, and that PtCesA5, 7 and 8 are constitutively expressed. However, more extensive studies need to be performed in order fully determine their function.

Sequence comparison alone does not shed light on whether genes might be transcriptionally silent or not. In order to elucidate whether these newly discovered CesA genes may be actively transcribed, corresponding EST sequences were retrieved from PopulusDB (Sterky et al., 2004). The CesA sequences were mapped to contigs in PopulusDB, which are representations of transcript sequences consisting of sets of overlapping clones. Alignments more similar than the applied limit of 94% identical nucleotides, over 95% of the contig length, were mapped. The identification of individual EST sequences in PopulusDB was not satisfactory due to the high sequence similarities between paralogous CesA gene pairs and the location of several EST sequences in conserved regions of the gene sequences. Yet, previous gene expression studies in hybrid aspen implies that both genes in at least one pair, CesA3-1 and CesA3-2, are actively transcribed (Djerbi et al., 2004).

The full length CesA gene sequences in hybrid aspen were further used to confirm the number of identified CesA genes in fully sequenced genomes from other plant taxa, thale cress (A. thaliana) and rice (Oryza Sativa). The analysis corroborates previous analyses, showing the presence of ten CesA family members in respective genome.

**Figure 8.3** Phylogenetic relationships of CesA genes from Populus trichocarpa (Pt), Populus tremula x tremuloides (Ptt), Populus tremuloides (Ptr), Arabidopsis thaliana (At), Gossypium hirsutum (Gh), Hordeum vulgare (Hv), Oryza sativa (Os), and Zea mays (Zm). The analysis is based on predicted amino acid sequences derived from CesA genes in fully sequenced genomes, full-length cDNA copies of putative CesA genes, and selected partial cDNAs (denoted by †). The numbers along the edges show the a posteriori probabilities of the partitions induced by the edges.
Concluding Remarks

The human genome project has provided large amounts of data that has truly opened the field of sequence analysis, sequence comparisons, and biocomputing. Knowledge encompassing computer programming and data analysis, as well as biotechnology, will be increasingly important for researchers of tomorrow.

The PrEST selection strategies described in this thesis (publication I, III, unpublished data) provide a framework for designing PrESTs for all previously defined categories of genes (see chapter 7.4). These strategies are built around experimentally defined criteria. Specifically directed software and sequence analysis tools developed to facilitate the move from a one-by-one sample analysis to high-throughput handling of data (publication I-III, unpublished data, related publication B), will be of great importance. Due to the wide range of unsolved problems and the uniqueness of research projects, generally applied systems will not be suitable in the majority of cases. Instead, most software solutions will need to be custom made for specific purposes.

Large scale proteomic efforts such as the HPR (see chapter 7.1, related publication A) will provide a vast amount of data and shed further light on the functions and localization of human gene products. Different antibody-based approaches need to be synchronized in order to avoid redundant target genes and allow for comparative studies. Furthermore, as the specificity of antibodies is of great importance, proper validations need to be performed by different assays in a systematic fashion. Cross-reactivities can thus be kept at a minimum and the quality assured antibodies will be of great value to the proteomic research community. In contrast, small scale in-depth investigations of individual target genes or gene families (publication IV) will still be performed, as these studies need to be done by researchers having expert knowledge regarding these targets. Collaborations between small-scale and large high-throughput projects will be advantageous for both parts. For instance, in-depth knowledge from small-scale projects might direct large-scale efforts towards identification of relevant target genes.

Currently, we only view the tip of the iceberg of functional genomic and proteomic research and the use of computational methods is now intrinsic. All future research initiatives will incorporate and integrate computer-assisted methods into their experimental setup. The level of complexity and the amount of the generated proteomic data will increase, and with that more challenging biological questions will need to be answered. As a consequence, bioinformatic approaches will be more prominent than ever before.
List of Abbreviations

2DE    Two-Dimensional gel Electrophoresis
3D     Three-Dimensional
3’     Three prime
5’     Five prime
A      Adenine
BIND   Biomolecular Interaction Network Database
BLAST  Basic Local Alignment Search Tool
BLASTP Basic Local Alignment Search Tool Protein
BLOSUM BLocks Of SUbstitution Matrices
C      Cytosine
cDNA   complementary DNA
CesA   cellulose synthase
DDBJ   DataBank of Japan
DIP    Database of Interacting Proteins
DNA    DeoxyriboNucleic Acid
EBI    European Bioinformatics Institute
EC     Enzymatic activity Classification
EMBL   European Molecular Biology Laboratory
EST    Expressed Sequence Tag
FASTA FAST All
G      Guanine
GO     Gene Ontology
GT     Glycosyl Transferases
HGP    Human Genome Project
HPR    Swedish Human Proteome Resource
HSSP   High Sequence Similarity Proteins
HVR    HyperVariable Region
IHC    ImmunoHistoChemistry
IMAC   Immobilized Metal ion Affinity Chromatography
kb     kilo bases
LIMS   Laboratory Information Management System
mAbs   monoclonal Antibodies
mRNA   messenger RibO nucleic Acid
MS     Mass Spectrometry
msAbs  mono-specific Antibodies
NCBI   National center for biotechnology information
pAbs   polyclonal Antibodies
PAM    Point Accepted Mutation
PCR    Polymerase Chain Reaction
PDB    Protein Data Bank
Pfam   Protein families database
PIR    Protein Information Resource
PrEST  Protein Epitope Signature Tag
PSI-BLAST Position-Specific Iterated BLAST
PTM    Post Translational Modification
PVDF   PolyVinylidene DiFluoride
RE     Repetitive Elements
RNA    RibO nucleic Acid
RT-PCR Reverse Transcriptase Polymerase Chain Reaction
SIB    Swiss Institute of Bioinformatics
SMART Simple Modular Architecture Research Tool
<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>Single Nucleotide Polymorphisms</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>TIGR</td>
<td>The Institute of Genomic Research</td>
</tr>
<tr>
<td>TMA</td>
<td>Tissue MicroArray</td>
</tr>
<tr>
<td>TMHMM</td>
<td>Trans Membrane Hidden Markov Model</td>
</tr>
<tr>
<td>UniParc</td>
<td>UniProt archive</td>
</tr>
<tr>
<td>UniProt</td>
<td>Universal Protein resource</td>
</tr>
<tr>
<td>UNIPROT</td>
<td>UniProt knowledgebase</td>
</tr>
<tr>
<td>UniRef</td>
<td>UniProt NREF</td>
</tr>
<tr>
<td>USCS</td>
<td>University of California Santa Cruz</td>
</tr>
<tr>
<td>VEGA</td>
<td>Vertebrate Genome Annotation</td>
</tr>
</tbody>
</table>
Acknowledgements

I am grateful to a lot of persons, as you all know this is no one-man show. Without the following persons at work and in my life, I doubt the research over the last five years and this thesis would have been possible.

Persons at the Institution of Biotechnology Royal Institute of Technology:

Prof. Mathias Uhlen (My very, very, very busy supervisor. You have given me invaluable guidance in the field of antibody-based proteomics, as well as teaching me the importance of presenting research data in a convenient way. Furthermore, you have provided me with splendid ideas regarding our research)

Dr. Fredrik Sterky (You have helped me tremendously over the last years, structuralizing our research and motivating me. I really appreciate that your door always is open for me. At times when I tend to forget the focus in our research and present drifted-away ideas, you always keep me back on track)

Prof. Stefan Ståhl (My former diploma work supervisor, who guided me in the field of biosynthesis of bioactive marine natural products, currently a sock-walking hardrocking Dean of School of Biotechnology. Initially, you provided me with a PhD position, and I am truly thankful to you)

Prof. Per-Åke Nygren (My current next-door-neighbour, for always giving direct and specific answers regarding my research), Assoc. Prof. Sophia Hober (For directing the lab-course in a structuralized and fun way), Prof. Joakim Lundeberg (For valuable advice concerning my research and for lending me your room when your were at daddy-leave)

Dr. Anja Persson, Dr. Jenny Ottoson, Dr. Henrik Wernerus and Dr. Cristina Al-Khalili Szigyarto (For bringing clarity to my questions)

Dr. Per Unneberg (For much advice and computational guidance during the old days, when I was a novice in the field of computers and bioinformatics)

Dr. Afshin Ahmadian and Erik Petterson (For showing me the world of parallel SNP amplification. Our meetings were the most laid-back I’ve ever encountered during my PhD studies)

My old room-buddies in B3:1049 (Esther Edlund-Rose, Ronny Falk, Olof Nord, Emma Lundberg, Tove Alm, and Cecilia Laurell. I still miss you, and the quite atmosphere we had. Tove and Esther, thank you for being the party-hardy girls that you are), My intermediate room-buddies in B3:1001’ (Lisa Berglund, Carl Hamsten, Johan Rockberg, Dr. Åsa Sivertsson, Dr. Cristina Al-Khalili Szigyarto, and Dr. Charlotte Agaton)

The entire HPR-crew at KTH and Uppsala University (For interesting, fun and crazy happenings in at HPR-annuals at Krusenberg, and at HUPO in Munich)

My Hedhammar and Max Käller (My two partners in thesis writing, thank you for discussions about references, figures and other things about writing a thesis nobody else cares about)
And of course, all of the remaining hard-working associates at the divisions of Proteomics, Gene Technology and Molecular Biotechnology division at the School of Biotechnology (You know who you are)

Moreover, The Knut and Alice Wallenberg foundation is gratefully acknowledged for financial support.

**The following people have contributed tremendously to my thesis:**
Prof. Mathias Uhlen, Dr. Fredrik Sterky, Dr. Henrik Wernerus, Dr. Soraya Djerbi, and Dr. Cristina Al-Khalili Szigyarto (For critical reading and fruitful discussions)

Håkan Gill (My redwine-loving, former next-door-neighbour. Thank you for bringing the third dimension to my thesis by providing me with awesome 3D pictures. You are truly a master at 3D!)

Assoc. Prof. Harry Brumer III (For corecting maj innglissh, and teaching me that N-terminals has nothing to do with biotechnology)

Cecilia Heyman at the library (For answering my many, many e-mails and providing me with publications that I couldn’t get hold of myself)

Furthermore, I would like to thank the following for providing me with material and insights: Johan Rockberg (the unpublished data section), Prof. Per-Åke Nygren (discussions about affinity reagents), Dr. Cristina Divne (protein structure pictures), Prof. Fredrik Pontén (TMA picture and the annotation of the staining), Dr. Peter Nilsson (PrEST array picture), Dr. Anja Persson (raw material for the PrEST statistics table), Marcus Johansson (For your helpful advice with InDesign), Dr. Afsin Ahmadian (discussions about the SNP section) and Dr. Margareta Ramström (MS spectra).

**My Friends:**
R&B: (It’s been, and still is, a BLAST to hang around with the crazy crew from R&B, and every event has had it’s fair share of relaxation, fun and crazyness. I was going to write extremeness, but maybe we’ll add that to the list after some years. Just to remained you: Stenskär, Gotland, Dalarna, Jönköping, Västerås, Florida and Sandhamn)
R&B is: André (The wine-knower who likes to shake his stuff to latino rhythms and sing Ronja Rumpdaughter songs), Ante (The old shrimp-farmer, who has a Matrix-inspired running-form, and burned his upper-lip once when drinking), Feffe (Knows his charter-destinations and luxury cars, he dances with one finger up i the air), Jake (He drives nothing but Alfa Romeo and his tough body withstand even several rounds of ammunition), Martin (Vice president in R&B. He is a newlywed marathån runner, who has actually no taste in music), Niclas (The designer/bonus-dad/carpenter, who like to hang out at Knife-south and likes to dance riverdance), Ola (The man with the broken arm, the proud owner of a pizza-racer, which he once forgot where it was), Patric (co-founder of R&B, A.K.A ‘The guy in cruel condition’, drives a Mercedes, lives the high-life and hangs out close to swampen), and Sparre (A.K.A. Goose and Spirre, who is still in the military service, that is, a TopGun instructor)
I’m looking forward to new events, R&B forever!
Kristina Tunkrans (the attorney and DJ-extravaganza. You have supported me tremendously over the last years!)

Daniel, zlow-roc Johansen (The only time I’ve seen him pissed off is during gym class when he wasn’t allowed to drink soda. Thanks for awesome Miami-trips, boat-trips, parties, T2-review and dinners)

Magnus von Bahr (The second sickest party dancer, thanks for fishing trips, dinners at Rörstrandsgatan and crazy party trips)

Tomas Eklund (A.K.A Tjomme, Lance Armstrong, thank you for the opportunity of having my party at your place. I’ll do my best to raise the roof!)

**My Family:**
The Djerbis, Taieb & Enikô, Samir & Emili (For for nice dinners and get-togethers)

Mounira & Martin (For weekends at Stenskär, pleasant dinners and for X-box/Singstar nights)

Jonas (My oldest brother, thank you for looking after me all these years, even if I didn’t like that when I was younger), Maria (For moral support), Nora & Hanna (For being who you are)

Fredrik (My older brother, thank you for your support the last years and for watching Adam when I really needed to work), Eva (For nice get-togethers at Stenskär and Åre, Good luck with your research), Hugo & Filippa (For bringing joy and fun)

Mom Eva and Dad Rolf (My two heroes! As much support I’ve had from you all my life is more than a son can ask for. Thank you so much for taking care of Adam when I wrote my thesis. Can you believe, even though I wasn’t the most promising student in school, I still managed to accomplish this. I owe a lot to you)

Adam (A.K.A. Mats Jr., Knas-fnas, Klanton, Fjanton. When you say: *"I love you on all planets, even on the sun’* it makes me warn inside. I hope I haven’t neglected you too much when writing my ‘blue book’.)

Sosso (My ideal-princess and soulmate. I sure hope our time together only has started and that we will have many years together in the future. Thank you for all your help with Adam, you are surely the best bonus-mom anyone can ask for in the world. I love you)
Populärvetenskaplig
Sammanfattning (In Swedish)


Under år 2003 startades ett samarbetsprojekt mellan Kungliga Tekniska Högskolan och Uppsala Universitet. I detta projekt kallat HPR (the Swedish Human Proteome Resource; se figur 7.1), kartläggs systematiskt de mänskliga proteinerna och deras uttryck med hjälp av vävnadsmikromatriser (se figur 3.3). Dessa matriser består av prover från 48 olika typer av vävnader och 20 cancertyper. Således kan proteiner studeras med avseende på var de är uttryckta.


Utöver det mänskliga genomten, har flera andra organismer avrassmassa kartlagts. Bland annat poppel, som är en modellorganism för träd. Det som skiljer träd från andra växter är deras förmåga att bilda ved. Cellulosa är en huvudkomponenten i ved och dessutom en viktig råvara i t.ex. pappersmassa- och textilindustrin. En bioinformatisk (se nedan) studie av jättepoppel (Populus trichocarpa; artikel IV), har tidigare okända cellulosasynasgener identifierades och antalet uppskattats till 18. Tidigare studier i växter har visat på ungefär tio stycken cellulosasynasgener.
Arbetet i denna doktorsavhandling har fokuserats på studier som baserats på offentligt tillgängliga genomsekvanser, vilka sedan bearbetats och analyserats med hjälp av datoringiska metoder. Denna metodik kallas bioinformatik och innefattar principer hämtade från matematik, statistik, datoring och biologi. De arbeten som beskrivs i denna doktorsavhandling innefattar utveckling av analysprogram och databasimplementeringar (artikel I, III, opublicerat data) samt bearbetning av DNA- och proteinsekvanser. De sekvanser som har analyserats härstammat från människa (artikel I-III) och jättepoppelträd (artikel IV). Dessa analyser har omfattat parvisa- (artikel I, III, opublicerat data) och multipla upplinjeringar samt fylogenetiska studier (artikel IV).
References


Appendices


