Insights into the Evolution of Protein Domains give rise to Improvements of Function Prediction

Dissertation zur Erlangung des

naturwissenschaftlichen Doktorgrades

der Bayrischen Julius-Maximilians-Universität Würzburg

vorgelegt von

Birgit Pils

aus Bad Soden im Taunus

Würzburg 2005

Eingereicht am: .....

Mitglieder der Promotionskommission:

Vorsitzender: .....

Gutachter: Prof. Dr. Jörg Schultz

Gutachter: Prof. Dr. Jürgen Kreft

Tag des Promotionskolloquiums: .....

Doktorurkunde ausgehändigt am: .....

ERKLÄRUNG

Hiermit erkläre ich ehrenwörtlich, daß ich die vorliegende Dissertation selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Die Dissertation wurde bisher weder in gleicher noch ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt.

Außer dem Diplom in Biochemie von der Universität Witten/Herdecke habe ich bisher keine weiteren akademischen Grade erworben oder versucht zu erwerben.

Würzburg, August 2005 Birgit Pils

Acknowledgements

Many individuals have supported me during the development of this thesis whom I would like to express my deepest gratitude.

First and foremost, I would like to thank my mentor Prof. Dr. Jörg Schultz for his inspiring and knowledgeable guidance throughout my research. Without his expertise in bioinformatics and his encouragement, this thesis would never have been completed in its present form and I gratefully acknowledge his support.

I am also thankful to Prof. Dr. Jürgen Kreft for his time and willingness to serve on my thesis committee.

It is a great pleasure to thank my collaborators, Dr. Richard Copley and Dr. Alexander Heyl, for their help and advice during my research.

I wish to thank Prof. Dr. Thomas Dandekar and all members of the Department of Bioinformatics at the University of Würzburg for their inspiration and motivation during the last two years and for the pleasant working environment. Especially, I thank Juilee, Julia, Torben and Stefan for being more than just colleagues and Karin, for being the greatest office mate I can imagine. I also would like to thank Esther and Karin for proof-reading this thesis.

Many special thanks are due to the members of the Vingron Department at the Max Planck Institute for Molecular Genetics, Berlin, during the time from May 2002 to October 2003 for their stimulating discussions and the friendly atmosphere. In particular, I would like to thank Steffi, Wasinee, Jochen, Tim, Claudio and Alexander for their after hours commitment.

I would like to thank Gudrun, Sven, Rupert and Holger for their support and friendship and thanks to all other friends who I have encountered during the course of this PhD.

Many thanks to the generous funding of the Japan Society for the Promotion of Science (JSPS) that allowed me to explore bioinformatic research in Japan. I would like to express my sincere gratitude to Prof. Dr. Hiroyuki Toh for accepting me in his group and for introducing me to the scientific and cultural way of Japanese life. My stay at the University of Kyoto has been an enriching and unforgettable experience and I owe a big “Thank you” to all members of the Toh group for providing a smooth and enjoyable stay.

Above all, I would like to thank my family for their unlimited support throughout my studies.

Table of Contents

1 General Introduction

Computer aided assignment of protein function has become indispensable with the beginning of the post-genomics era. Computational methods offer a fast alternative to tedious and expensive experimental studies and can easily deal with the vast amount of uncharacterised sequences generated by various genome projects. The field of genomics was born with the sequencing of the relatively small genomes of *Haemophilus influenza* (FLEISCHMANN *et al.* 1995), *Mycoplasma genitalium* (FRASER *et al.* 1995) and *Saccharomyces cerevisiae* (GOFFEAU *et al.* 1996). Due to improvements in sequencing techniques and bioinformatic tools in the late 90s the field received a big boost and the genetic code of several metazoa was rapidly released. The sequenced metazoa include the human (LANDER *et al.* 2001; VENTER *et al.* 2001) as well as the most important model organisms like mouse (WATERSTON *et al.* 2002), the fruitfly *Drosophila melanogaster* (ADAMS *et al.* 2000), the worm *Caenorhabditis elegans* (*C. ELEGANS* SEQUENCING CONSORTIUM 1998), and the puffer fish *Takifugu rubripes* (APARICIO *et al.* 2002). In addition to these metazoan genomes, many other eukaryotic genomes as well as several hundreds of prokaryotic genomes have been sequenced[1,2]. The Genomes OnLine Database[3] (GOLD, BERNAL *et al.* 2001) lists almost 1500 genome projects as of July 2005.

Although we have deciphered many genomes and gained knowledge on their composition we are far from understanding how their genes and protein content function at the cellular context, which has been an expected outcome of the sequencing projects ten years ago. With more genomes being sequenced, the number of uncharacterised sequences is growing exponentially. Until today the assignment of protein function to the numerous predicted protein sequences remains a great challenge. The passed decade has been marked by great efforts towards the improvement of protein annotation. Many automated pipelines have been developed, which work at a genome wide scale to predict protein function. These approaches are well suited for rapid processing of the large amount of uncharacterised sequences, but they are also limited in their specific description of protein function. For understanding the protein complement of an organism, detailed functional

[1] http://www.tigr.org/
[2] http://www.ncbi.nlm.nih.gov/Genomes/
[3] http://www.genomesonline.org

analysis is necessary and presents a research area of growing importance. In a way, the pace of functional protein annotation can be compared to the pace of sequencing the human genome. After teething troubles, the sequencing process reached a reasonable speed and lead to the rapid completion of a draft sequence. However, it took comparably long to finish the last few percent of the genome. Similarly, we now have many excellent methods to achieve a rough idea on a protein's function but the in-depth analysis of protein function will certainly require a longer period.

This chapter will introduce computational methods to predict a protein's function and will illuminate the evolution of proteins. Issues in the assignment of protein function are raised and methods to predict functional sites are introduced. Evolutionary processes represent an important factor in the development of new protein functions and understanding the evolution of proteins contributes to enlightening the function of the protein. Automatic approaches for function prediction are usually based on sequence information and bear advantages and disadvantages that will be discussed. As sequence analysis is best performed if the protein is divided into smaller functional units, protein domains and domain databases are described. Protein domains are central to this thesis and all analyses presented in this work are based on these functional units. Finally, this chapter will give an outline of this thesis.

1.1 Prediction of protein function

The annotation of protein function with computational methods has become a very important field to understand how proteins work together and form complex biological systems (for review see PONTING 2001; ROST *et al.* 2003; GABALDON and HUYNEN 2004; VALENCIA 2005). A first approach in the assignment of function is usually to transfer information from evolutionary close homologues that have been studied experimentally, because it is assumed that homologous sequences share similar functions (BORK and KOONIN 1998). The degree of sequence similarity sheds light on the level of conservation of the protein's properties. While highly similar sequences might even share the affinity for same substrates, more diverged sequences may only have the catalytic mechanism in common. Sequence information and functional descriptions are stored in primary databases, of which the three most important are GenBank at the National Center for

Biotechnology[4] (NCBI, BENSON *et al.* 2005), the EMBL Nucleotide Sequence Database[5] at the European Molecular Biology Laboratory (EMBL, KANZ *et al.* 2005) and the DNA Databank of Japan[6] (DDBJ, TATENO *et al.* 2005). These databases can be searched for similar sequences with tools like FASTA (PEARSON and LIPMAN 1988), BLAST (ALTSCHUL *et al.* 1990) or PSI-BLAST (ALTSCHUL *et al.* 1997). It is important to estimate the biological significance of the sequences identified by similarity searches with statistical tools. Sequence similarity is assessed as an alignment score based on the local alignment algorithm, gap penalties and a substitution matrix (VINGRON and WATERMAN 1994). These scores are used to calculate an expectation value (E-value) indicating the number of sequences with better or similar scores to occur in the database by chance. However, even low E-values cannot be blindly trusted and the sequence alignments have to be visually checked to confirm homology. The E-value is strongly dependent on the size of the database and can be very low for small sequence sets. In addition, the E-value does not consider non-random amino acid compositions and apparently significant hits with a high occurrence of a particular amino acid might be identified.

Several problems can arise during the annotation process and lead to the incorrect interpretation of the novel protein sequence (BORK and BAIROCH 1996; SMITH and ZHANG 1997; BORK and KOONIN 1998; DOERKS *et al.* 1998; ANDRADE *et al.* 1999). Annotation of protein sequences from genome data requires the reliable detection of gene structures and open reading frames by gene prediction algorithms. Incorrect gene prediction is followed by error-prone prediction of protein function. Errors in the gene prediction process arise from multiple start and stop codons within the open reading frame, which result in truncation or elongation of the protein sequence. Especially in intron-rich eukaryotes, the gene structure is often ambiguous and leads to the translation of intronic sequences or the omittance of exons. In addition, the existence of several alternative splice variants complicates the correct prediction of the protein product.

Another source of problems can arise from the transfer of function from homologous sequences. Very often the most similar sequence identified in the databases matches only part of the query sequence. Then only this part of the sequence can be annotated with a function and the remaining part of the sequence, for which no

[4] http://www.ncbi.nlm.nih.gov/
[5] http://www.ebi.ac.uk/embl/
[6] http://www.ddbj.nig.ac.jp/

homologous sequence was detected in the database search, is ignored (figure 1). Conversely, the query sequence can match only part of the database sequence. This can lead to incorrect assignment if the function is transferred from the unmatched region of the protein sequence. But even if the complete sequence can be aligned to a database sequence with high similarity, the function can be different due to few mutations at key residues.

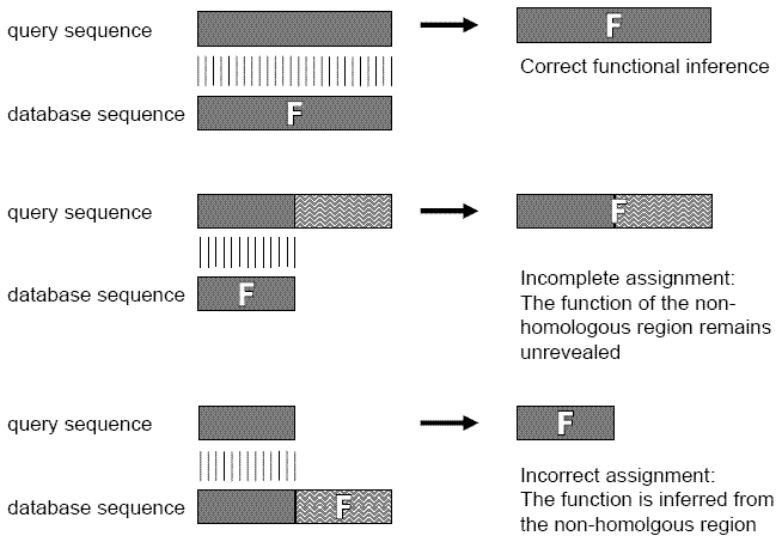


Figure 1: Problems arising from functional inference based on sequence homology (modified from PONTING *et al.* 2000)

The problem of partial homology can easily be avoided by analysing protein domains separately. Domains present evolutionary, structural and functional independent units of proteins. They are also called the "lego set" of nature (DAS and SMITH 2000). Each domain has a unique function such as catalysing a chemical reaction or binding a specific ligand. Proteins are very variable in their composition and arrangement of domains. The function of the whole protein is composed of the individual functions of the domains. Since the domain offers many advantages in the prediction of protein function it will be explicitly discussed in the following (see chapter 1.5).

One has to keep in mind that the functional inference from homology is based on a very small dataset of experimentally annotated proteins, from which the function is extrapolated on homologous sequences. The amount of experimentally annotated protein

sequences accounts for probably less than 5% of the total number of known sequences (VALENCIA 2005). Consequently, the outcome of this approach is very often misleading and erroneous. In order to reduce the risk of falsely annotated sequences, one could limit the transfer of function to orthologous sequences, because it is assumed that function is conserved from the last common ancestor. Orthologous sequences are related by a speciation event, whereas paralogous sequences arose from intra-genome duplications. Following the intra-genomic duplication process, the redundant gene is freed from selective pressure and can acquire new functions or change its expression pattern or specificity. It is often observed that paralogous genes accumulate mutations to change their substrate specificity like enzymatic domains modifying similar but different molecules. For example the more than 500 different protein kinases in the human genome all catalyse the transfer of a phosphate group from a donor to a receiver domain, but differences exist in the kind of cofactor and the protein, which is being modified (LEONARD *et al.* 1998).

Unfortunately, the identification of orthologous genes is not straightforward. Often one gene from one organisms matches several other genes from another organism. This is especially a problem in multi-cellular organisms with large genomes. Here, multiple gene duplications have lead to many-to-many relationships among orthologous genes, demanding a closer classification. Different relationships arise from different timings of the duplication and speciation event. The speciation event can be followed by intra-genomic duplication leading to orthologous genes. Alternatively, the duplication event can precede the speciation event. Sonnhammer has suggested the term inparalogous for genes that have recently undergone the duplication process and outparalogous for ancestral duplications (figure 2, SONNHAMMER and KOONIN 2002). Orthologous genes that have arisen from ancestral duplications followed by speciation (outparalogous genes) are assumed to be more conserved, thus, they should be better suited for functional prediction of protein function.

A good example of the adaptation of inparalogous genes present the lactate dehydrogenase from *Trichomonas vaginalis*, which is closer related to malate dehydrogenase, than to any other dehydrogenase within the genome. This indicates that these genes arose from a recent duplication event, followed by acquirement of new substrate specificity of one gene copy (WU *et al.* 1999). To clearly differentiate between orthologous genes and paralogous genes other information like the genomic context can be

used (BHARATHAN *et al.* 1999). The conservation of gene order around homologous genes is an additional evidence for a common ancestry.

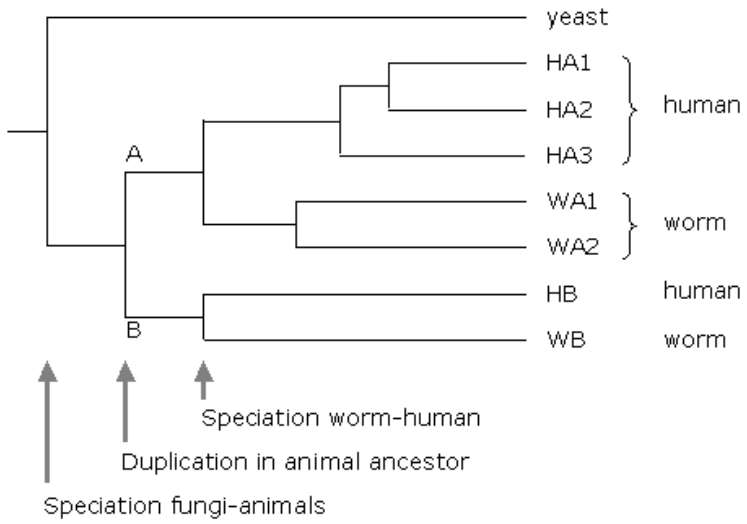


Figure 2: Evolutionary relationships among inparalogous and outparalogous genes. The ancient gene is inherited in yeast, worm and human. An early duplication in the animal lineage was followed by a second round of duplication after the divergence of human and worm. The more recent duplication occurred independently in the human and worm lineages. In this scenario the human genes HA1, HA2 and HA3 are all co-orthologous to the WA1 and WA2 worm genes. According to the terminology introduced by Sonnhammer, the human HA1, HA2 and HA3 genes are inparalogues when comparing human and worm, whereas the HA1/HA2/HA3 and HB genes are outparalogues (modified from (SONNHAMMER and KOONIN 2002))

The homology-based approach to predict a protein's function described above compares protein sequences with the aim to identify sequences with common evolutionary ancestry and thus, very likely common molecular function. Very often, it is not possible to identify any similar sequences. Usually, 30% to 50% of all predicted gene products of a completely sequenced genome do not match any sequences in the databases or only with very small sequence identity, which does not allow the reliable prediction of the gene product's function (YAKUNIN *et al.* 2004). Thus, additional information has to be taken into account, such as temporal or spatial regulation of the protein, interaction partners, gene neighbourhood, phenotype of the gene-knockout or structural information. Especially three-dimensional protein structures from large-scale structural genomics projects have been very beneficial for the investigation of the molecular function of proteins. Similar to the international organized genome projects, several initiatives pursue the aim to determine

all structural folds of nature's protein repertoire. For example the RIKEN Structural Genomics/Proteomics Initiative[7] (RSGI) in Yokohama, Japan, or the Division of Structural and Functional Genomics[8] of the Wellcome Trust Centre for Human Genetics in Oxford, England, have begun with high-throughput analyses to determine protein 3D structures. In order to combine the international efforts in structural genomics, the Brookhaven Protein Data Bank[9], (PDB, BERMAN *et al.* 2000) serves as depository for experimentally determined 3D structures. Comparison of protein structures can be used to infer protein function in an analogous manner as comparison of protein sequences (THORNTON *et al.* 2000). Global similarities can reveal the biological task of the protein; just as well as local similarities can identify motifs of known biological function. Examples for structural motifs are the helix-turn-helix motif (PABO and SAUER 1992), which is frequently found in DNA-binding proteins, the calcium binding EF hand (YAP *et al.* 1999) or the catalytic triad (BLOW *et al.* 1969; WRIGHT *et al.* 1969), located in the catalytic center of many enzymes. Even if no similarities to other known structures are detected, the 3D structure can still shed light on the functional properties of the protein. Since catalytic sites of enzymes are usually located in clefts on the protein surfaces, these sites can be identified by analysing the 3D structure. It has been shown that over 70% of enzyme's catalytic sites correspond to the largest cleft on the protein surface (LASKOWSKI *et al.* 1996). Structural information is also very helpful in the prediction of functional sites in the protein (see chapter 1.3).

Other successful methods for protein function prediction use genomic information assuming that the genes of functionally interacting proteins tend to be associated with each other on the genome level (for review see GALPERIN and KOONIN 2000; HUYNEN and SNEL 2000; MARCOTTE 2000; VALENCIA and PAZOS 2002). Thus, the information revealed regards a cellular or higher order function such as in which pathway the protein is involved. These methods are gaining in attraction with more genomes being completely sequenced. Several types of information can be exploited from the genomes. First, gene fusion can indicate a related function (ENRIGHT *et al.* 1999; MARCOTTE *et al.* 1999). Here, homologues are searched, which are fused in one genome, but occur as individual genes in another genome. The evidence for a related function is especially strong if the genes are orthologues. Second, genes conserved as pairs or clusters within a genome tend to

[7] http://www.rsgi.riken.go.jp
[8] http://www.strubi.ox.ac.uk
[9] http://www.rcsb.org/pdb/

physically interact (MARCOTTE *et al.* 1999). The most prominent example of functional related gene clusters in prokaryotic genomes are operons, which embrace multiple genes under a single transcriptional regulator, whereas the overall conservation of gene order is relatively low if compared *Escherichia coli*, *Haemophilus influenzae*, and *Mycoplasma genitalium* (KOLSTO 1997) or nine bacterial and archeal genomes (DANDEKAR *et al.* 1998). But the genes that are conserved as part of a cluster often encode subunits of larger protein complexes (MUSHEGIAN and KOONIN 1996). And third, co-occurrence of genes across genomes (phylogenetic profiles) can indicate a functional interaction (HUYNEN and BORK 1998; PELLEGRINI *et al.* 1999).

The database STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a comprehensive collection of all kinds of direct and indirect protein asscociations (VON MERING *et al.* 2005). The information is not only retrieved from genomic context analysis but also from high throughput experimental data and literature mining. The size of the protein associations exceeds manifold information stored in primary databases and the information can be easily transferred onto orthologous genes from species not contained in the STRING database.

It is important to note that many proteins are not only multi-domain in character but also multi-functional. Even one structural fold can carry out more than one function. And even one individual protein can carry out more than one function controlled by environmental factors. A special class of these multi-functional proteins has been termed moonlighting proteins. They can change their function depending on the time, pH value or location (JEFFERY 2003). The phosphoglucose isomerase catalyzes glucose-6-phosphate to fructose-6-phosphate if found intracellular. Its extracellular counterpart, neuroleukin functions as cytokine and promotes the survival of neurons (SUN *et al.* 1999). Similarly tryptophanyl-tRNA synthetase (TrpRS) is responsible for loading the tryptophane specific tRNA inside the cell, but in the extracellular compartment it inhibits blood vessel growth as angiostatic cytokine (YANG *et al.* 2004). The performance of more than one function by a single protein suggests that the protein fold has two or more functional sites. With large-scale approaches to study protein function, the moonlighting proteins would not be detected with their full abilities. Thus, we will have to face detailed small-scale approaches in the future to finish our encyclopedia of proteins.

1.2 Prediction of Functional Sites

Even though the cellular function of a protein family is known, it is desirable to identify the individual functional sites of a protein to reveal its molecular mechanism. In sequence analysis the multiple sequence alignment of the protein family can deliver first hints. Completely conserved positions often point out functional important sites. For example, catalytic positions can easily be spotted in the alignment assuming that these sites are absolutely conserved in the whole family. Functional important sites, which are restricted to a subclass, like substrate specific positions, are less conserved in the full alignment and are harder to detect. Similarly, positions involved in the interaction with other macromolecules are able to co-evolve with the interaction partner and will not be conserved in the alignment.

In order to detect functional important sites, many methods have been proposed in the last years that aim for finding subclass specific positions or so called tree-determinant positions (for review see LICHTARGE and SOWA 2002; WHISSTOCK and LESK 2003). First approaches studied physico-chemical properties of amino acids in a multiple sequence alignment to identify interesting subgroup specific positions (LIVINGSTONE and BARTON 1993). Casari and colleagues described the sequence-space method. They project proteins and amino acid residues as vectors in multi-dimensional space and use principal component analysis to identify clusters of residues that are conserved within a subfamily (CASARI *et al.* 1995). A method that has gained a lot of popularity and has been modified many times is the evolutionary trace (ET) method. Originally introduced by Lichtarge and colleagues, this method correlates global and local patterns of amino acid conservation with the classification into subgroups provided by a phylogenetic tree and searches for clusters of conserved subclass-specific positions on a 3D structure representing the protein family (LICHTARGE *et al.* 1996). Further improvements of the ET method include weighting of sequences depending on their similarity to other family members (weighted evolutionary tracing, LANDGRAF *et al.* 1999), searching for clusters instead of single positions by including surrounding residues revealed from the structure in the calculation of a regional conservation score, using evolutionary data to reduce the effects of taxonomically uneven sampling of sequences (ARMON *et al.* 2001) and considering gaps resulting from insert and deletion events (MADABUSHI *et al.* 2002).

1.3 Evolution of proteins

As most domains are found in all three kingdoms of life, they must have a very ancient origin (PONTING *et al.* 1999). The strong conservation over evolutionary time shows that these domains function in fundamental cellular processes such as translation or central metabolic pathways. For example, enzymatic domains adopting TIM barrel or Rossmann-like folds must have been generated more than three billion years ago in the common ancestor of eukarya, bacteria and archea. Also non-enzymatic domains like the PDZ domain, von Willebrand factor A or cystathionine β-synthase are found in all three forms of life (PONTING 1997; PONTING *et al.* 1999). New inventions of domains occur mainly along with speciation events. Chromatin associated domains or domains functioning in ubiquitin mediated proteolysis or apoptosis have arisen in the eukaryotic lineage and are not found in prokaryotes. Many extracellular domains are only found in metazoa and are absent in all other species. The invention of new extracellular domains has occurred concurrently with the evolution of multi-cellular organisms, for which the communication between the individual cells was absolutely required. Although many domains seem to be specific for a certain phylogenetic lineage at first sight, a common evolutionary origin with domains from other lineages can be inferred with better algorithms to compare structures and sequence similarities. The cytokine IL-1α domain, functioning in immune response was thought to have been generated in chordates. But recently, a sequence and structure comparison with more sensitive tools revealed that this cytokine arose from a protein precursor homologous to invertebrate growth factors and actin-binding proteins found in slime-mold and fungi (PONTING and RUSSELL 2000). Similar, many isolated domain families can be grouped to families that have arisen from one common ancestor (BECKMANN *et al.* 1998; SHAPIRO and SCHERER 1998; TODD *et al.* 2001).

New domains arise mainly through gene duplications or genome duplications. Large-scale duplications like genome duplications, chromosomal duplications or sub-chromosomal duplications are the most efficient way to rapidly increase the amount of raw material for the development of new genes (WOLFE 2001). It is assumed that two rounds of genome duplications followed by loss of some genes and functional differentiation occurred early in the vertebrate lineage (SIDOW 1996), but this '2R' hypothesis is controversially discussed (DURAND 2003). However, any of these duplication processes

lead to two, originally equal copies of which one is freed from selective pressure. The consequence is the more rapid accumulation of mutations, which might lead to the development of a new function. If the new acquired function is advantageous for the organism the gene will come under selective pressure again to maintain the beneficial gain ("neofunctionalization"). On the other hand, the accumulation of mutations can result in a non-functional pseudogene (LYNCH and CONERY 2000).

Duplication events are a very important mechanism to increase the complexity of the protein complement in the cell. Families like the olfactory receptors have arisen from numerous duplication events followed by diversification (GLUSMAN *et al.* 2001; ZHANG and FIRESTEIN 2002). The olfactory receptors present the largest gene family in the mammalian lineage with more than 1000 genes (GLUSMAN *et al.* 2001; ZHANG and FIRESTEIN 2002). However, great differences in the selective pressure exist between mammalian genomes. The mutation rate of human olfactory genes is much higher than in other sequenced mammals and has lead to the inactivation of about 60%, whereas the mouse genome counts only 20% of pseudogenes among the olfactory receptors (ROUQUIER *et al.* 1998; ZHANG and FIRESTEIN 2002).

Besides duplication, the protein complement of the cell is shaped by other genetic events such as unequal recombination, circular permutation or exon shuffling. These mechanisms can lead to multi-domain proteins or the exchange of domains in favour of a new arrangement. The hypothesis that domains are encoded by individual exons and that proteins with new domain combinations are linked by recombining different exons (GILBERT 1978) could only be confirmed for extracellular proteins (PATTHY 1987). In addition, domains can also be received from the environment. Lateral transfer (also referred to as horizontal transfer) is a mechanism by which an organism acquires genetic information from another organism (PONTING *et al.* 1999). This mechanism is prevalently found in bacteria. Signalling domains have been frequently transferred from eukaryotes to bacteria. Horizontal gene transfer in metazoa is less likely, because it would require a transfer into the germline. Similar to horizontal gene transfer, mobile elements can be transferred by infection with viruses or plasmids.

The genetic events mentioned above explain how homologous domains could have evolved from a common ancestor domain and also how multi-domain proteins might have arisen. Very little is known about how domains evolved originally. It is assumed that there

is a limited number of different folds (CHOTHIA 1991; ORENGO *et al.* 1994). Recently it has been suggested that the modern proteins we see today evolved from smaller structural units made up from short polypeptides (antecedent domains segments) that could fold and perform a beneficial function (LUPAS *et al.* 2001). The theory of an ancient peptide world is supported by two findings: First, many proteins contain copies of a homologous repeat. Folds like the β-trefoil fold, β-propellers or triple β-spirals contain several homologous regions that probably arose from intrachain duplication. Second, short highly similar motifs are found in different, non-homologous proteins, as for example, the helix-hairpin-helix (HhH) motif or the Asp-box, which is found in at least eight distinct protein families (COPLEY *et al.* 2001). The ancient peptide world might have produced homo- or hetero-multimers from short domain segments and longer proteins might have arisen by fusion of the substructures.

The protein complement of the cell is not only influenced by the birth of new domains. Other genetic events can lead to the death of domains. Without any selective pressure, the domain will accumulate mutation until the conversion into a pseudogene, or deletion can cause the disappearance of the domain.

1.4 Evolutionary forces on proteins

Once a functional protein is established, it changes by random drift and natural selection. The most frequent changes are point mutations, inserts or deletions, which can change the protein tremendously, often far beyond the point where any sequence similarity is still detectable. The fate of the protein is dependent on the way these evolutionary changes affect the function of the protein. Advantageous changes become fixed under positive selection, deleterious ones are eliminated and neutral changes lead to polymorphisms in the population.

The fate of protein domains is determined by mutations in the genomes of individuals. After a substitution is manifested in the genome, the occurrence of the substitution in the population accumulates at a rate that is determined by random genetic drift and natural selection. However, an advantageous change in the genome, which is subject to positive selection, will spread faster in the population and a deleterious change can be fixed in the population but will require more generations. Positive or negative

selection pressures can be measured by comparing the rates of synonymous and non-synonymous substitution rates. Due to the degenerative amino acid code, substitutions can be classified as synonymous, when the coding amino acid is not changed by a mutation in the nucleotide sequence, or non-synonymous when the nucleotide substitution changes the amino acid. The amino acid level is an important indicator of natural selection because all important functions are performed by proteins rather than by nucleotide sequences (YANG 2001).

The comparison of synonymous and non-synonymous substitution rates allows studying the effect of positive or negative selective pressures on genes. It is well suited to investigate the fate of duplicated genes and it even allows estimating at what time in evolution the duplication must have occurred. The calculation of the number of non-synonymous substitutions per non-synonymous site (K_a) and the number of synonymous substitutions per synonymous site (K_s) is based on the alignment of homologous sequences at the codon-level. The ratio of K_a/K_s, often referred to as ω, reflects the type of selection that has been acting on the genes since the last common ancestor. Several methods for the estimation of the K_a/K_s ratio have been developed (NEI and GOJOBORI 1986; ZHANG *et al.* 1998; YANG and BIELAWSKI 2000). For most genes the ration of K_a/K_s is relatively small, below 0.15, because synonymous changes occur much more frequently than non-synonymous. This is probably caused by the rather deleterious nature of non-synonymous mutations. Without any selective pressure, it is assumed that mutations of both kinds are equally likely. Then the K_a/K_s ratio would be ~1. Interestingly, some genes exhibit a K_a/K_s ratio that is significantly higher than 0.15 and often closer to 1. These genes have been subject to adaptation and have accumulated amino acid changes at a higher rate.

In contrast to estimate evolutionary rates for complete genes, it is also possible to detect single amino acid sites that are under positive selection within a set of homologous sequences. This is of great interest for the detection of functional sites within protein domains. Methods are available that estimate K_a and K_s at single codon sites (NIELSEN and YANG 1998; SUZUKI and GOJOBORI 1999). The condition for a reliable estimation is that the genetic distance between the investigated sequences is small, yet large enough to observe substitutions. With increasing genetic distance the rate of nucleotide substitutions also increases and the number of true substitutions might be underestimated. The effect of too many substitutions for a signal observation is called substitution saturation and results

from substitutions that can revert back to the original nucleotide, substitutions that occur several times at the same nucleotide (multiple hits) or from the same substitution in different lineages (parallel substitutions). Checking aligned nucleotide sequences for saturation is therefore an important step before drawing conclusions on their evolutionary history. If the nucleotide sequences seem to be diverged beyond the point of substitution saturation it is recommended to analyse protein sequences instead. Protein distances and evolutionary site rates can be estimated with a maximum likelihood approach implemented in TREE-PUZZLE (STRIMMER and VON HAESELER 1996).

Positions identified by studying evolutionary site rates correspond to regions of structural or functional constraints. Structural constraints arise from folding or packing properties of the protein, whereas functional constraints are based on catalytic or interacting abilities. It has been shown that the regions under strongest constraints are identical with the functional regions of a protein (SIMON *et al.* 2002). Doubts about the type of constraints can easily be removed by including structural data in the analysis. In fact, many methods use this advantage and map evolutionary site rates onto the 3D structures. As side-effect, clusters subject to stronger constraints are identified, which represent putative functional regions.

1.5 Domains as functional units of proteins

The most useful feature of proteins in sequence analysis is, however, the protein domain. Protein domains have first been mentioned in reports on the 3D structure of lysozyme (BLAKE *et al.* 1965) and ribonuclease (KARTHA *et al.* 1967). They are defined as compact structural units of proteins and also form functional units. Many methods exist that define domains based on geometric measures of the compactness within the 3D structure (ISLAM *et al.* 1995; SIDDIQUI and BARTON 1995; SOWDHAMINI and BLUNDELL 1995; SWINDELLS 1995). As domains are independent elements and are able to exist apart from the rest of the protein sequence, it is also possible to define domains by sequence comparison and identify homologous domains that are present in different molecular context (figure 3).

Domains usually consist of amino acid sequences of 20 to 400 residues in length. The fold can be stabilized by metal ions like zinc or disulfide bridges. Some domains are

not stable in isolation but can form stable globular structures as dimers or multimers. The β-propeller for example consists of 6 to 8 WD40 domains. The repetition of small domains can give rise to very large proteins like the giant muscle protein titin (PFUHL and PASTORE 1995). As mobile elements, domains are often found in combination with various other domains in different proteins. For example, the Src-homology 2 (SH2) domain or pleckstrin homology (PH) domain appear in many proteins involved in signalling processes and can be combined with enzymatic domains like kinases, phosphatases or other ligand specific domains (figure 3). Especially in Eukaryota, which have a larger number of longer proteins than Prokaryota, many proteins are built from domains in a modular architecture (LANDER *et al.* 2001). It has been estimated that two thirds of prokaryotic proteins and 80% of all eukaryotic proteins are multi-domain in character (GERSTEIN 1998; TEICHMANN *et al.* 1998). Knowledge of the domain structure of the protein is an essential step in studying the function of the protein, as the overall function of the protein is the sum of the individual functions of all domains. Ignoring the modular character of the protein can even lead to falsely annotated proteins (GALPERIN and KOONIN 1998).

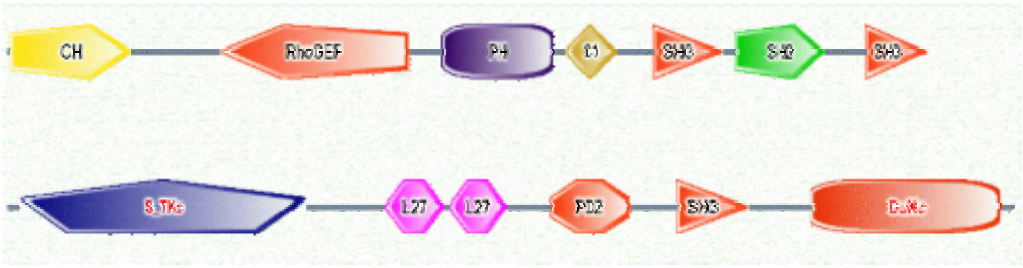


Figure 3: Domain organisation of SH3 containing proteins. The figure shows that protein domains can occur in different molecular context. The bubble representation show the Mus musculus GTP exchange factor VAV-3 (Q9ROC8) and the Mus musculus peripheral plasma membrane protein CASK (ENSMUSP00000033321), respectively.

In the last years, the discovery of new domains has been accelerated by the enormous effort of the sequencing projects. The gained data on protein domains and related data such as structural information and literature had to be collected in databases. Domain databases have become a very important step in protein analysis and are described in detail below. Now, it seems likely that most of the common domains have already been identified. Chothia estimated the number of different folds to be around 1000 only

(CHOTHIA 1992). This explains the reduced rate of new domain discoveries in the last years.

1.6 Protein Domain Databases

The modular character of proteins demanded the organization of information on protein domains in databases (for review on protein family databases see REDFERN *et al.* 2005). Even though the domain refers to a structural unit, it is more convenient to annotate proteins at the sequence level so that domains are usually represented by multiple sequence alignments in the databases. One of the best curated databases for protein domains is SMART (simple modular architecture research tool, SCHULTZ *et al.* 1998), which has been used for all analyses presented in this thesis. Originally, SMART focused on eukaryotic signalling domains, as these were underrepresented in other domain databases. Today, SMART contains a wider spectrum of protein domains from all kingdoms of life.

For each domain found in SMART a set of representative homologous sequences has been gathered. These sequences are first automatically aligned and then the quality of the alignment is improved by manual editing. The family alignments are available from the SMART website. SMART also offers an online tool for the identification of domains in protein sequences. This search tool is based on Hidden Markov Models (HMMs) and aligns the query sequence against a library of HMMs representing the domain families. HMMs are a very sophisticated way to detect homologous sequences in a database or to generate a multiple sequence alignment and they are often used in sequence analysis (figure 4, BALDI *et al.* 1994; KROGH *et al.* 1994; EDDY 1996). As HMMs are very sensitive and can even identify distant homologues (SCHAFFER *et al.* 1999) they are superior to other sequence similarity tools such as BLAST (ALTSCHUL *et al.* 1990) or PSI-BLAST (ALTSCHUL *et al.* 1997). The success of the HMM is dependent on the selection of protein sequences for the alignment and training of the model and increases with a better coverage of the family.

Classical HMMs are composed of three different states corresponding to the behaviour of sequences in a multiple sequence alignment: A match state describes the probability of the amino acid found at this position, an insert state allows the insertion of an amino acid and a delete state describes a gap in the alignment. For each state, the HMM

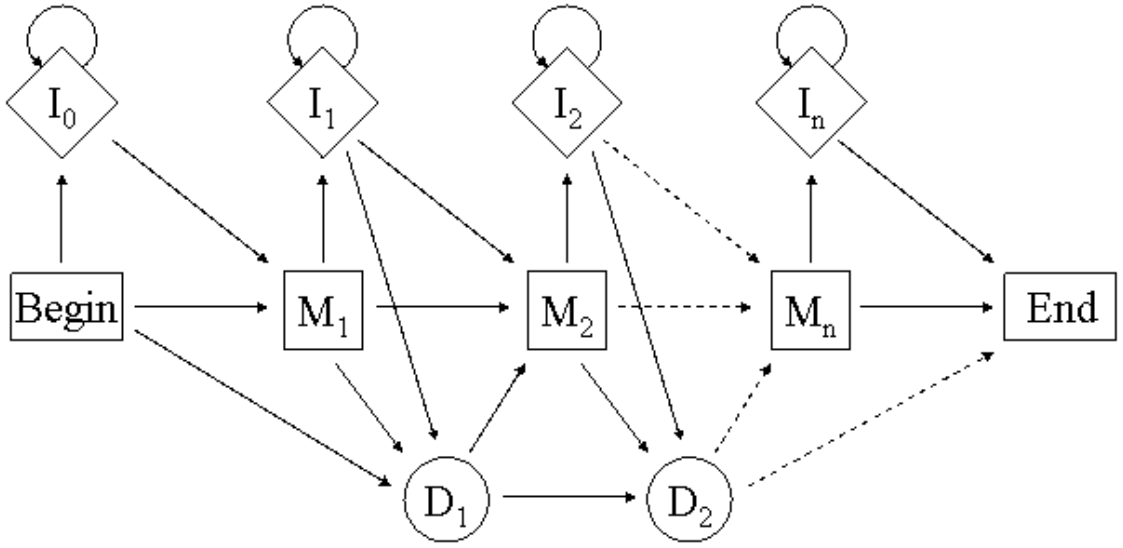


Figure 4: The hidden Markov Model describes a multiple sequence alignment by a combination of different states and transition probabilities indicated by arrows. The columns of an alignment correspond either to match states (rectangles) with associated amino acid distributions, insert states (diamonds) or delete states (circles) representing gaps in the alignment.

includes a probability for each possible transition to the next state of the HMM and in case of match states, the HMM includes an emission probability underlying the amino acid distribution at this alignment position. The undisputable advantages of HMMs and the availabilty of user-friendly software packages like HMMER[10] (EDDY 1998) or SAM (KROGH *et al.* 1994; HUGHEY and KROGH 1996) have resulted in their rapid spread among domain databases.

The Pfam database (Protein families database of alignments and HMMs) includes protein families and domains, and offers a greater coverage of domains over all kingdoms of life (BATEMAN *et al.* 2002). In contrast to SMART and Pfam, which are manually curated domain databases, ProDom is based on an automatic approach to detect domains in protein sequences from the SWISS-PROT and TrEMBL databases (BOECKMANN *et al.* 2003). Other useful resources are meta-databases, such as InterPro and CDD (conserved domain database, MARCHLER-BAUER *et al.* 2005), as they allow the simultaneous search of several primary domain and motif databases. At present, InterPro includes information from PROSITE, PFAM, PRINTS, ProDom, SMART, TIGRFAMs, PANTHER, PIRSF, Gene3D and Superfamily (see table 1 for references). The CDD database is mainly based on

[10] http://hmmer.wustl.edu/

information from SMART, Pfam and COGs (Clusters of ortholgous groups of proteins, TATUSOV *et al.* 2000) but also provides in-house models for domain classification (MARCHLER-BAUER *et al.* 2005).

In addition to these databases, which are built on sequence similarities, other databases use structural information to classify proteins. This is especially advantageous comparing diverged proteins, because the structure is changing more slowly than the sequence. Even when two sequences have changed far beyond the point where sequence similarity is still detectable, a common evolutionary origin can often be inferred by comparing there structural folds. The structural classification uses experimentally determined structures from the Brookhaven Protein Data Bank (PDB). The SCOP (Structural Classification Of Proteins) database (MURZIN *et al.* 1995; BRENNER *et al.* 1996) groups proteins into hierarchical levels of families, superfamilies, folds and a class at the highest level of the hierarchy, which is based on the arrangement of secondary structure elements. These hierarchical levels reflect structural and evolutionary relationships of the proteins. The CATH (Classification by Class, Architecture, Topology, and Homology) database (ORENGO *et al.* 1997) is based on a similar classification of protein structures.

Table 1: Databases mentioned in this chapter

Name	Description	Reference
BLOCKS	ungapped blocks in families defined by the Prosite catalog	http://blocks.fhcrc.org/ (HENIKOFF and HENIKOFF 1996)
CATH	Hierarchical classification of protein structures	http://www.biochem.ucl.ac.uk/bsm/cath/cath.html (PEARL *et al.* 2005)
Gene3D	Protein families of completely sequenced genomes, supplement of CATH database	http://www.biochem.ucl.ac.uk/bsm/cath/Gene3D/ (BUCHAN *et al.* 2003)
INTERPRO	meta database; allows to scan several domain and motif databases simultaneously	http://www.ebi.ac.uk/interpro (MULDER *et al.* 2005)
Panther	HMM based database on protein families and subfamilies modelled by functional divergence	https://panther.appliedbiosystems.com (MI *et al.* 2005)
Pfam	profiles derived from alignment of protein families, each one composed of similar sequences and analyzed by hidden markov models	http://www.sanger.ac.uk/Pfam (BATEMAN *et al.* 2004)
PIRSF	family and superfamily classification based on sequence alignment	http://pir.georgetown.edu/pirsf/ (WU *et al.* 2004)
PRINTS	protein fingerprints or sets of unweighted sequence motifs from aligned sequences families	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html (ATTWOOD *et al.* 1999)
ProDom	groups of sequence segments or domains from similar sequences found in SwissProt database by BLASTP algorithm; aligned by multiple sequence alignment	http://protein.toulouse.inra.fr/prodom.html (BRU *et al.* 2005)
Prosite	groups of proteins of similar biochemical function on basis of amino acid patterns	http://www.expasy.ch/prosite (HULO *et al.* 2004)
SCOP	Structural Classification of Proteins	http://scop.mrc-lmb.cam.ac.uk/scop/ (ANDREEVA *et al.* 2004)
SMART	modules described by multiple sequence alignments from family members identified by PSI-BLAST and analyzed by HMMs	http://smart.embl-heidelberg.de (LETUNIC *et al.* 2004)
Superfamily	HMM library of proteins of known structures based on SCOP	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/ (GOUGH *et al.* 2001)
Tigr fams	HMM library of protein families	http://www.tigr.org/TIGRFAMs/index.shtml (HAFT *et al.* 2003)

1.7 Project Outline

Goal of this project was to get insights into the function of protein domains and improve the annotation of proteins by computational tools. The field of protein annotation achieved obvious improvements with the analysis of independent domains in the last years. However, more and more examples of protein families with differing functions at the domain level are reported in the literature lately. Several domains have been subtyped using sequence information combined with structural data. The Ras association domain (RA, KALHAMMER *et al.* 1997), Src homology 2 domain (SH2, KIMBER *et al.* 2000) or Pleckstrin homology domain (PH, ISAKOFF *et al.* 1998) are only a few examples. This shows a trend moving from large-scale analyses towards more detailed small-scale analyses to describe subfamilies that share specific functions.

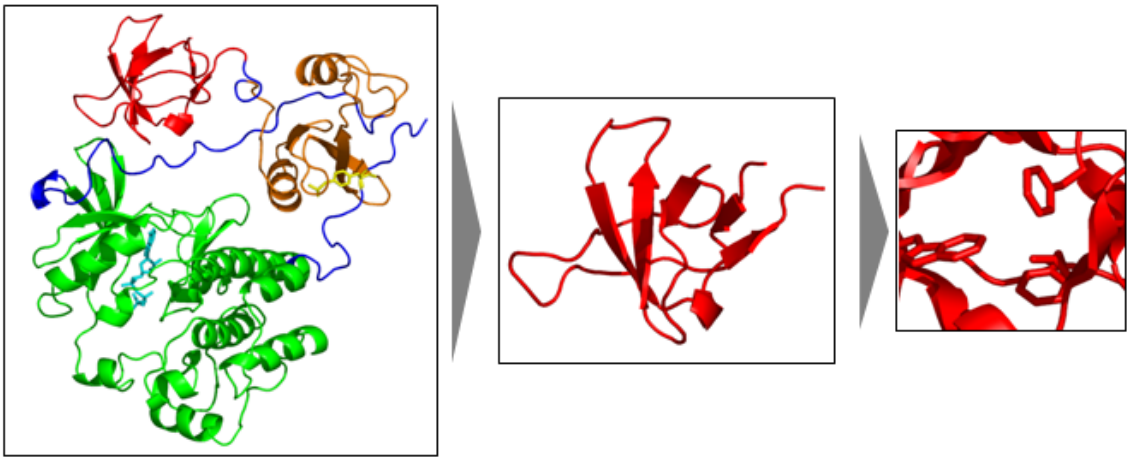


Figure 5: Zooming in: Protein function annotation has moved from the protein level via domains to the amino acid level. The figure on the left presents the Src tyrosine kinase, a signalling enzyme containing a SH3 (red), SH2 (orange) and kinase (green) domain. The figure in the middle focuses on the SH3 domain and the figure on the right shows a close-up of the SH3 domain's center.

In some cases, functional differences of protein domains are even observed in subfamilies when protein domains are classified based on sequence and structure information. Enzymatic domains can be converted into catalytically inactive domains by a single amino acid substitution if the substitution involves a catalytically essential residue. For example, protein tyrosine phosphatases usually function as antagonists of growth stimuli and can suppress tumor growth. However, a phosphatase with an opposite behaviour has been identified. This phosphatase carries a substitution at a catalytic position

and is able to transform cells (CUI *et al.* 1998). These functional differences can only be detected at the amino acid level (figure 5). The huge difference in the cellular behaviour demonstrates that such approaches are very important to shed light on the complex protein interplay in cells and whole organisms. Throughout this thesis, special emphasize was on the analysis of functional sites, including catalytic and ligand-binding sites as positions of great importance in the interaction network and cellular pathways. Wherever possible when using protein domains in the various analyses, the domain representing family alignments were retrieved from the SMART database, as the currently best maintained domain database containing manually improved alignments of highest quality (MARCHLER-BAUER *et al.* 2002).

Chapter 2 focuses on the assignment of catalytic functions to a prior little understood protein. The N-Acetyl-β-D-Glucosaminidase (GlcNAcase) is an important enzyme in signal transduction pathways and works as a molecular switch similar to kinases and phosphatases by transient modification of signalling proteins with O-linked N-acetylglucosamine (O-GlcNAc). The enzyme was described to comprise two different catalytic activities, a hyaluronidase and glucosaminidase activity. It was further assumed that both activities are carried out by the same catalytic center. The aim of this analysis was to map the catalytic activities to protein regions and identify putative catalytic residues. Simultaneously, the evolution of the glucosaminidase family was studied.

Continuing the analysis of signalling domains, an extracellular domain of a sensory receptor from plants was studied in chapter 3. Here, the problem of lacking a three-dimensional structure had to be overcome to identify ligand-binding positions specific for plant sequences. The successful prediction of functional sites was only possible because this domain can bind very different ligands in plants and bacteria, and this preference is also reflected in the phylogenetic tree of the family.

In order to study properties of functional sites and gain knowledge for the prediction of functional residues, the protein tyrosine phosphatase family was used as case study (chapter 4). Initially the data set was restricted to catalytic sites only and amino acid substitutions at important catalytic positions were investigated. The catalytic center is very sensitive to substitutions and usually underlies strong selective pressure. However, several domains were identified with substitutions at catalytic positions that presumably turn the enzymatic domain into a catalytically inactive one. Interestingly, these inactive

phosphatase domains are very often combined with an adjacent active phosphatase domain and seem to be redundant. These tandem phosphatase domains led to the question, whether loss of the catalytic center occurred simultaneously with the evolution of a new functional center. To study functional differences at the amino acid level, evolutionary site rates of amino acid substitutions were compared between the active and inactive domains. A cluster of higher conserved positions was identified in the inactive domains that could function as a regulatory center. Following this first case study, a large-scale analysis of substitutions at catalytic positions of enzymatic domains was performed (chapter 5). The analysis was based on all enzymatic SMART domains with known catalytic mechanism. This study also gave insight into the function of inactive enzymatic domains. The overrepresentation of inactive domains among signalling domains suggests a role in regulatory networks and several examples in the literature support this finding.

A broader analysis on functional sites is described in chapter 6. Here, the dataset was enlarged to include ligand-binding sites extracted from experimental validated protein interaction data. Emphasis was on understanding the properties of functional sites. The sites were analysed for preferences towards specific amino acids. Besides, the usage of certain locations for the functional site was investigated.

The overall aim of these studies was to understand how protein domains function in their cellular environment and also, to improve protein function annotation. The results presented here are a significant step towards the understanding of the properties and differences between domain families. Progress has especially been made in the evolution of signalling domains. It is planned to present the data gained by this analysis to the scientific community in form of online search tools. The first step has already been done by the integration of catalytic sites in the SMART database. The data will improve the characterization of unknown protein sequences.

2 Prediction of structure and functional residues for O-GlcNAcase, a divergent homologue of Acetyltransferases

2.1 Abstract

N-acetyl-β-D-glucosaminidase (O-GlcNAcase) is a key enzyme in the posttranslational modification of intracellular proteins by O-linked N-acetylglucosamine (O-GlcNAc). Here, we show that this protein contains two catalytic domains, one homologous to bacterial hyaluronidases and one belonging to the GCN5- related family of acetyltransferases (GNATs). Using sequence and structural information, we predict that the GNAT homologous region contains the O-GlcNAcase activity. Thus, O-GlcNAcase is the first member of the GNAT family not involved in transfer of acetylgroups, adding a new mode of evolution to this large protein family. Comparison with solved structures of different GNATs led to a reliable structure prediction and mapping of residues involved in binding of the GlcNAc modified proteins and catalysis.

2.2 Introduction

The function of many proteins is regulated by posttranslational modifications. This is especially apparent in the case of signal transduction, where the activity of proteins has to be fine-tuned to react accordingly to intra- and extracellular signals.

Only recently it became clear, that in addition to phosphorylation of proteins also their glycosylation by O-linked β-N-acetylglucosamine (O-GlcNAc) might be involved in signal transduction (WELLS *et al.* 2001). In contrast to the more complex N- and O-linked glycosylation of extracellular proteins, this modification is added to serine or threonine residues of intracellular proteins. A plethora of proteins from species all over the eukaryotic kingdom has been reported to contain O-GlcNAc. The development of novel

methods usable in proteomics approaches like antibodies or mass-spectrometry can be expected to further increase this number.

These studies might add to the increasing evidence of an implication of O-GlcNAc into different human diseases like cancer (SHAW *et al.* 1996), diabetes (VOSSELLER *et al.* 2002) and neurodegenerative diseases. Tau for example, which is a major component of neurofibrillary tangles in Alzheimer's diseased brains, is multiply O-GlcNAc modified in normal brains, whereas it is highly phosphorylated in association with Alzheimers disease (ARNOLD *et al.* 1996).

Not only this case links glycosylation and phosphorylation, many different proteins are modified by both systems. As the same types of amino acids are modified, reciprocity was suspected. Indeed it was shown in different cases like the estrogen receptor (CHENG *et al.* 2000) and the protooncogene c-Myc (CHOU *et al.* 1995), that the same residues are either phosphorylated or glycosylated. A fine tuned regulation is crucial for this 'ying-yang' relationship. The responsible enzymes are the O-GlcNAc transferase (OGT) and the N-acetyl-β-D-glucosaminidase (O-GlcNAcase). The OGT consists of an N-terminal part build of TP repeats and a C-terminal catalytic unit. In depth sequence analysis revealed that the catalytic unit is homologous to the glycogen phosphorylase superfamily (WRABL and GRISHIN 2001). Less is known about the O-GlcNAcase, which was originally cloned as a hyaluronidase and only later shown to additionally possess glucosaminidase activity. The region responsible for the O-GlcNAcase activity has not been mapped and neither the substrate binding site nor catalytic residues have been described, hindering detailed experimental characterisation of the GlcNAcase and restricting the analysis of O-GlcNAc modification.

2.3 Materials and Methods

Members of the O-GlcNAcase family were found by BLASTp (ALTSCHUL *et al.* 1990) searching with the human MGEA5 sequence against NCBI's non-redundant database. To identify known domains, sequences were checked against against SMART (LETUNIC *et al.* 2002) and Pfam (BATEMAN *et al.* 2002). The C-terminal sequence of MGEA5 (700-916) was further analysed by PSI-BLAST. The obtained putative O-GlcNAcase sequences were aligned with ClustalX (THOMPSON *et al.* 1997) and the

resulting alignment was manually optimised in Seaview (GALTIER *et al.* 1996). Secondary structure elements were predicted by Jpred (CUFF *et al.* 1998). To identify homologs with known structures, intermediate sequence search (PARK *et al.* 1998) was used.

Within the N-acetyltransferase family three additional proteins were selected for a structural alignment: The closely related GNA5 histone acetyltransferase (PDB Id. 1I12) and two Aminoglycoside N-acetyltransferases (1bo4 and 1b87) because of their abilities to bind GlcNAc similar substrates. Pairwise structural alignments performed by VAST (MADEJ *et al.* 1995) were obtained from NCBI's MMDB (WANG *et al.* 2002). By comparing each of the four structures to each other a multiple alignment was build based on secondary structure elements.

2.4 Results

Searches with MGEA5 (gi: 10835356), the first cloned O-GlcNAcase (GAO *et al.* 2001), against different domain databases (Pfam, BATEMAN *et al.* 2002; SMART, LETUNIC *et al.* 2002) did not reveal any conserved motif, although Comtesse et al. report a C-terminal acetyltransferase domain found by SMART (COMTESSE *et al.* 2001). Still, a sequence search using blastp (ALTSCHUL *et al.* 1990) against the non-redundant subset of GenBank found different, significant homologous proteins. These include bacterial hyaluronidases as reported in Heckel et al. (HECKEL *et al.* 1998) (Evalue from 10-41 to 3 x 10-8) and a putative acetyltransferase (E = 3 x 10-8). Inspection of the pairwise alignments revealed two regions of homology, separated by a linking sequence. The N-terminal region was homologous to the hyaluronidases, whereas the similarity to the acetyltransferases was restricted to the C-terminus.

Iterative Blast searches with the C-terminal part (AA 700 to 916) found different acetyltransferases. All found proteins belong to the GCN5-related family of acetyltransferases (GNAT), for which a couple of structures have been solved (for review see DYDA *et al.* 2000). Yet, direct searches with the C-terminus of MGEA5 against the sequences of proteins with known structure did not reveal significant similarities. This link could only be established using intermediate sequence searches (PARK *et al.* 1998), that is, searching with sequences retrieved in previous searches. For example, the putative acetyltransferase sequence from *Streptomyces coelicolor* (gi: 7160091) found in a search

with MGEA5 (E=7 x 10-12) did detect the PDB sequence 1CJW, a serotonin acetyltransferase (E=0.091).

To further strengthen the link between the O-GlcNAcases and acetyltransferases selected members of the GNAT family were aligned according to their structures. A secondary structure prediction based on an alignment of the C-terminal region of the O-GlcNAcase family revealed a significant congruence between the secondary structure elements (Figure 6). Taken together, these results show that the O-GlcNAcase is a divergent member of the GNAT family.

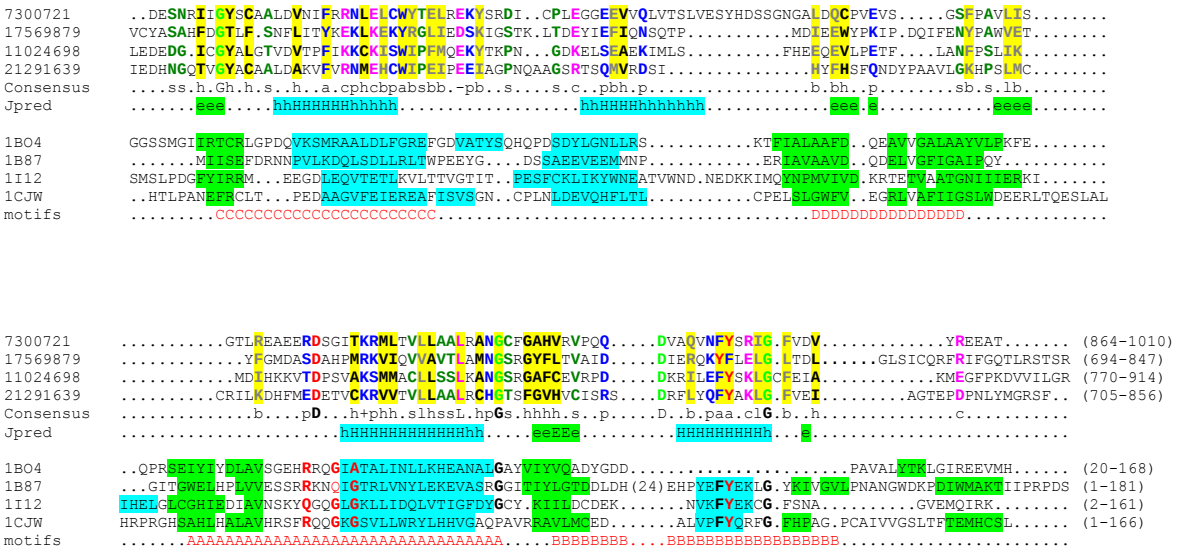


Figure 6: Multiple alignment of O-GlcNAcase sequences and structurally characterised members of the acetyltransferase superfamily. Sequences are labeled with Genbank or PDB identifier on the left side, the region of the sequence used in the alignment is given in parenthesis on the lower right side. Numbers in parenthesis within the alignment indicate the length of an insert. The upper four sequences (7300721: Drosophila melanogaster , 17569879: Caenorhabditis elegans, 11024698: Homo sapiens 21291639: Anopheles gambiae) are representatives of the O-GlcNAcase family. Sequences from Mus musculus (gi: 14329484) and Rattus norvegicus (gi: 16943639) are not shown, as they are in the aligned region nearly identical to the human sequence. Conserved residues are coloured and the 80% consensus determined by Chroma (GOODSTADT and PONTING 2001) is given below (capital letters represent aminoacids, lower case letters: a aromatic, c charged, s small, p polar, b big, h hydrophobic, - negative, + positive). A consensus secondary structure prediction for these sequences was obtained from Jpred and is shown in the corresponding line. α-helices are denoted by h/H, β-sheets by e/E. Capital letters indicate a reliability score higher than 6. The lower four sequences (1BO4: aminoglycoside 3-N-acetyltransferase, 1B87: aminoglycoside 6-N-acetyltransferase, 1CJW: serotonin N-acetyltransferase, 1I12: glucosamine-phosphate N-acetyltransferase) belong to the family of Gcn5-related N-acetyltransferases. Structural elements are hallmarked by blue background for helices or green background for β-sheets. Conserved functional residues are highlighted by bold red characters, conserved residues by bold characters. Beneath the alignment positions of the four conserved regions are shown that are described as motif A, B, C and D (NEUWALD and LANDSMAN 1997).

The degree of conservation between the GlcNAcase family and the acetyltransferases varies in different regions of the alignment (Figure 6). The C-terminal region, motif B, is highly conserved between the GlcNAcases and a subfamily of acetyltransferases including glucosamine-6-phosphate N-Acetyltransferases (PDB id. 1I12) and Serotonin N-Acetyltransferases (PDB id. 1CJW). This region with the consensus FYxxxG is not present in other GNATs with differing substrates like the Aminoglycoside 3'N-Acetyltransferases. Only in the *Caenorhabditis elegans* sequence, two of these residues have been 'swapped'. Less conservation can be found in motif A containing the AcetylCoA binding site of the GNATs which can be described by the consensus sequence R/QxxGxG/A (NEUWALD and LANDSMAN 1997). This motif is not present in the corresponding region of the GlcNAcases, which in contrast contain a conserved D not found in the GNATs.

2.5 Discussion

2.5.1 Structure and functional residues of O-GlcNAcase

The human gene encoding an O-GlcNAcase, MGEA5, was cloned by two independent studies, one purifying a hyaluronidase (HECKEL *et al.* 1998), the other a glucosaminidase (GAO *et al.* 2001). As the substrates of these enzymes share some similarity, it was speculated that both reactions are performed by one, unspecific catalytic center (COMTESSE *et al.* 2001). Here we showed that MGEA5 contains two regions similar to distinct enzymes. The N-terminal region is homologous to bacterial hyaluronidases, whereas the C-terminus shows homology to acetyltransferases. Evidence that the C-terminal part contains the glucosaminidase activity comes from the observation that one of the acetyltransferases most similar to MGEA5 is GNA1. This protein, whose structure has been solved (HANOVER 2001), is involved in the biosynthesis of UDP-GlcNAc by catalyzing the formation of GlcNAc6P. This led to the question, whether the ability to bind GlcNAc might be conserved in MGEA5. The region of GNA1 responsible for interaction with GlcNAc6P, commonly denominated Motif B (NEUWALD and LANDSMAN 1997), is highly conserved within the O-GlcNAcases but not in other acetyltransferases with strongly differing substrates like aminoglycosids (Figure 6, sequence 1bo4). The most

prominent feature is the conservation of a tyrosine residue, which is involved in the binding of GlcNAc and the catalytic mechanism of GNA1 (HANOVER 2001). This tyrosine as well as two other positions is conserved within all members of the O-GlcNAcases. Taken together, this led to the prediction, that the C-terminal, acetyltransferase homologous region of MGEA5 is responsible for binding O-GlcNAc modified proteins. Furthermore, the conserved tyrosine might be involved in the catalytic function of MGEA5.

Contrasting the well conserved proposed GlcNAc binding site, regions which are in the classical acetyltransferases involved in Acetyl-CoA binding have diverged in the GlcNAcases. Most of the interactions between the acetyltransferases and Acetyl-CoA are performed by a helix termed motif A (Figure1). Within this motif, residues which can be described by a Q/RxxGxG pattern are involved in direct interaction (PENEFF *et al.* 2001). This pattern is not conserved in the GlcNAcases. Instead, they contain a conserved D within the region of this pattern. As, according to the structures of the GNATs, this residue is solvent accessible and close to the proposed catalytic tyrosine, it might either be involved in binding of the O-GlcNAcase modified protein or directly in the catalytic process.

Additional experimental evidence for a C-terminal location of the glucosaminidase comes from a splice variant of MGEA5, which misses the C-terminal region (MGEA5s). Whereas the full-length variant is mainly localised in the cytoplasm, the short form has a nuclear localisation (COMTESSE *et al.* 2001). Together with the finding that major parts of the O-GlcNAcase catalytic activity are present in the cytoplasm (WOLF *et al.* 1998), this indicates that the catalytic activity is localised in the C-terminal region of MGEA5, which is homologous to acetyltransferases.

2.5.2 O-GlcNAcases as linkers of different regulatory processes?

Two catalytic functions have been ascribed to MGEA5, a hyalurinodase (HECKEL *et al.* 1998) and a glucosaminidase (WELLS *et al.* 2002) activity. The mapping of two independent catalytic domains raises the question, why these are co-localised in the same protein. The presence of intracellular hyaluronan, the substrate of hyaluronidases, was described only recently and an involvment in regulatory processes is assumed (LEE and SPICER 2000). The GlcNAc modification also takes part in regulatory mechanisms and is

highly interwoven with phosphorylation (WELLS *et al.* 2001). This might indicate that MGEA5 links two regulatory mechanisms. With the mapping of the responsible catalytic domains, both processes can now be studied independent of each other.

The prediction of functional residues for the glucosaminidase will allow creating non-functional mutants. As these should not be able to remove GlcNAc groups from modified proteins, they will influence the cellular GlcNAcase modification state. Complementary to described specific inhibitors, these mutants might therefore be useful in the determination of proteins, which undergo glycosylation in response to various stimuli.

2.5.3 A novel mechanism in the evolution of Acetyltransferases

In the evolution of novel enzymes one can imagine two different scenarios. Either a protein which has a defined catalytic activity evolves a new substrate affinity or an enzyme with a given substrate affinity changes the underlying catalytic mechanism. All so far described members of the GCN5-related acetyltransferase family transfer an acetyl group from Acetyl-CoA to a wide range of substrates like small molecules and proteins using a conserved catalytic mechanism (DYDA *et al.* 2000). On the structural level, GNATs are related to N-myristoyltransferases (BHATNAGAR *et al.* 1999). Although the transferred molecule is changed, the major catalytic mechanism and the mode of cofactor binding are conserved (BHATNAGAR *et al.* 1998). This indicates that the prevalent mode of evolution in this superfamily is the change of substrate affinity by keeping catalytic mechanism and cofactor binding constant. The membership of O-GlcNAcase to the family of acetyltransferases adds a new mode of evolution to this family. Not only a novel catalytic mechanism has developed but also cofactor-binding capabilities were lost.

This gives rise to questions about the possible starting point for this novel invention. One step in the biochemical pathway leading to UDP-GlcNAc, the substrate that is transferred by the OGT to a protein, is the addition of an acetyl group to Glc6P, giving GlcNAc6P. The protein catalysing this step, GNA1, belongs to the GNAT family and has an affinity for GlcNAc (HANOVER 2001). Starting from this protein, the O-GlcNAcase activity might have evolved by changing the underlying catalytic mechanism. Thus, one could imagine a recruitment of this protein from the metabolic pathway to the regulatory O-GlcNAc pathway.

3 Prediction of Cytokinin-binding sites in the CHASE domain

3.1 Introduction

Cytokinins are plant hormones that regulate many important developmental and physiological processes. In the last years several reports demonstrated that cytokinin signals are perceived and transduced by phosphorelay systems (for review see HEYL and SCHMULLING 2003). The "His to Asp phosphorelay" or "two component system" is based on two types of signal transducers, histidine kinases and response regulators, between which the phosphotransfer is mediated (PARKINSON and KOFOID 1992; ALEX and SIMON 1994). The two component system is found in many lower eukaryotes and prokaryotes, but is unique to plants among higher eukaryotes. In *Arabidopsis thaliana*, binding of cytokinin to the extracellular sensor domain of the membrane-located histidine kinase leads to dimerization of the receptor and, subsequently, to autophoshorylation of a conserved histidine residue in the intracellular transmitter domain. This phosphoryl group is then transferred to an aspartate residue within the receiver domain, which is located at the carboxy terminus of the receptor. In the following, the signal is transmitted into the nucleus via a second round of phosphoryl-shuttling between histidine and aspartate residues. This involves the transfer of the phosphoryl group onto a histidine of a phosphotransfer protein and finally onto an aspartate residue of a response regulator (SHEEN 2002).

Promoted by the *Arabidopsis thaliana* genome project, several cytokinin receptors have now been identified. The CRE1/AHK4/WOL, AHK2 and AHK3 proteins are all known to be involved in cytokinin signalling (ARABIDOPSIS GENOME INITIATIVE 2000; INOUE *et al.* 2001). These proteins contain a large extracellular domain and the intracellular histidine kinase and receiver domain. The extracellular domain is approximately 270 amino acids long and flanked by transmembrane regions. This domain has been recently described by two individual groups and termed CHASE (cyclases/histidine kinases associated sensory extracellular) domain because of its presence in extracellular regions of sensory proteins in combination with a intracellular cyclase or histidine kinase

(ANANTHARAMAN and ARAVIND 2001; MOUGEL and ZHULIN 2001). Both groups identified several homologues of the extracellular domain of the cytokinin receptor CRE1 in plants, slime molds and bacteria such as cyanobacteria and proteobacteria. Archae bacteria seem to lack this sensory domain completely. This phylogenetic distribution suggests that eukaryotes might have acquired this domain by horizontal transfer from bacteria. The genetic event by which plants have acquired the CHASE domain might have been the same by which the plant lineage has obtained chloroplasts, which are of cyanobacterial origin.

The type of ligand activating the various CHASE domains in different species remains to be investigated. In case of *Arabidopsis*, the CHASE domain was described to be able to bind cytokinin (INOUE *et al.* 2001). Different plant CHASE domains have different binding affinities for the various kinds of cytokinins (SPICHAL *et al.* 2004). Experiments in *E. coli* have shown that cytokinins can be of the isopentenyl or zeatin type (SUZUKI *et al.* 2001). The domain can also be activated by certain derivates of diphenylurea (YAMADA *et al.* 2001) and the CHASE domain has been reported to bind small peptide molecules. The DhkA receptor kinase from *Dictyostelium discoideum* can be activated by the SDF-2 peptide. (WANG *et al.* 1999). Another CHASE domain in *Dictyostelium discoideum* appears together with an adenylyl cyclase and is dependent on discadenine, a cytokinin-like molecule (COTTER *et al.* 1999). It seems that different versions of the CHASE domain can bind adenine derivatives like cytokinin as well as small peptide ligands.

Although potential ligands of the CHASE domain have been identified, little is known about the functional sites of this domain, nor its three-dimensional structure. So far, only one functional residue of the CHASE domain has been suggested by mutation experiments. Threonine at position 301 in the extracellular domain of AHK4/WOL appears to bind cytokinin as inferred from loss of function when mutated to isoleucine (MAHONEN *et al.* 2000).

Here, we study amino acids important for cytokinin dependent activation of the plant phosphorelay system by analysing evolutionary rates of amino acid substitutions in functionally different subgroups of the CHASE domain. We predict several sites presumably involved in ligand-binding. Subsequently, positions of interest were experimentally mutated and their effects on cytokinin binding were studied.

3.2 Methods

A Hidden Markov model (HMM) was built from the multiple sequence alignment reported by Anantharaman and Aravind (ANANTHARAMAN and ARAVIND 2001) using the HMMER package (http://hmmer.wustl.edu/). With this model, Genbank's non-redundant database (August 2003) was searched for sequences containing the CHASE domain. 56 sequences with E-values below 0.001 were found. To identify additional plant sequences, Genbank's EST database (est_others, August 2003) was searched with the AHK4/WOL part comprising the CHASE domain. A multiple sequence alignment was built from sequences spanning the majority of the CHASE domain with hmmalign (HMMER package) according to the CHASE HMM. The alignment was manually optimised. A phylogenetic tree was calculated with CLUSTAL W (THOMPSON *et al.* 1994).

Five stable subgroups of the tree were used for the further analysis: Excluding identical sequences, subgroup A contains sequences from plants like *Arabidopsis*, rose, tomato, rice and millet (gi18379305\175-388, gi28301941\110-321, gi12060392\86-298, gi18421494\302-526, gi22353121\154-378, gi20279446\50-274, gi9802528\163-389, gi20161916\151-377, gi14513877 (EST), gi9415459 (EST), gi33217468 (EST), gi14823726 (EST), gi7265090 (EST), gi16247084 (EST)), while the other subgroups contain sequences form proteobacteria: subgroup B (gi23055097\38-240, gi27364428\14-215, gi13471907\45-248, gi15966662\49-253, gi23027857\50-253, gi23000375\48-254, gi15601713\42-248, gi28900053\42-248, gi27367985\42-248, gi23013402\61-262), subgroup C (gi23058768\88-299, gi26988822\85-301, gi28869512\121-331, gi23468834\87-298), subgroup D (gi22968449\79-295, gi15599307\77-286, gi21885294\85-301, gi21243225\83-298, gi21231797\83-298, gi23471793\76-296), subgroup E (gi23030255\93-313, gi23053590\87-308, gi32475880\90-306, gi24372138\85-305, gi22982057\71-291). EST sequences were translated into amino acid code.

For each subgroup, evolutionary rates were estimated with TREE-PUZZLE v5.1 using the quartet puzzling algorithm under the substitution model of Jones-Taylor with an 8 site-rate category discretized gamma model (YANG 1994; STRIMMER and VON HAESELER 1996).

3.3 Results and Discussion

3.3.1 Computational prediction

In order to identify putative ligand-binding positions in the plant CHASE domain, homologous sequences of the CHASE domain originating from the AHK4/WOL histidine kinase of *Arabidopsis thaliana* were searched. 56 sequences originating from proteobacteria, cyanobacteria, the slime-mold *Dictyostelium* and plants were retrieved from Genbank's non-redundant database and, additionally, 6 sufficiently long plant sequences were identified in Genbank's EST database. An unrooted phylogenetic tree of the CHASE domains and the species of origin is shown in figure 7.

The underlying assumption of this analysis was that ligand binding sites are well conserved within functional subgroups of the CHASE domain and that different subgroups bind biochemically different ligands. Evidence for diverse ligand specificity in the CHASE domain was found in the literature. CHASE domains from plants seem to bind cytokinin, while at least some of the bacterial CHASE domains appear to prefer small peptides (COTTER *et al.* 1999; WANG *et al.* 1999; SUZUKI *et al.* 2001; YAMADA *et al.* 2001). This functional difference could be reflected by different patterns of amino acid conservation. By searching for positions that are evolving at a very slow rate of amino acid substitutions among plant sequences in contrast to bacterial sequences, we should be able to identify positions involved in cytokinin binding. Bacterial amino acid positions corresponding to cytokinin-binding positions in plants can present various behaviours. They could also be conserved, but very likely contain an amino acid with different biochemical properties to adjust to the nature of the peptide ligand. The behaviour of a position that is conserved in two subgroups but with different amino acids in the different subgroups is also termed type II functional divergence. On the other hand, the corresponding position in a bacterial subgroup might not at all be involved in the interaction with a ligand and might be occupied by a wide variety of amino acids. In this case the subgroups would differ in the degree of conservation caused by the altered functional constraint. This behaviour is referred to as type I functional divergence. Functional divergence can be easily investigated by comparing evolutionary rates of amino acid substitutions between different subgroups. Conserved positions present very small evolutionary rates, whereas high rates indicate loss of selective pressure and loss of function.

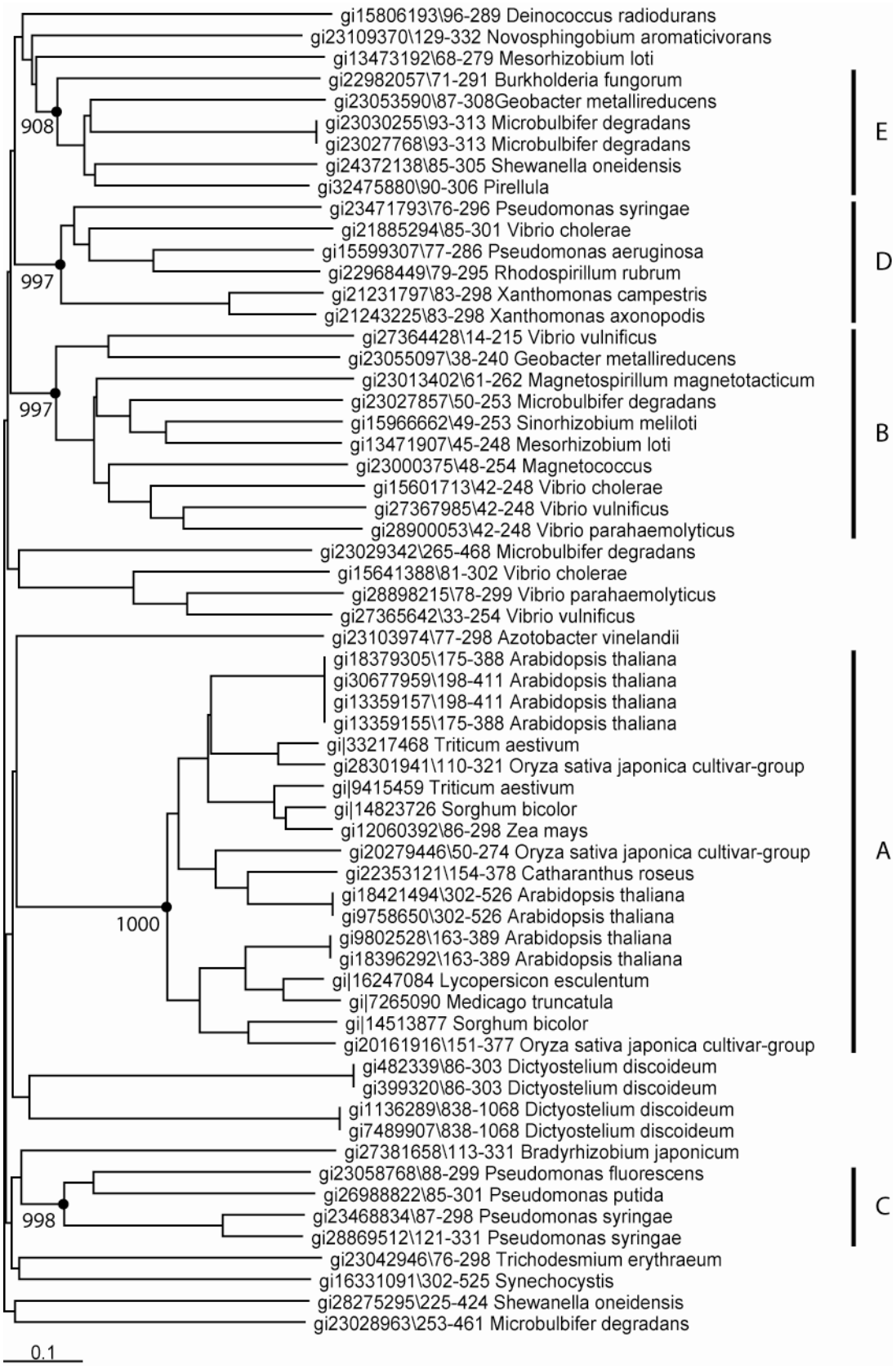


Figure 7: Phylogenetic tree of the CHASE domain. The tree was generated with CLUSTAL W. Bootstrap values are indicated for the subgroups used in the analysis. 1000 bootstrap resampling steps were performed.

Figure 8: Section of CHASE alignment containing sequences used in the analysis of evolutionary site rate comparison (above and next page). Classification into subgroups is indicated to the right of the alignment. The site rate category for each investigated subgroup and for each alignment position is given below the alignment. The categories range from 1 (slow rates of evolution) to 8 (fast rates of evolution). Putative ligand-binding sites identified in the analysis are indicated by triangles above the alignment.

We searched positions that expose either type I or type II functional divergence in the multiple sequence alignment (figure 8) and predicted positions that are putative ligand-binding sites in the plant CHASE domain. All positions refer to the *Arabidopsis thaliana* cytokinin receptor CRE1/AHK4 (gi30677959). The evolutionary rate category for these positions in the different subgroups is displayed in table 2. In the following, the most likely cytokinin-binding positions are listed, ranked by their support from the functional divergence analysis.

1. The most promising candidate is position 317, occupied by threonine in the cytokinin receptor CRE1/AHK4. This position is absolutely conserved in all investigated subgroups and all subgroups evolve at slow rates at this position. In contrast to the threonine found in plants, other subgroups present a conserved arginine or small hydrophobic residue (type II functional divergence). The biochemical different properties at this position suggest involvement in ligand-binding and that this residue has been modified in different subgroups to recognize specific ligands.
2. Another striking residue is tryptophane at position 244, because it is conserved in all plant sequences, but variable in all other CHASE domains. This position is fast evolving in all other subgroups and occupied by various amino acids. The aromatic side chain of tryptophane seems to be well suited to complement the basic structure of cytokinin and thus to provide stable and specific interaction between cytokinin and its receptor.
3. Position 304 is well conserved in subgroup A, containing plant sequences. The phenylalanine might also facilitate specific ligand binding. Other subgroups favour small hydrophobic residues. It is interesting to note, that phenylalanine 304 is located in the vicinity of threonine 301, which has been suggested to bind cytokinin.
4. Although position 305 is not absolutely conserved in its amino acid it still seems highly interesting when compared to other subgroups. CHASE domains originating from plants favour positively charged amino acids like arginine at this position, in contrast to other subgroups, which display hydrophobic or even negatively charged amino acids.
5. Position 297 exhibits functional divergence of type I as it is highly conserved in the cytokinin-binding subgroup but fast evolving in other subgroups. In addition, there

are striking differences in the biochemical properties at this position. In subgroup A, this position is occupied by lysine. All other subgroups present mainly negatively charged amino acids.

Table 2: Categories of evolutionary sites rates of the different investigated subgroups for putative functional sites. The numbers represent one of the 8 categories of the discrete gamma distribution, where 1 refers to the slowest evolving category and 8 to the fastest evolving category. Positions and amino acids refer to the CRE1/AHK4 histidine kinase (gi30677959).

	T-317	W-244	F-304	R-305	K-297
Subgroup A	1	1	1	6	1
Subgroup B	1	8	1	7	6
Subgroup C	1	8	6	8	3
Subgroup D	1	2	2	1	6
Subgroup E	1	8	1	7	8

The positions presented here seem to be the most promising candidates for the cytokinin interaction. It must be noted that position 301, which leads to complete loss of function if the threonine residue is mutated to isoleucine, was not detected by our approach. This can be explained by the high conservation of amino acids at this position. All investigated subgroups scored in the category of slowest evolutionary rates. In addition, this position did not present functional divergence of type II as all sequences exhibit either serine or alanine residues. This indicates that position 301 is important for the function of the CHASE domain in general, for example for the correct folding of the structure, for passing the signal through the membrane or for binding a ligand, independent from the type of ligand.

3.3.2 Experimental verification

The positions proposed to play a role in cytokinin-binding were experimentally analysed (Alexander Heyl, personal communication). Various constructs of the *Arabidopsis thaliana* gene CRE1/AHK4 were generated, in which the suggested cytokinin-binding positions were substituted by alanine. The impact of these mutations on the ligand-binding ability and on the activity of the CRE1/AHK4 receptor was investigated in *E. coli* cells with trans-zeatin as ligand.

Activity of the receptor is easily detectable with a blue/white screen. The CRE1/AHK4 receptor expressed in *E. coli* can employ the bacterial phosphorelay system and activate the endogenous *cps* operon (SUZUKI *et al.* 2001). Using a *cps::lacZ* fusion gene, the down-stream effects of the CRE1/AHK4 receptor can be visualized. If *E. coli* cells are grown on media containing 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal), the external cytokinin induces the expression of the β-galactosidase reportergene, which catalyses the conversion of X-Gal into the insoluble Indigo dye, thus turning the colonies blue. A more precise analysis of the specific ligand-binding was performed in an *in vitro* binding assay. Radiolabeled cytokinin bound to the bacterial membrane fractions containing the CRE1/AHK4 was measured.

The results of these experiments are summarized in figure 9 (by courtesy of Nicola Nielsen). For each construct two different *E. coli* clones were analyzed. Four controls were used, which are displayed on the left side of the figure. The wild-type CRE1/AHK4 receptor shows the expected result in the blue/white screen and its ability to bind cytokinin was used as reference in the following. The empty vector delivered negative results in both

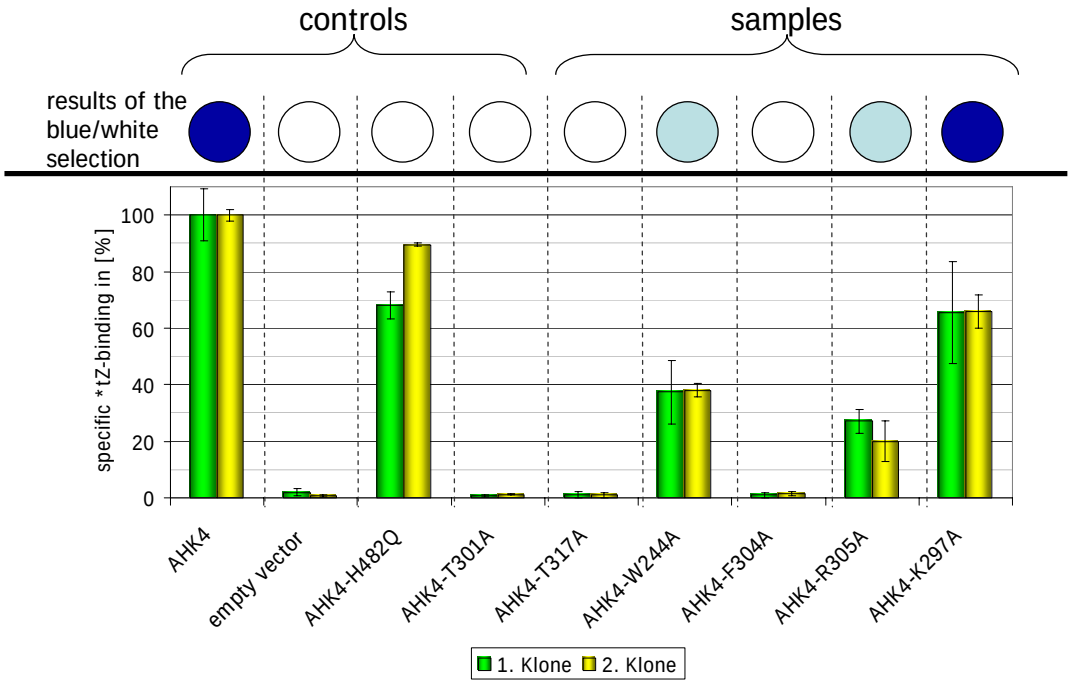


Figure 9: Results of the biochemical cytokinin binding assay. Binding ability of the CHASE domain was tested in a blue/white screen in E. coli cells and in an in vitro assay with radiolabeled ligand. The binding affinity was compared to the wild-type receptor, which was set to 100%. Two clones were tested for each construct indicated by green and yellow bars. (Courtesy of Nicola Nielsen)

experiments, hence *E. coli* cells are not capable to sense trans-zeatin. The CRE1/AHK4-H482Q construct carries a mutation in the intracellular kinase domain. Since cytokinin can still bind to the CHASE domain, the *in vitro* binding assay was positive, but the signal could not activate downstream targets and resulted in white *E. coli* colonies in the blue/white screen. The results for the CRE1/AHK4-T301A construct were negative in both assays, as expected from reports in the literature (MAHONEN *et al.* 2000). Among the constructs containing mutations at putative cytokinin-binding position, different results were obtained. Two constructs, the CRE1/AHK4-T317A and CRE1/AHK4-F304A mutant, produced negative results, indicating that these mutations lead to complete loss of function in the CHASE domain. Two mutations show partial impairment of function, presented by the CRE1/AHK4-W244A and CRE1/AHK4-R305A constructs. Here, the colonies turned light blue and the binding assay resulted in binding efficiencies between 25-50% of the wild-type protein. The CRE1/AHK4-K297A clone is as efficient in signal sensing as the wild-type AHK4. This mutation had only a minor effect on the ligand-binding and the in vitro assay showed 50-85% efficiency.

3.4 Conclusions

The computational analysis of evolutionary rates in the CHASE domain identified five putative functional residues, of which four residues lead to complete or partial impairment of cytokinin-binding if mutated. The negative experimental result of the binding assay does not necessarily mean that the mutated residue cannot interact with the cytokinin ligand any more. Mutation of a structural important residue could also cause a loss of function if the domain is not correctly folded. However, structural important sites should be conserved in the type of amino acid in all sequences of the CHASE domain. The functional sites identified here are either variable in bacterial sequences or they are conserved in a different type of amino acid. In addition, many bacterial sequences exhibit an alanine residue at positions corresponding to the identified functional sites. If alanine is tolerated at this position in bacterial sequences it is unlikely that it promotes incorrect folding or instability of the domain. These two observations strongly argue for the identification of real cytokinin-binding sites.

The experimental verification of the computational predicted functional sites demonstrates the accuracy of theoretical prediction methods. Taking advantage of the phylogenetic relationship within protein families, the development of specific functions in subfamilies can be traced. The successful outcome of the study presented here is encouraging to apply this method to other protein families in order to focus on promising candidate residues and to speed up tedious and pricey experimental studies.

4 Evolution of the multi-functional protein tyrosine phosphatase family

4.1 Abstract

The protein tyrosine phosphatase (PTP) family plays a central role in signal transduction pathways by controlling the phosphorylation state of serine, threonine and tyrosine residues. PTPs can be divided into dual specificity phosphatases and the classical PTPs, which can comprise of one or two phosphatase domains. We studied amino acid substitutions at functional sites in the phosphatase domain and identified putative non-catalytic phosphatase domains in all subclasses of the PTP family. The presence of inactive phosphatase domains in all subclasses indicates that they were invented multiple times in evolution. Depending on the domain composition, loss of catalytic activity can result in different consequences for the function of the protein. Inactive single domain phosphatases can still specifically bind substrate and protect it from dephosphorylation by other phosphatases. The inactive domains of tandem phosphatases can be further subdivided. The first class is more conserved, still able to bind phosphorylated tyrosine residues and might recruit multi-phosphorylated substrates for the adjacent active domain. The second has accumulated several variable amino acid substitutions in the catalytic center indicating a complete loss of tyrosine binding capabilities. To study the impact of substitutions in the catalytic center to the evolution of the whole domain, we examined the evolutionary rates for each individual site and compared them between the classes. This analysis revealed a release of evolutionary constraint for multiple sites surrounding the catalytic center only in the second class, emphasizing its difference in function compared to the first class. Furthermore, we found a region of higher conservation common to both domain classes, suggesting a new regulatory center. We discuss the influence of evolutionary forces on the development of the phosphatase domain, which has led to additional functions like the specific protection of phosphorylated tyrosine residues, substrate recruitment and regulation of the catalytic activity of adjacent domains.

4.2 Introduction

Protein tyrosine phosphatases (PTPs) regulate physiological processes common to all metazoa, including growth, differentiation, metabolism, the cell cycle and cytoskeletal function. Together with tyrosine kinases, they control the phosphorylation state of tyrosine and serine/threonine residues of signalling proteins in highly specific reactions. The level of protein phosphorylation is highly dynamic and any disturbance can lead to severe malfunction of the eukaryotic cell. Increased level of protein phosphorylation results in abnormal proliferation and many cancer types show a mutation or deletion of a PTP gene (SIMINOVITCH *et al.* 1999). In contrast to protein tyrosine kinases, that have a growth promoting potential, PTPs can act as tumor suppressors and inhibit cell growth (DAHIA 2000; WU *et al.* 2003). PTPs have also been implicated in B- and T- lymphocyte activation and insulin signalling, so that PTPs represent attractive drug targets for a wide variety of diseases, such as cancer, inflammation, diabetes and obesity (JUSTEMENT 2001; VAN HUIJSDUIJNEN *et al.* 2002; ASANTE-APPIAH and KENNEDY 2003).

Functionally, two types of PTPs, conserved in sequence and structure can be distinguished: The classical PTPs that are specific for tyrosine residues and the dual specificity phosphatases (DSPs), which can additionally dephosphorylate serine and threonine residues.

Obviously, the catalytic residues of an enzyme are key to its molecular function. Therefore, it came as a surprise, when a member of the DSP family, Sfb1, with a replacement of the catalytically essential cysteine was described (CUI *et al.* 1998). As expected, this protein had lost its enzymatic activity, raising the question about its molecular function. Experimentally, it could be shown, that Sbf1 has maintained its ability to stably bind phosphorylated substrate, protecting the substrate from other phosphatases at this specific site. Due to its antagonistic mechanism, this phosphatase has been termed 'anti-phosphatase' (DE VIVO *et al.* 1998; HUNTER 1998). Sbf1 function differs also on the cellular level. In contrast to active phosphatases exposing growth inhibitory behavior, it shows transforming abilities (CUI *et al.* 1998). Following this first description, a similar substitution of the catalytic cysteine was found in another subgroup of the DSPs, called STYX (WISHART *et al.* 1995). Within the classical PTPs, substitutions of functional residues have been described in the receptor protein tyrosine phosphatase (RPTP) family. Most RPTPs contain two phosphatase domains, of which the second phosphatase domain

is inactive or remains with very low activity due to substitutions at functional sites. Although the detailed function of this catalytically inactive domain is not yet totally understood, it is of absolute importance for the function of the receptors. Partial or entire deletion of the second domain completely abolishes or severely reduces activity of the first domain, so that a role in regulating the catalytic activity or substrate specificity of the first domain has been hypothesized (STREULI *et al.* 1990; JOHNSON *et al.* 1992). Furthermore, specific interaction of domain II with domain I leading to active site blocking of domain I (BILWES *et al.* 1996; MAJETI *et al.* 1998; BLANCHETOT and DEN HERTOG 2000) has been shown experimentally. Inactive phosphatase domains have also been studied experimentally in the single phosphatase domain RPTPs (KAMBAYASHI *et al.* 1995; CUI *et al.* 1996). Here, the PTP typical signature with the catalytic cysteine is present, but two other catalytically important residues are substituted, leading to loss of phosphatase activity.

On the basis of these more anecdotal reports, we analyze here on a genomic scale how frequent substitutions of functional residues in the PTP family are and whether they were invented multiple times in evolution. Furthermore, we use site-specific evolutionary rates to unravel the evolutionary implications of these substitutions for the whole domain, leading to the prediction of distinct functional classes and the delineation of an additional functional site.

4.3 Methods

4.3.1 Data Sets

Protein sequences for the whole genomes of *Homo sapiens* (version 9.30), *Mus musculus* (version 9.3), *Fugu rubripes* (version 10.2), *Anopheles gambiae* (version 9.1) and *Caenorhabditis elegans* (version 12.95) were retrieved from the ENSEMBL website (http://www.ensembl.org/). The *Drosophila melanogaster* predicted protein sequences (version 3) were obtained from BDGP (http://www.fruitfly.org/) and the Ciona intestinalis predicted protein sequences (version 1.0) from JGI (http://www.jgi.doe.gov/). These datasets were searched with Hidden Markov Models (HMMs) (HMMER: http://hmmer.wustl.edu/) specific for the classical PTPs and DSPs, respectively, which

were built from the family multiple sequence alignment retrieved from SMART (http://smart.embl-heidelberg.de/). For subclassification, SMART combined E-value thresholds were used (SCHULTZ *et al.* 1998). This schema allows to set two E-value thresholds for each subfamily HMM of a larger, homologous family. The first is given by the E-value of the best scoring family member not belonging to the actual subfamily, the second by the worst hit before the first non-family member. Searching with all subfamily specific HMMs assigns a sequence to a subfamliy if its E-value is lower than the according subfamily threshold. Sequences, which cannot be assigned to any subfamily but with an E-value lower than the family cutoff in at least one search are assigned to the family without any subfamily classification. These sequences were marked as 'undefined specificity' and not used in the further analysis. We cleaned the obtained data set by filtering the sequences for alternative splice variants and for fragments. Only the best scoring hit per gene was used for further analysis and sequences that did not cover the complete HMM, tolerating an uncovered interval of 10 amino acids at the beginning and end of the profile, were excluded as they might represent fragments.

4.3.2 Scan for inactive phosphatases

For each PTP subclass, a multiple sequence alignment of all found members was created according to the family alignment in SMART, using HMMalign (HMMER). The alignments were manually curated to remove unnecessary gaps and to connect interrupted secondary structure elements, partially with the help of secondary structure information, using a representative structure from PDB (1LAR for the classical PTPs, 1VHR for the DSPs). Knowing the positions of functional residues from literature, we were able to scan the sequences for substitutions at these sites and to extract the substituted amino acid from the alignment. If these sites were occupied by gaps, absence of the functional residue was not considered in the further analysis.

4.3.3 Evolutionary Rate Analysis

After definition of the subclasses (D2A, D2B, membrane proximal and cytosolic), we compared the evolutionary rates of all sites between the subclasses. Therefore, we selected the following genes from human and mouse for further analysis: *Homo sapiens*: ENSP00000175756, ENSP00000246887, ENSP00000248594, ENSP00000256635,

ENSP00000262539, ENSP00000263708, ENSP00000311857; *Mus musculus*: ENSMUSP00000022508, ENSMUSP00000025420, ENSMUSP00000027633, ENSMUSP00000029053, ENSMUSP00000029433, ENSMUSP00000030556, ENSMUSP00000048119. The evolutionary site rates were estimated with TREE-PUZZLE v5.1 (STRIMMER and VON HAESELER 1996) using the quartet puzzling algorithm under the substitution model of Jones-Taylor-Thornton with an 8 site-rate category discretized gamma model (YANG 1994). This model sets the average rate of a site to 1 and assigns each site to one of 8 rate classes. As the rate of one class can differ between analyses of different subfamilies, we considered a site as differentially evolving if its rate was below 0.8 in the conserved subfamily and above 1.6 in the fast evolving one. These values were chosen as within all analyses they covered classes 1-4 for the conserved domain and classes 7-8 in the fast evolving domain (total rates range approximately from 0 to 3 within all analyses). We performed the analyses separately for mouse and human and accepted only sites which were classified as differentially evolving in at least 3 of the 4 possible comparisons.

4.4 Results and Discussion

4.4.1 Amino acid substitutions at functional sites

As a first step to analyze the evolution of tyrosine phosphatases, we searched all to date sequenced metazoan genomes for phosphatase domains using specific HMMs for the classical PTPs and DSPs created from the SMART family alignments. Table 3 shows the presence of phosphatase domains in different genomes after filtering for fragments and alternative splice variants. Phosphatase domains that could not be clearly assigned to either one subclass are listed as “undefined specificity” phosphatases. The number of phosphatase domains found in human, mouse and pufferfish, which have a comparable proteome size, are similar for the total number as well as within the subtypes. Drosophila and Anopheles show the same ratio of phosphatases to the proteome as seen in vertebrates. Contrasting this, the genomes of *Caenorhabditis elegans* and *Ciona intestinalis* contain a substantial increase of phosphatases. Although the Ciona proteome is comparable in size to the Drosophila proteome, Ciona has more than twice as many phosphatases and almost as

many as vertebrates, caused by an expansion in the tandem domain RPTPs. An even larger expansion can be observed in *C. elegans*, whose genome contains twice as many classical PTPs as human or mouse. Here, the multiplications fall into the class of the single domain PTPs.

Table 3: Phosphatase domains in metazoa.

Note – The numbers of gene predictions were obtained from the ENSEMBL web site, in case of Drosophila from Flybase and for Ciona from Dehal and colleagues (DEHAL *et al.* 2002).

	DSPs	PTPs	undefined specificity	total	gene predictions
Homo sapiens	24	39	29	**104**	24847
Mus musculus	23	33	27	**93**	24948
Fugu rubripes	27	35	24	**86**	35180
Caenorhabditis elegans	11	91	36	**138**	19942
Ciona intestinalis	10	37	11	**58**	15852
Anopheles gambiae	8	18	7	**33**	14658
Drosophila melanogaster	8	20	16	**44**	13639

One aim of the analysis was to investigate the extent of substitutions in functional sites within the phosphatase domain family. We focused on sites directly involved in catalysis or substrate binding as it can be expected that substitutions in these sites will lead to a loss of catalytic activity. Within the PTP subfamily, these sites are (here and in the following, all positions regarding a classical PTP refer to pdb structure 1LAR): C-216, for attacking the phosphorous atom nucleophilically to form a phosphoenzyme intermediate, D-184 and Q-260 to position and polarize the active water molecule that dephosphorylates the phosphoenzyme in a second nucleophilic reaction, the substrate binding R-222 and the aromatic tyrosine or phenylalanine at position 49 that forms a stack with the phenyl ring of the phosphotyrosine (STUCKEY *et al.* 1994; FAUMAN *et al.* 1996; PUIUS *et al.* 1997; PANNIFER *et al.* 1998). Catalytic functional sites in the DSP family are (positions regarding a DSP refer to pdb structure 1VHR): C-123 and R-129 analog to the PTPs, D-91 that acts as general acid in the dephosphorylation step and as general base in the hydrolysis of the phosphoenzyme, the latter reaction is supported by S-130 (YUVANIYAMA *et al.* 1996;

PUIUS *et al.* 1997). Using a multiple sequence alignment of all obtained PTP and DSP sequences, respectively, we were able to extract substitutions at these positions.

The analysis reveals that all subclasses of the DSPs and classical PTPs, including the membrane-proximal, membrane-distal and cytosolic domain, carry substitutions in functional positions.

Table 4 shows the distribution of phosphatase domains among subclasses for each investigated species and the number of domains with substitutions at functional sites within each subclass. The number of putative inactive domains varies strongly among the different species, however, it accounts for a significant part of all subclasses, even in the single domain phosphatase subfamily. One might argue that the catalytically inactive proteins are non-functional relicts in the genome, but their conservation between species and also within subfamilies, strongly indicates their functional importance. In many of the observed cases the protein contains either a single non-active phosphatase domain or if it contains two, both are inactive. Depending on the type of substitution, these proteins are good candidates as potential anti-phosphatases.

An extremely high portion of amino acid substitutions is found in the membrane-distal domain of RPTPs, in which all sequences feature two or more substitutions. Based on the type of substitutions, we split this group into two subfamilies. This split is supported by a phylogenetic analysis, which revealed a monophyletic origin of the subfamilies. One subclass (D2A) shows a very high degree of conservation in these substitutions, with the catalytic aspartic acid (position 184) replaced by glutamic acid and the substrate binding tyrosine (position 49) either replaced by valine or leucine. The fact that these substitutions in functional residues maintain the biochemical properties of the original amino acids suggests that this domain might be able to carry out the dephosphorylation reaction. Indeed, experiments with HPTPα, a receptor type protein tyrosine phosphatase carrying a D→E and Y→V substitution at the functional sites in the membrane-distal domain shows a small rest activity, even if D2 is expressed by itself (WANG and PALLEN 1991). Still, mutation of these sites to the amino acids found in active phosphatases lead to a full recovery of catalytic activity (LIM *et al.* 1998). These experimental results on the one hand corroborate the functional importance of the identified sites, on the other they hint that there is still selective pressure on the catalytic center of D2A.

Table 4: Number of PTP domains and of putative inactive domains among the different PTP subclasses.

Note – Subclassification of PTPs according to the phylogenetic tree of the multiple sequence alignment. Unequal quantities of D1 and D2 domains are probably caused by errors in gene predictions. In case of DSPs and cytosolic PTPs, the number of domains corresponds to the number of genes, while the PTP-D1 and PTP-D2 domains originate from genes expressing tandem domain phosphatases.

	DSP		Cytosolic PTPs		PTP-D1		PTP-D2A		PTP-D2B	
	total	inactive	total	inactive	total	inactive	total	inactive	total	inactive
Homo sapiens	24	4	28	8	11	0	5	5	7	7
Mus musculus	23	4	22	5	11	1	5	5	7	7
Fugu rbripes	26	2	19	6	10	2	4	4	2	2
Caenorhabditis elegans	11	1	85	53	3	1	0	0	3	3
Ciona intestinalis	10	4	11	5	14	2	3	3	9	9
Anopheles gambiae	8	0	9	4	4	1	1	1	4	4
Drosophila melanogaster	8	0	12	4	4	1	1	1	3	3
sum	110	15	209	98	43	4	18	18	25	25

In the second subclass (D2B) substitutions in functional sites are more frequent and more heterogeneous than in subclass D2A. The high number of substitutions and the high variety in amino acids makes it seem unlikely that these domains maintained their catalytic activity. Indeed, Streuli et al. experimentally showed that the D2B domain of CD45 completely lost its phosphatase activity (STREULI *et al.* 1990). The fact that the inactive phosphatase domain has not been lost during evolution and that orthologues are found in all species investigated in our study excludes the mutation to a non-functional 'pseudodomain'. On the contrary, since the PTP structure is still conserved, a specialization on other functions is likely to have occurred during evolution.

In summary, substitutions in functional residues reside in all subclasses of the phosphatase family. This raises the point, whether the inactive phosphatase was invented once or multiple times. If invented once, its widespread presence would indicate an origin before the split into the subclasses. Assuming this monophyletic origin would imply, that

inactive phophatases of different subclasses are more related to each other than to other members of the subclass, which is not the case. Furthermore, it was shown that the major subclasses evolved mononphyletically (ANDERSEN *et al.* 2001). Therefore, we conclude that the invention of inactive phosphatases happened multiple times independently during evolution. The percentage of non-functional phosphatases in metazoan genomes is surprisingly high. The regulation of signalling pathways by protection of phosphorylated serine/threonine or tyrosine residues might therefore turn out to be an important mechanism to modulate signalling pathways. It has to be further investigated whether the phenomena of non-functional enzymes is restricted to the phosphatase family or whether other signalling enzymes show a similar behaviour.

4.4.2 Analysis of altered evolutionary rates between phosphatase subclasses

The striking high number of amino acid substitutions in functional residues in the membrane-distal phosphatase domain of receptor PTPs, which is associated with loss of activity, raises the question how on the one hand these substitutions evolved and, on the other, how they influenced the evolution of the whole domain. As a change in function should be mirrored in a change of evolutionary constraints at the involved sites, we compared the site-specific evolutionary rates between the active (cytosolic and membrane-proximal) and the inactive (D2A and D2B) PTP subclasses, a method which has been used to identify functionally important sites (GU and VANDER VELDEN 2002; BLOUIN *et al.* 2003). Due to the lack of a representative quantity of PTP sequences in the single subclasses of most species evolutionary rates could only be estimated reliably for human and mouse. Indeed, we found a substantial number of sites with changing evolutionary constraints (Table 3). To further understand how these work together, we mapped these sites onto the structure of the PTP domain (pdb 1LAR). This revealed that these sites cluster within two regions. One group is located around the catalytic center, while the other is located on the backside of the protein (figure 10). In the following section we discuss these regions separately.

Table 5: Sites with altered evolutionary rates between the PTP subclasses.

Note – Each comparison was performed for human and mouse. Sites with altered evolutionary rates correspond to the 75% consensus of the 4 resulting site rates for each position. Comparison A and B reveal sites that are evolving faster in the inactive domains D2A and D2B than in the active membrane-proximal domain (D1) and the cytosolic domain of single domain phosphatases (cyto). In contrast, comparison C and D show sites that are more conserved in the inactive domains D2A and D2B. All positions refer to pdb structure 1LAR.

Comparison	Figure	Sites with altered evolutionary rates
D2A > D1 D2A > cyto	10A	62, 109, 156, 185, 210
D2B > D1 D2B > cyto	10B	49, 55, 107, 122, 127, 156, 172, 180, 184, 185, 230, 254, 260, 261, 268
D2A < D1 D2A < cyto	10C	40, 41, 83, 100, 164, 175, 193, 203, 207, 240, 245, 256
D2B < D1 D2B < cyto	10D	36, 37, 40, 68, 85, 134, 203, 240, 245, 250, 256

4.4.3 Fast evolving sites around the catalytic center

The comparison of evolutionary site rates of the membrane-distal domains D2B versus the active membrane-proximal domains and the cytosolic domains revealed 15 sites, which are fast evolving in domain D2B, but conserved in the other domain subclasses (table 3, figure 10B). Most of these are located around or within the catalytic center. For example, one of the sites (260) is, in the corresponding active domain, occupied by the catalytic aspartate and two other sites (49 and 185) are involved in substrate binding. Since domain D2B has accumulated various substitutions in its functional sites and has lost catalytic activity, it is expected that the selective constraint on the catalytic center was relaxed, which consequently allowed the surrounding sites to evolve at a faster rate. The fact that these sites are not only fast evolving in the inactive domain D2B but also are conserved in the active domains in addition with their appropriate location on the surface of the domain, suggests their role in substrate binding in the active domains.

The evolutionary events in D2B after domain duplication might have been triggered by a single mutation of a functional residue, which led to a non-catalytic domain. Subsequently, the evolutionary constraint of the catalytic center was relaxed, leading to the accumulation of mutations in the surrounding.

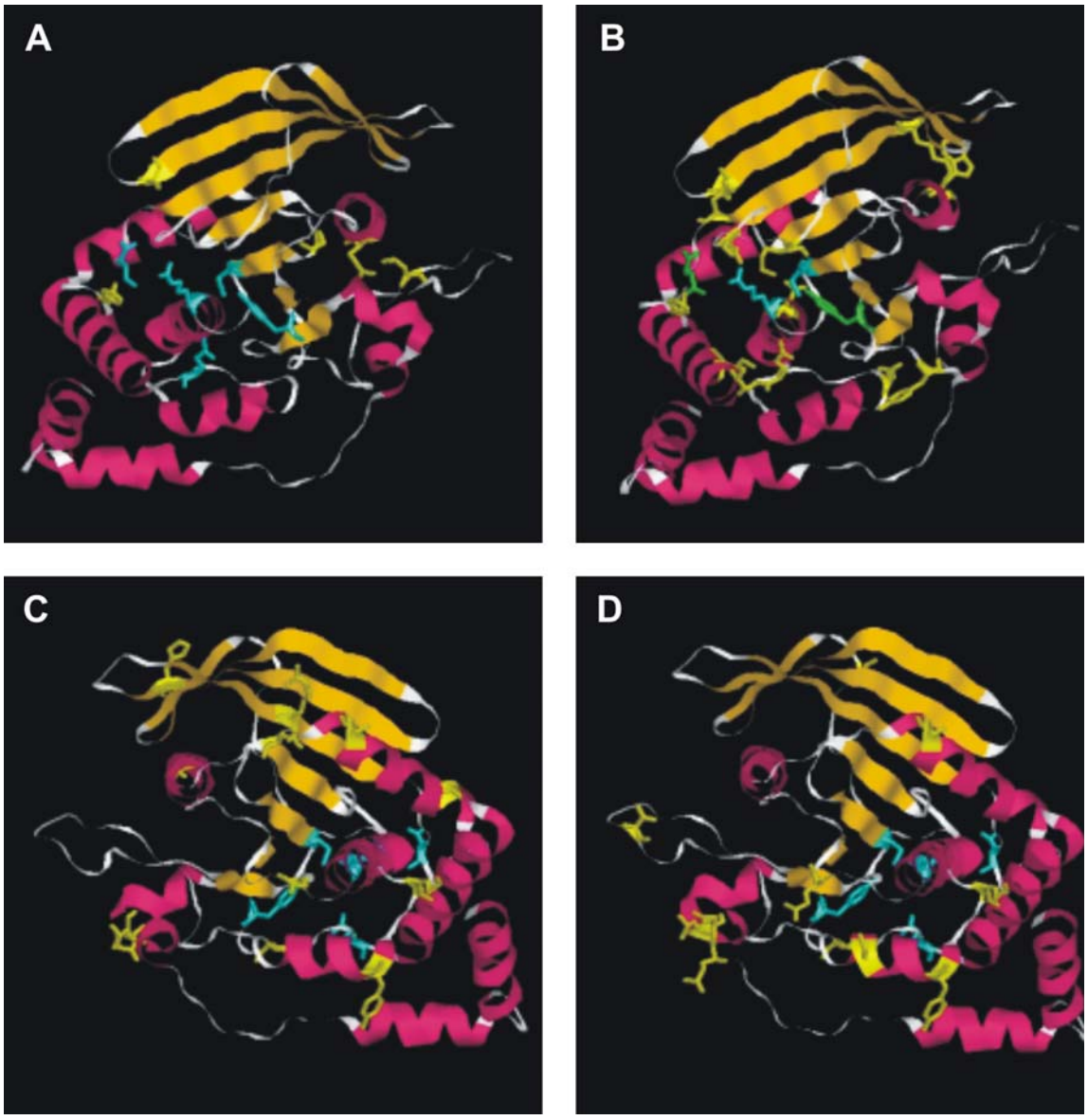


Figure 10: Sites with altered evolutionary rates mapped onto the tertiary structure

Functional sites and sites with altered evolutionary rates were mapped onto the pdb structure 1LAR. Catalytic and substrate binding residues are colored blue, residues at sites with altered evolutionary rates yellow, and residues, which belonging to both categories are colored green. A and B show a view on the catalytic center of the domain, C and D show the backside of the domain. A. Sites fast evolving in D2A but conserved in the active domains. B. Sites fast evolving in D2B but conserved in the active domains. C. Sites more conserved in D2A than in the active domains. D. Sites more conserved in domain D2B than in the active domains.

In contrast to domain D2B there are only 5 sites that are fast evolving in D2A but conserved in the active domains (figure 10A). One of them (position 185) is located next to the functional site 184, occupied by aspartate in the active domain, but replaced by glutamate in domain D2A. This mutation might have freed the immediate surrounding

from selective constraint and allowed a faster evolutionary rate at site 185. The functional residues in the catalytic center of D2A are affected by only two amino acid substitutions. The catalytic aspartate is replaced by glutamate acid and the substrate binding tyrosine either by valine or leucine. These substitutions maintain the biochemical properties and, although the catalytic activity of this domain is barely detectable, it is still able to stably bind its substrate (BLISKA *et al.* 1992). The analysis demonstrates, that the catalytic center is still under selective pressure, because residues that are fast evolving in D2B and predicted to function in substrate binding, are conserved in D2A. This is confirmed experimentally by regaining a catalytically fully active domain if the two substituted functional residues are converted to their original amino acids (LIM *et al.* 1998). We conclude that the catalytic center of D2A in contrast to D2B plays a pivotal role in the function of tandem domain phosphatases, leading to the question what this function is. As the domain has lost its catalytic activity but still can bind to phosphotyrosine, one could assume two complementary molecular functions. First, the domain could function as 'anti-phosphatase' as described for DSPs (CUI *et al.* 1998). Second, it might work as an adaptor domain for phosphotyrosine substrates, similar to SH2 and PTB domains, revealing an additional function of the PTP domain family.

4.4.4 Slow evolving sites on the backside of the domain

The complete loss of evolutionary pressure on the catalytic center of the D2B family leads to the question, what the function of this domain is and whether there is a similar role of the D2A domains. If a new function was acquired, this should be reflected in novel conserved sites. Therefore, we searched for sites that evolve at a higher rate in the active domains while they are conserved in D2A or D2B. The comparison found 11 sites conserved in D2B (figure 10D) and 12 sites conserved in D2A (figure 10C), of which 5 sites are found in both analyses (Table 3). Almost all sites are located on the surface of the 'backside' of the domain. This could indicate that a new functional center has evolved in this region. The solvent exposure as well as the nature of the conserved amino acids might hint that this region is involved in protein-protein interactions. Indeed, interactions of the membrane-distal and proximal domains have been described recently. The direct interaction of the membrane-distal domain with the membrane-proximal domain stabilizes the enzyme and enhances catalytic activity (FELBERG and JOHNSON 2000). This effect can be abrogated by deletion of the two carboxy-terminal alpha helices of the membrane-distal

domain (244-278) (JOHNSON *et al.* 1992). These two helices host two residues (245 and 256), which are significantly more conserved in D2A and D2B than in the active domains and one residue (250) that is more conserved if D2B is compared against the active domains. Another highly conserved site (240) found in both comparisons is preceding the two helices. These sites might play an important role in the interaction between the phosphatase domains. The additional sites with altered selective constraint might contribute to the stable binding, but are not sufficient for stable interaction without presence of the two carboxy-terminal helices. Experimental mutation of these sites might give further insight into the molecular mechanism of regulation of RPTPs.

Together with the variation of the catalytic site, our results indicate, that domain D2A and D2B have distinct influences on the activity of membrane-proximal domains. We suggest that both domains can control activity of the first domain by interaction of residues from the 'backside' of the membrane-distal domain and residues from the membrane-proximal domain. In addition D2A can regulate substrate specificity of the membrane-proximal domain and remain associated with the substrate protein, which is accomplished by the inactive catalytic center of D2A.

The results of our analyses allow delineating a possible scenario for the evolution of the membrane-distal domain of RPTPs. The overlap of conserved residues on the ‘backside’ of both, D2A and D2B, indicates that their common ancestor already had evolved this novel functional site. Whether the membrane-distal domain of the first RPTP was still active remains unclear, but the complete absence of a domain without substitutions of functional residues within the catalytic center hints, that it indeed was inactive. This ancestral RPTP gave rise to one lineage with a conserved catalytic center that is still able to bind substrate (D2A) and one lineage that accumulated substitutions around the catalytic center, completely loosing the substrate binding function (D2B).

4.4.5 Conclusions

Our analysis of the PTP family illustrates how a closely related domain family can evolve multiple different molecular functions. On the catalytic site, the family varies in the specificity of substrates, allowing the dephosphorylation of a wide range of phosphoproteins as well as phosphoinositides (MAEHAMA and DIXON 1998). Loss of catalytic activity opens the opportunity to evolve novel functions at the catalytic site. In

single domain phosphatases, this event led to the evolution of proteins antagonizing phosphatase function, the anti-phosphatases. The inactive domain in tandem domain phosphatases functions in substrate recognition and binding to multiple phosphorylated proteins. Here, it might work as a competitor for other phoshpotyrosine binding domains as SH2 and PTB. In addition to these changes within the catalytic site, loss of catalytic activity also enabled the evolution of a novel functional site within the domain and specialization on regulatory functions. In summary, our analysis shows how evolution can create novel functionality based on an existing, well-adapted enzyme, illustrating the versatility of the PTP family.

5 Inactive Enzyme-Homologues find new Function in Regulatory Processes

5.1 Summary

Although the catalytic center of an enzyme is usually highly conserved, there have been a few reports of proteins with substitutions at essential catalytic positions, which convert the enzyme into a catalytically inactive form. Here, we report a large-scale analysis of substitutions at enzymes' catalytic sites in order to gain insight into the function and evolution of inactive enzyme-homologues. Our analysis revealed that inactive enzyme-homologues are not an exception only found in single enzyme families, but that they are represented in a large variety of enzyme families and conserved among metazoan species. Even though they have lost their catalytic activity, they have adopted new functions and are now mainly involved in regulatory processes, as shown by several case studies. This modification of existing modules is an efficient mechanism to evolve new functions. The invention of inactive enzyme-homologues in metazoa has thereby led to an enhancement of complexity of regulatory networks.

5.2 Results and Discussion

The catalytic center of an enzyme is usually highly conserved to maintain its ability for efficient catalysis. A mutation in the catalytic center would not only destroy the sensitive arrangement of catalytic and substrate binding residues and suggest loss or diminishment of catalytic activity, but also lead to a decreased fitness of the organism. Thus, the catalytic center of an enzyme is subject to high selective pressure. If selective pressure is lost, for example due to an inactivating mutation, the corresponding non-functional gene would very likely be lost during the process of evolution. However, several enzyme-homologues have been observed that carry substitutions at sites corresponding to catalytic sites in the active enzyme (WISHART *et al.* 1995; CUI *et al.* 1998). Recently, we discussed inactive subclasses of protein tyrosine phosphatases, which are involved in

regulatory tasks (PILS and SCHULTZ 2004).These inactive phosphatases substituted at least one essential catalytic amino acid with a non-functional one. According to the conservation of the catalytic center, two subclasses can be distinguished. One subclass of inactive phosphatases has preserved its ability to specifically bind substrate and competes with the active phosphatases for phosphorylated tyrosine residues. Due to loss of catalytic activity, the substrate is trapped in the binding pocket and protected from other active phosphatases. The other subclass, found in tandem domain phosphatases, has evolved a new regulatory center and functions as regulatory subunit of the adjacent active phosphatase domain.

Recently, Bartlett and colleagues described three additional pairs of active and inactive enzyme-homologues (BARTLETT *et al.* 2003). In two cases, these inactive enzyme-homologues function in regulation. The SOD1 copper chaperone (CCS) is very similar to the enzyme superoxide dismutase (SOD1) in sequence and structure, but CCS has lost its ability to bind metal and its catalytic activity. Instead, CCS has specialized on regulating the activity of SOD1 (LAMB *et al.* 2000). Phospholipase A2 (PLA2) and its inhibitor represent an analogous case. Both proteins share a common ancestor, but the PLA2 inhibitor has lost one of the three essential catalytic amino acids. Correspondingly, it evolved a new function to support the unstable conformation of phospholipase A2 by forming a heterodimer with the enzyme. In the bound form the PLA2 inhibitor covers the catalytic center of the enzyme and can regulate the activity of phospholipase A2 (DEVEDJIEV *et al.* 1997).

These anecdotal reports of inactive enzyme-homologues raised the question whether there is a general trend for recruiting inactive enzyme-homologues for regulating their active counterparts. To address this, we systematically investigated substitutions at catalytic sites of a large set of enzymatic sequences. A substantial number of sequences can be scanned efficiently by annotating conserved protein domains with essential catalytic positions. We achieved this by transferring the annotation of a single enzymatic sequence onto the Hidden Markov Model (HMM) of a domain, which represents a large number of family members. Using the annotated HMM, all sequences represented by the HMM can be rapidly scanned for the presence of catalytic amino acids. Essential catalytic amino acids were manually extracted from reports on the catalytic mechanism based on structural analysis and represent amino acids that interact directly or indirectly with the substrate

during enzymatic catalysis, thus are indispensable for the full function of the enzymatic domain (see legend of table 6 for details).

As a source of conserved domains, we used the SMART[11] database (LETUNIC *et al.* 2004) and extracted all 92 enzymatic domains. The catalytic mechanism and hence the catalytic sites could be assigned for 52 of these domains from the primary literature (table 9). For the remainder, no catalytic sites could be assigned for the following reasons. Firstly, a rather large portion of enzymatic domains do not include any well-characterized sequences or lack a 3D structure that gives information on the sterical orientation and function of the amino acid residues. Secondly, some domains contain subfamilies, which use distinct catalytic strategies, and different amino acids at different positions are involved in the distinct mechanism, so that the catalytic positions are not conserved throughout the family, which was one of our criteria for a catalytic residue. Of these 52 annotated enzymatic domains, the 47 domains, which are present in metazoa, built the core set of our analysis.

Currently, most methods for the distinction of functionally divergent subgroups within one domain family rely on phylogenetic information (SJOLANDER 2004). But, as the activity of an enzyme depends on a few residues, inactive enzymes can, as we have shown for the tyrosine phosphatases (PILS and SCHULTZ 2004), evolve multiple times within one family. Therefore, this functional classification does not follow phylogeny, leading to numerous mis-annotations. Using the site-specific annotations presented here can eliminate this difficulty and should be of general use for the automatic prediction of a protein's function. Details of all annotations are available[12] (table 9).

Taking advantage of the high conservation around the active center, catalytic positions could easily be transferred to the HMM and back to any other family member. However, for several sequences we found gaps at catalytic positions and the alignment showed deletions of exon-sized blocks. As this suggests errors in the gene prediction process rather than reflect evolutionary deletion, we omitted these sequences from our analysis. Table 5 shows the number of sequences, which contain gaps at catalytic positions for each of the analysed proteomes, and which represent non-functional domains at this stage. An especially high portion of non-functional sequences in relation to the proteome

[11] http://smart.embl-heidelberg.de/
[12] http://www.biozentrum.uni-wuerzburg.de/bioinformatik/projects/inactive_enzymes.html

was observed in *Anopheles gambiae* and *Fugu rubripes*. This might be caused by the lack of specially trained gene prediction programs for these species or by lower sequence coverage.

Table 6: Enzymatic domains in metazoan genomes.

Whole genomes of Caenorhabditis elegans (version 18.102.1), Anopheles gambiae (version 18.2a.1), Drosophila melanogaster (version 18.3a.1), Fugu rubripes (version 18.2.1), Mus musculus (version 18.30.1), Rattus norvegicus (version 18.3.1) and Homo sapiens (version 18.34.1) were retrieved from Ensembl (http://www.ensembl.org/). Each peptide set was searched for the presence of the 47 enzymatic SMART domains with known catalytic sites with HMMsearch (HMMER: http://hmmer.wustl.edu/) using an HMM built from the family multiple sequence alignment retrieved from SMART, and a domain-specific E-value threshold as described by Schultz et al. (SCHULTZ *et al.* 1998). To avoid bias caused by alternatively spliced transcripts, only the best scoring hit per gene was used for further analysis. Additionally, sequences had to cover the complete HMM, allowing an unmatched region of 10 amino acids at the beginning and end of the HMM versus the sequence alignment. Knowing the catalytic positions in the HMM, the sequences obtained by the HMMsearch could easily be scanned for presence of all catalytic essential amino acids. The table shows the results of scanning metazoan proteomes for 47 annotated HMMs. Given are the amount of non-functional domains, the amount of active domains and the amount of inactive domains due to substitutions at catalytic sites, as well as the percentage of inactive domains from all functional (active and inactive) domains. Non-functional sequences contain gaps at catalytic positions and were excluded from the further analysis. Here, lack of conservation in the catalytic region is presumably caused by inaccurate gene prediction.

Species	Non-functional Domains	Active Domains	Inactive Domains	% inactive
Caenorhabditis elegans	15	634	114	15 %
Anopheles gambiae	67	636	116	15 %
Drosophila melanogaster	27	683	119	15 %
Fugu rubripes	82	788	102	11 %
Mus musculus	39	997	128	11 %
Rattus norvegicus	23	948	109	10 %
Homo sapiens	17	1013	131	11 %

5.2.1 Inactive enzyme-homologues are the rule, not the exception

If an amino acid was present in the predicted catalytic site, we checked whether it was of the type found in catalytic active members or whether there was a substitution, indicating a loss of activity. These substitutions do not necessarily turn off catalytic activity completely, but even minor changes in the amino acid composition can highly reduce catalytic activity so we assume sequences with substitutions at catalytic positions are inactive domains. Table 7 shows the numbers of active domains and domains with substitutions at catalytic positions for each domain and species investigated. As a major result, it turns out that inactive enzyme-homologues are present in all of the investigated species and in most of the enzymatic domains. In contrast to metazoa, yeast only contains a negligibly small number of inactive domains (data not shown). The phylogenetic consistency in metazoa (Table 6) argues for the invention of inactive enzymatic domains simultaneously with the evolution of metazoan, possibly indicating an increased complexity of regulatory networks.

Table 7: Amount of active/inactive enzymatic domains in metazoan genomes (next page)

An annotated HMM containing the catalytic sites was generated for each SMART domain in order to scan for presence of active and inactive domains in metazoan genomes. For the identification of catalytic sites, sequences containing the here investigated SMART domains were derived and the primary literature linked to these sequences was inspected. Described catalytic positions, usually from structure reports, were assigned to the corresponding sequence. Our criteria for a catalytic amino acid residue were direct or indirect involvement in the catalytic mechanism and selection was very parsimonious to avoid false positives in our analysis. A direct effect on catalysis can be as acid or base catalyst, as nucleophilic agent forming a covalent bond with the substrate or as stabilisator of the transition state. Ranked among indirect effects are binding of metal ions, which are directly involved in catalysis or electrostatic effects, needed for example to lower the pKa of adjacent residues and increase the reactivity. Carbonyl or amino groups of the peptide backbone are sometimes described as catalytic groups, as for example in the serine proteases, where the amino groups of serine and glycine form the oxyanion hole and stabilize the transition state. Still, we did not consider these residues in our analysis since they are arbitrarily exchangeable and mutation does not necessarily affect the activity of the enzyme. Thus, selection of catalytic sites was very conservative, choosing only amino acids, which are unambiguously described as playing an important role in the catalytic mechanism and whose mutation would have a dramatic effect on the enzyme's catalytic activity. Sequences with known catalytic sites were aligned to the domain specific HMM using HMMalign (HMMER) and the positions of the catalytic sites were transferred to the HMM position, generating a set of annotated HMMs of enzymatic SMART domains. After detection of a catalytic domain in a protein, this strategy could be reversed and the position of the catalytic sites could be transferred onto its sequence. To avoid the annotation of domain families with subfamilies using distinct catalytic mechanisms, catalytic positions which were not conserved in the HMM, characterized by a probability of less than 0.5, were excluded from further analysis. Ensembl identifiers of inactive enzyme-homologues are available at http://www.biozentrum.uni-wuerzburg.de/bioinformatik/projects/inactive_enzymes.html.

	Domain	Description	*C. e.*	*A. g.*	*D. m.*	*F. r.*	*M. m.*	*R. n.*	*H. s.*
Signalling Domains	acidPPc	Acid phosphatase homologues	4/1	6/2	8/2	7/9	8/9	8/8	7/9
	CASc	Caspase, interleukin-1 beta converting enzyme homologues	3/0	10/0	5/0	6/0	8/1	8/0	11/1
	CYCc	Adenylyl- / guanylyl cyclase	29/13	19/8	25/13	21/15	18/9	17/9	18/11
	DSPc	Dual specificity phosphatase	10/1	5/2	6/2	25/2	20/6	23/5	22/5
	G_alpha	G protein alpha subunit	20/1	7/0	7/0	18/0	16/0	15/0	17/0
	GuKc	Guanylate kinase homologues.	6/2	6/2	7/2	26/3	19/1	19/2	22/2
	HECTc	Domain Homologous to E6-AP Carboxyl Terminus	9/0	14/0	14/0	20/0	23/0	23/0	26/0
	LMWPc	Low molecular weight phosphatase family	0/0	1/0	2/1	0/0	4/0	3/0	1/0
	PLAc	Cytoplasmic phospholipase A2	0/0	0/0	0/0	1/0	1/0	2/0	5/0
	PLCXc	Phospholipase C, domain X	5/1	3/0	3/0	15/4	10/2	11/1	12/2
	PLDc	Phospholipase D. Active site motifs	7/2	5/0	6/0	8/1	8/2	6/2	7/2
	PP2Ac	Protein phosphatase 2A homologues	41/3	9/0	17/0	13/0	15/0	14/0	13/0
	PTPc	Protein tyrosine phosphatase	37/48	11/9	11/8	27/15	23/15	29/15	29/16
	RAB	Rab subfamily of small GTPases	25/2	23/0	23/7	56/1	50/1	53/2	61/3
	RAS	Ras subfamily of RAS small GTPases	6/3	5/5	5/5	7/9	13/12	12/9	14/9
	RHO	Rho (Ras homology) subfamily of Ras-like small GTPases	7/0	6/0	6/0	18/6	14/5	13/4	17/4
	RHOD	Rhodanese Homology Domain	11/9	5/2	6/2	5/10	9/15	9/14	8/16
	S_TKc	Serine/Threonine protein kinases	124/3	90/4	97/4	164/6	227/7	241/5	238/8
	TyrKc	Tyrosine kinase	60/3	24/0	28/0	69/2	71/1	74/1	83/1
	UBCc	Ubiquitin-conjugating enzyme E2, catalytic domain homologues	21/4	23/4	27/5	37/3	65/11	35/6	36/7
Nuclear Domains	35EXOc	3'-5' exonuclease	1/3	2/0	5/0	1/0	2/1	2/1	1/1
	CPDc	ctd-like phosphatases	4/1	5/0	6/3	9/1	8/1	4/0	7/1
	EXOIII	exonuclease domain in DNA-polymerase α and ε chain, ribonuclease T, etc.	10/0	8/0	7/0	4/0	11/1	12/1	17/0
	GIYc	GIY-YIG type nucleases (URI domain)	1/0	1/0	1/0	0/0	0/0	1/0	0/0
	POLAc	DNA polymerase A domain	2/0	1/0	2/0	2/0	3/0	3/0	3/0
	POLBc	DNA polymerase type-B family	5/0	4/0	4/0	3/0	4/0	4/0	4/0
	POLXc	DNA polymerase X family	0/0	0/0	0/0	3/0	4/0	4/0	4/0
	TOP1Ac	Bacterial DNA topoisomerase I DNA-binding domain	2/0	3/0	2/0	2/0	2/0	1/0	2/0
	TOP4c	DNA Topoisomerase IV	3/1	2/0	1/0	2/0	2/0	1/0	2/0
	TOPEUc	DNA Topoisomerase I (eukaryota)	1/0	1/0	1/0	3/0	2/0	3/0	2/0
	TOPRIM	topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins	2/0	3/2	2/0	1/0	2/0	2/0	1/0
Extracellular Domains	Aamy	Alpha-amylase domain	0/0	1/0	7/0	0/0	0/0	0/0	0/0
	CysPc	Calpain-like thiol protease family	10/2	5/0	3/1	11/0	11/1	11/1	10/1
	DNaseIc	deoxyribonuclease I	0/0	0/0	0/0	2/1	4/0	4/0	4/0
	LYZ1	Alpha-lactalbumin / lysozyme C	0/0	3/2	11/4	1/0	3/4	4/5	4/5
	NUC	DNA/RNA non-specific endonuclease	1/0	1/4	2/5	2/5	1/4	1/4	1/5
	PA2c	Phospholipase A2	2/0	3/3	1/3	3/1	10/1	10/1	7/1
	Pept_C1	Papain family cysteine protease	21/7	6/2	8/2	9/2	19/3	15/2	11/1
	RNAse_Pc	Pancreatic ribonuclease	0/0	0/0	0/0	0/0	7/1	11/0	8/0
	Tryp_SPc	Trypsin-like serine protease	8/3	168/65	167/50	47/6	107/9	86/6	99/14
	TSPc	tail specific protease	0/0	0/0	0/0	0/0	1/3	1/3	1/3
	ZnMc	Zinc-dependent metalloprotease	43/1	17/0	15/0	32/0	26/0	21/0	27/0
Other Domains	alkPPc	Alkaline phosphatase homologues	0/0	9/0	12/0	2/0	5/0	3/0	4/0
	NDK	nucleoside diphosphate kinase	1/0	2/0	3/0	8/0	11/2	11/1	11/1
	PlsC	Phosphate acyltransferases	15/0	14/0	11/0	14/0	15/0	17/0	15/0
	PSN	Presenilin/signal peptide peptidase family	5/0	3/0	3/0	5/0	6/0	6/0	7/0
	TGc	Transglutaminase/protease-like homologues	1/1	3/0	2/0	10/0	7/0	8/1	10/1

Some domains present an unexpectedly high ratio of active versus inactive domains. These domains are primarily used in signalling pathways, but surprisingly, there are also several domains secreted to the extracellular compartment. The group with the least inactive sequences is formed by the nuclear domains. This could be caused by the fact that enzymatic domains found in the nucleus mainly function in DNA/RNA synthesis, but are not involved in the regulation of these processes. A substitution of a catalytic amino acid in a nuclear domain involved in such fundamental processes is certainly deleterious and unlikely to be fixed in evolution.

5.2.2 Inactive signalling enzymes

Within the analysed domains, those involved in signalling processes revealed the highest number of inactive homologues. To gain insight into their molecular and cellular function, in the following we discuss some families with an unexpected high number of members with substitutions at catalytic sites. Probably most outstanding is the family of small GTPases. Of all Ras proteins found, almost 50 % carry a substitution at the position of the catalytic glutamine, which plays a central role in the mechanism of GTP hydrolysis. Within the Rho family, the inactive proteins account for 25%. Although mutations of small GTPases often lead to transforming behaviour of the cell, there are some experimental studies proving that the substitution of the catalytic amino acid investigated by us does not lead to oncogenic properties, but instead acts antagonistically towards the active form of the enzyme. Ras-family members with the catalytic glutamine substituted by threonine for example can inhibit the growth-stimulating property of Ras, if overexpressed (KITAYAMA *et al.* 1989). Analogous, the Rho family, known to organize the formation of the cytoskeleton, includes members with a glutamine to serine substitution at the catalytic position, which leads to the disassembly of the actin filament structures, while active Rho proteins promote the actin assembly and cell adhesion (NOBES *et al.* 1998). Thus, inactive Rho family members could function as negative regulators.

The sulfurtransferse domain family (RHOD) also contains about 50% of inactive members. In almost all cases the catalytically essential cysteine is replaced by aspartic acid. The maintenance of selective pressure on the active center suggests that the domain developed a new function carried out by the former catalytic site. RHOD domains are very often combined with phosphatase domains or occur in tandem, so that it is likely that the

inactive RHOD domain modulates activity of the active domain or functions in substrate recruitment (BORDO and BORK 2002).

5.2.3 Inactive extracellular enzymes

Extracellular enzymes also present a high number of sequences with substitutions at catalytic sites, although to a lesser extent than signalling domains. An unexpected case is the trypsin-like serine protease family (Tryp_SPc), as proteases are not involved in any regulation tasks. Still, we find inactive proteases caused by various substitutions in all of the investigated species. Interestingly, the amount of trypsin-like serine proteases indicates a gene expansion in *Anopheles gambiae* and *Drosophila melanogaster*, including both active as well as inactive proteases. Inactive proteases might have developed a new function advantageous to insects. Possibly, these inactive proteases compete with active proteases for specific substrate and antagonize the effect of the active enzymes.

Another protease family with inactive domains are the tail-specific proteases (TSPc). The four domains found in *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* are all expressed from the same gene, but only one domain remained catalytically active. In all three species the catalytic activity resides in the third domain. The function of the highly conserved three other domains is still to be elucidated, but it is likely that the tail specific proteases represent a case similar to the protein tyrosine phosphatases, in which a second inactive domain can regulate activity of the catalytic active domain.

The lysozyme (LYZ) domain comprises two families with a common evolutionary origin, lysozyme and lactalbumin. Only lysozyme is catalytically active in the hydrolysis of polysaccharides. Lactalbumin, responsible for milk production in mammals, has no known catalytic activity and representatives account for the high number of inactive members in this family. Instead, lactalbumin forms a regulatory subunit of lactose synthase and also changes the substrate specificity of galactosyltransferase in favour of glucose (QASBA and SAFAYA 1984).

5.2.4 Are inactive enzyme-homologues encoded by pseudogenes?

The examples above demonstrate that enzymes can be integrative parts of cellular networks even after having lost their catalytic activity. Still, the high number of inactive

homologues within all genomes came as a surprise and could be caused by pseudogenes instead of actively expressed ones. Pseudogenes can arise from gene duplication events, which lead to one gene copy freed from selective pressure and allow the adoption of new functions (MIGHELL *et al.* 2000). In case of mutations in the catalytic center, the duplicated gene would lose its catalytic activity and, following genetic drift, would accumulate more mutations, if it does not present a selective advantage for the organism. To exclude that inactive enzyme-homologues are encoded by non-functional pseudogenes we retrieved their putative orthologues wherever possible from Ensembl[13]. Table 8 lists the number of orthologues found for inactive enzyme-homologues in related species and the numbers of inactive orthologues. Since the assignment of orthologues is based on best reciprocal blast hits, an orthologue could not always be identified and even if so, the orthologue might not be the complete gene and might lack functional parts. Despite these difficulties, a large number of orthologues are also catalytically inactive enzyme-homologues. The fact that inactive enzyme-homologues are well conserved between *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Fugu rubripes* and also between *Drosophila melanogaster* and *Anopheles gambiae* strongly argues against pseudogenes, because these would have been lost during the time passed since the last phylogenetic split between the species, and demonstrates that inactive enzyme-homologues are expressed from functional genes.

[13] http://www.ensembl.org/

Table 8: Orthologues of genes encoding inactive domains in related species.

Orthologues were retrieved from the Ensembl database and represent best reciprocal BLAST hits. The orthologous sequences were then examined for the presence of inactive domains. Orthologous genes might lack some domains if the orthology assignment is based on a local alignment, resulting in unequal numbers of orthologous genes (column 4) and orthologous genes containing inactive domains (column 5). For example, the human gene ENSG00000145348 encodes three domains, a protein kinase domain at the N-terminus, then a TBC domain and a C-terminal catalytically inactive Rhodanese domains. The orthologous mouse gene ENSMUSG00000028030 contains only the two N-terminal domains and is lacking the Rhodanese domain, and the difference of these orthologous genes presumably originates from incomplete gene prediction. Caused by genes encoding multiple inactive domains, the amount of genes encoding inactive domains does not match the total amount of inactive domains from table 5.

	Genes encoding inactive domains		Orthologues	Orthologues encoding inactive domains
		M. m.	117	97
Homo sapiens	126	***R. n.***	109	89
		F. r.	84	57
		H. s.	113	98
Mus musculus	125	***R. n.***	104	90
		F. r.	81	59
		H. s.	100	89
Rattus norvegicus	106	***M. m.***	100	90
		F. r.	76	56
		H. s.	72	63
Fugu rubripes	101	***M. m.***	69	58
		R. n.	68	57
Anopheles gambiae	115	***D. m.***	61	52
Drosophila melanogaster	118	***A. g.***	63	47

Table 9: Catalytic sites of SMART domains extracted from the literature.

Domain	Catalytic amino acid and position in HMM
PLAc	S-134
G_alpha	R-194, Q-227
PlsC	H-7, D-12
RAS	K-16, Q-61
CASc	H-89, C-139
PLCXc	H-16, H-92
PSN	D-165, D-273
TOPRIM	E-7, D-118, D-120
POLXc	D-206, D-208, D-321
TGc	C-9, H-65, D-87
TSPc	S-140
35EXOc	D-27, D-96, D-186
LIGANc	K-117
UBCc	C-82
Aamy	D-346, D-535
RAB	Q-59
Glyco_32	D-11
CysPc	C-87, H-279, N-312
HECTc	C-371
DSPc	D-75, C-113, S-120
TOP1Ac	Y-54
Pept_C1	Q-22, C-31, H-244
alkPPc	S-61
PTPc	D-199, C-250, Q-291
RHO	Q-57
NUC	D-84, H-87, E-128
Glyco_10	E-116, E-252
DNaseIc	E-97, H-173
NDK	H-119
PA2c	H-48, D-94
GuKc	D-105
EXOIII	E-8, H-174
TOP4c	Y-116
CYCc	N-190, R-194
LMWPc	C-7, R-13, D-124
RNAse_Pc	H-12, K-44, H-135
S_TKc	E-39, D-119, K-121
TOPEUc	Y-466
RHOD	C-56
LYZ1	E-35, D-53
acidPPc	H-82, H-141, D-145
Tryp_SPc	H-36, D-81, S-185
53EXOc	K-80
POLBc	D-353, D-521, D-523
PLDc	H-6, D-13
CPDc	D-8, D-10
GIYc	R-26
POLAc	D-22, D-237, E-238
ZnMc	H-138, H-142, H-148
PP2Ac	D-74, H-104
TyrKc	D-125

5.2.5 Concluding remarks

We performed a large-scale analysis of substitutions at catalytic positions of enzymatic domains, which resulted in the detection of numerous inactive enzyme-homologues. Although we cannot exclude that the detected proteins possess a catalytic rest activity, our criteria were carefully and very conservatively chosen to minimize false positives. A classification by function and subcellular localization revealed that catalytically inactive domains are very often found in signalling pathways, less in extracellular functions and hardly in nuclear processes. This can be explained by the vital importance of nuclear enzymes in contrast to signalling enzymes. In nuclear enzymes, lack of enzymatic activity would be fatal, while signalling enzymes can more easily tolerate an inactivating mutation and use the new protein variant to increase the repertoire of signalling modules. A closer look at the different inactive enzyme families has shown that they are mainly involved in regulatory processes. This can be achieved directly, by modulating the activity of the homologous active domain as a regulatory subunit or indirectly by antagonizing the cellular effects of the active domain. In the latter case, it is likely that the former catalytic center is still under selective pressure, so that it can mimic the original active center and bind specifically to its substrate. In accordance to the trend observed in this work, we suggest that the inactive enzyme-homologues of unknown function are also involved in regulation processes. Usage of inactive enzyme-homologues for regulatory tasks shows that evolution has been very economic in creating new regulatory control elements. Choosing from the available repertoire of domains and replacing just one single amino acid at the right position can be sufficient to reverse the effect of the original domain, as it has been described for the phosphatases. In this way, the evolution of inactive enzyme-homologues added a new level of complexity to regulatory networks. Inactive enzyme-homologues can regulate other enzymes, modulate existing signalling pathways and they increase the network of possible protein-protein interactions within the cell. The number of inactive enzyme families and their phylogenetic conservation demonstrates that this is a major principle in regulation. Although this mechanism has been underestimated in the past, it proves the importance of inactive enzyme-homologues in regulatory processes.

6 Variation in structural location and amino acid conservation of functional sites in protein domain families

6.1 Abstract

6.1.1 Background

The functional sites of a protein present important information for determining its cellular function and are fundamental in drug design. Accordingly, accurate methods for the prediction of functional sites are of immense value. Most available methods are based on a set of homologous sequences and structural or evolutionary information, and assume that functional sites are more conserved than the average. In the analysis presented here, we have investigated the conservation of location and type of amino acids at functional sites, and compared the behaviour of functional sites between different protein domains.

6.1.2 Results

Functional sites were extracted from experimentally determined structural complexes from the Protein Data Bank harbouring a conserved protein domain from the SMART database. In general, functional (i.e. interacting) sites whose location is more highly conserved are also more conserved in their type of amino acid. However, even highly conserved functional sites can present a wide spectrum of amino acids. The degree of conservation strongly depends on the function of the protein domain and ranges from highly conserved in location and amino acid to very variable. Differentiation by binding partner shows that ion binding sites tend to be more conserved than functional sites binding peptides or nucleotides.

6.1.3 Conclusions

The results gained by this analysis will help improve the accuracy of functional site prediction and facilitate the characterization of unknown protein sequences.

6.2 Background

Protein function is determined by the spatial configuration and type of amino acids at functional sites. Knowledge of functional sites provides valuable information for the assignment of molecular function, potential physiological binding partners and hence drug design. Tasks performed by functional sites range from the binding of small molecules like ions, cofactors, metabolic substrates or high molecular weight compounds such as nucleic acids and peptide chains, to catalysing chemical reactions in the active centre of enzymes.

The exponentially growing number of uncharacterised protein sequences in the public databases has turned the development of automatic identification of functional sites into an important research field and many computational methods focusing on this area have been described in recent years (for review see LICHTARGE and SOWA 2002; CAMPBELL *et al.* 2003; JONES and THORNTON 2004). In contrast to structural approaches that search for ligand binding pockets on the protein surface using molecular modelling (BHINGE *et al.* 2004; LAURIE and JACKSON 2005), network analysis (AMITAI *et al.* 2004), or compare the protein surface to structures with known interacting sites (STARK *et al.* 2004; KINOSHITA and NAKAMURA 2005), many methods are based on a set of homologous sequences combined with evolutionary or structural information. The evolutionary trace (ET) method (LICHTARGE *et al.* 1996; MADABUSHI *et al.* 2002), for example, searches for a structural cluster of conserved residues. Beginning with a sequence identity tree from a set of homologous proteins, the tree is scanned for subgroup-specific residues, which are invariant within the subgroup but vary between subgroups. These residues, called evolutionary trace residues, and the residues that are invariant in all sequences are then mapped onto a representative 3D structure and clusters of high ranking residues, corresponding to the inner nodes of the tree, are searched. These clusters usually coincide with the functional center of the protein. Improvements of the ET method use sequence weights based on their similarity (weighted evolutionary tracing) and an amino acid substitution matrix to account for biochemically similar amino acids in the identification of

the trace residues (LANDGRAF *et al.* 1999), they consider the evolutionary distance between proteins due to the phylogenetically biased databases (ARMON *et al.* 2001) or allow different rates of amino acid substitutions at protein sites (PUPKO *et al.* 2002). A similar approach is focusing more on structural information and calculates a conservation score at each position under consideration of the behaviour of spatial neighbours (LANDGRAF *et al.* 2001).

Most of the above mentioned methods assume that functional sites are under high selective pressure and conserved within the protein, so that functional sites can easily be detected by lower rates of amino acid substitutions. However, functional sites can vary in subfamilies and homologous protein sequences can perform different functions using a different set of functional residues. Accordingly, interaction interfaces can vary in their location in distant homologues and this has to be considered if interaction interfaces are inferred from homologous proteins (ALOY *et al.* 2003; REKHA *et al.* 2005).

Prior to their prediction, it is necessary to understand the arrangement and properties of functional sites, as well as how protein families and single sequences differ in their use. Effort towards this direction has been made by several groups, who studied physicochemical properties of protein-protein interaction interfaces found in homodimer, heterodimer or intra-chain domain complexes (JONES and THORNTON 1997; LIJNZAAD and ARGOS 1997; LARSEN *et al.* 1998; LO CONTE *et al.* 1999; VALDAR and THORNTON 2001), as well as protein-DNA interaction interfaces (LUSCOMBE and THORNTON 2002).

Here, we perform a large-scale analysis of functional sites extracted from experimentally determined protein-ligand complexes stored in the protein data bank (PDB) (DESHPANDE *et al.* 2005), grouped by the presence of conserved protein domains described in the SMART database (LETUNIC *et al.* 2004). The classification into protein families enables us to find differences in amino acid conservation and use of specific locations for functional sites between the families. The analysis shows that domains vary strongly in the conservation of interacting sites and provides useful information for the prediction of interacting sites based on homologous protein sequences.

6.3 Results and Discussion

Our analysis of interacting sites is based on conserved protein domains found in the structures of the protein data bank (PDB). Family sequence alignments of protein domains were retrieved from the SMART database and used to scan the protein sequences of the PDB with domain-specific Hidden Markov Models (HMMs). Wherever a domain was identified, all interactions between an amino acid belonging to this domain and any ligand were extracted and the position of the interacting amino acid was transferred onto the HMM consensus.

Table 10 lists the number of sequences with ligand interactions extracted from PDB and the number of interacting sites found in these sequences. The HMM consensus was used as a reference sequence to be able to compare the location of interactions among different sequences of a domain family. The positions in the consensus sequence correspond to positions shared by most sequences of the domain family (match states) and most domains interact only in regions for which an HMM match state exists. In contrast, if the interacting position corresponds to an insert state in the alignment to the HMM consensus, the information of interaction could not be transferred to the domain consensus. There are several domains, which have a comparatively large number of interacting sites located in loop regions. These amino acid sites are only present in subfamilies of the domain and are lacking an HMM match state, so that mapping the information of interaction onto the HMM was not possible. An overview of the number of interactions at HMM insert states for all investigated protein domains is given in table 10. In domains of the immunological system, e.g. the immunoglobulin (IG) and immunoglobulin V type (IGv) domains, up to 50% of the interacting sites are located in these regions. Strikingly, other extra-cellular domains (fibronectin type 3 (FN3), C-type lectin (CLECT), leucine rich repeat C-terminal domain (LRRCT)) also present a large number of interacting sites in loop regions. These domains all have a common involvement in highly specific recognition processes, where any restrictions in the choice of amino acid or structural constraints would be disadvantageous for the function of the domain. The loop regions without any match states exactly fulfil this condition and therefore biochemical properties of the domain can be fine-tuned to complement the appearance of the ligand. The increased use of variable loop regions for functional sites seems to be a common characteristic of extracellular highly ligand-specific domains.

6.3.1 Conservation of location of interacting sites

In order to compare the use of specific interacting positions within a domain family we introduced a score (ConsInt) to measure the conservation of interaction at a given site in the family alignment. The score reflects the importance of an amino acid site in domain interactions and ranges between 0 and 1. Scores greater than 0 mean that at least two non-identical sequences interact at this position, and a score of 1 arises if all sequences interact at this position. For the comparison of the interaction scores and amino acid conservation, we only calculated the score for sites, where the interaction is achieved by atoms of the amino acid side chain. These scores are generally smaller than those calculated for side chain and backbone interactions, because side chain interactions are only part of the total interactions and many sites are very flexible in the contribution of atoms to the interaction interface. Corresponding positions in homologous sequences can interact with side chain atoms in one sequence and exclusively with backbone atoms in the next sequence.

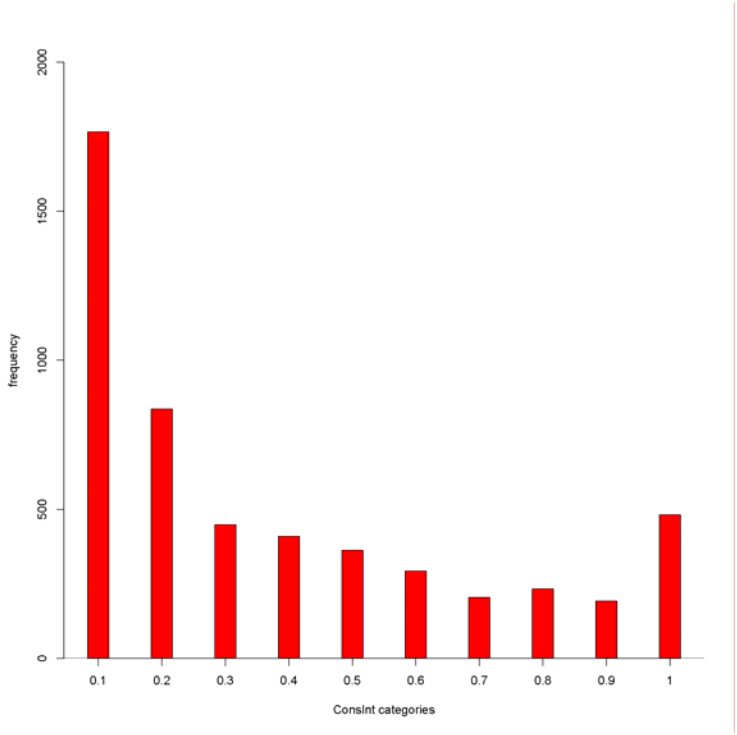


Figure 11: Distribution of interaction scores.

The interaction score reflects the importance of a functional sites in establishing an interaction. Surprisingly, only few interacting sites are absolutely conserved in their location within the whole protein family and characterized by high interaction scores. The majority of interacting sites feature small interaction scores. This shows that these sites are only used by a few sequences of the domain family for establishing an interaction, which can also be caused by the different nature of ligands.

The distribution of interaction scores is shown in figure 11. It is remarkable that there are only a few positions in all of the investigated domains with the maximum interaction score. Smaller interaction scores can emerge from the use of ligands with distinct functions in the PDB data, for example, if a domain is in complex with its native substrate or another time with a regulatory protein that binds to a remote part of the domain. Absence of the substrate in the latter case leads to lower interaction scores for important functional sites: even medium interaction scores can indicate significant interactions. A large proportion of amino acid sites have very low interaction scores, presenting sites that are only occasionally involved in interactions or that are specific to a subgroup. Subgroups tend to use the same functional positions, while the functional sites can vary between subgroups. This behaviour is also reflected in the distance trees of the domains. The carbohydrate binding RICIN domain in figure 12 exemplifies the divergence of functional sites within a domain family. By inspecting the arrangement of functional sites in the different RICIN sequences, the RICIN family can be divided into three main subgroups of distinct functions. While one subgroup possesses two carbohydrate binding sites and a peptide binding region (group II in figure 12), one subgroup is limited to the N-terminal carbohydrate binding region (group I) and one to the C-terminal carbohydrate binding region (group III). Absence of a carbohydrate binding site in the first or third group is unlikely to be an artefact of the crystallization process, since the domains were crystallized in complex with an adjacent sugar bound domain.

The classification observed in the interaction profile of figure 12 is also reflected in the cladogram given to the left of the interaction profile. Interestingly, the subgroups that preserved only one of the carbohydrate binding sites originate from proteins with tandem RICIN domains, which arose from gene duplications (VILLAFRANCA and ROBERTUS 1981), so that the proteins are again provided with two carbohydrate binding sites. A single RICIN domain can be divided into three subunits of approximately 40 amino acids in lengths that have evolved from an ancient galactose binding peptide (RUTENBER *et al.* 1987). These subunits represent the differently specialized binding sites in the domain family. In proteins carrying two RICIN domains, only the first subunit of the first RICIN domain and the last subunit of the second RICIN domain preserved their ability to bind carbohydrates, while one subunit has specialized on binding peptides (subgroup II in figure 12).

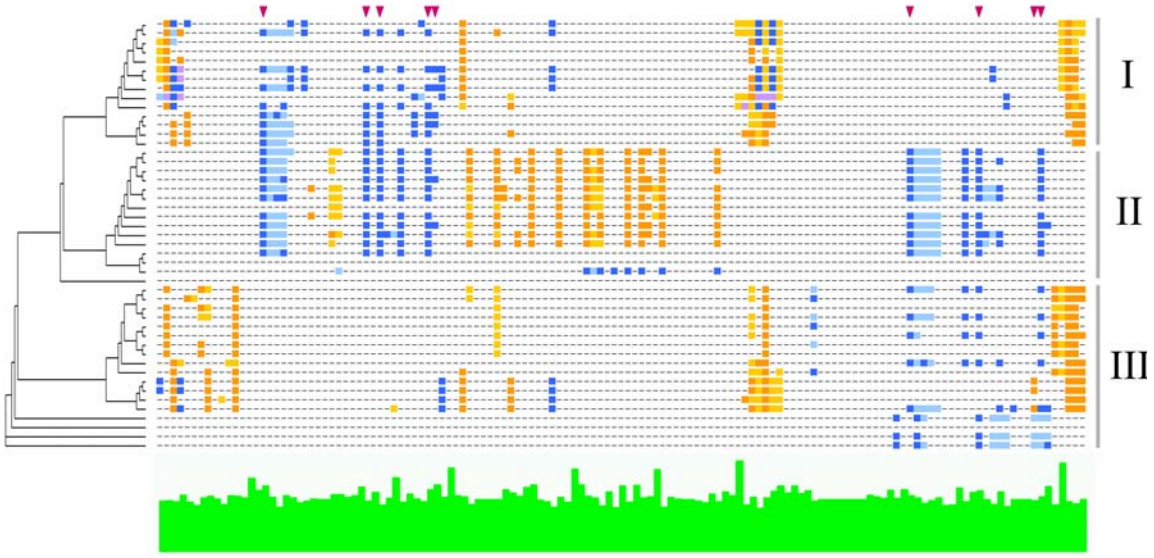


Figure 12: Interaction profile of the RICIN domain.

Alignment of positions corresponding to an HMM match state. Sites interacting with saccharides are indicated in blue, peptide interactions in orange, and sites interacting with both ligands, saccharides and peptides, are indicated in purple. Light colours represent backbone interactions, darker colours involve side chain atoms. The amino acid conservation is visualized by green bars below the alignment. Sugar binding sites described in the literature are indicated by red arrows above the alignment (PASCAL *et al.* 2001). Several positions (1, 3, 4, 22, 42, 58, 88, 90, 122) are located in the vicinity of a glycosylation site, but do not specifically interact with saccharides. The unrooted tree reflects the classification into three main subgroups with different interaction sites. Group II harbours two sugar-binding sites, group I and III originate from tandem RICIN domains, in which group I preserved the N-terminal sugar-binding site and group III the carboxy-terminal binding site. PDB identifiers from top to bottom: 1PC8 (B: 5-131), 1TFM (B: 5-131), 2MLL (B: 5-131), 1CE7 (B: 5-131), 1ONK (B: 9-135), , 1PUM (B: 9-135), 1M2T (B: 257-383), 1OQL (B:13-139), 1ABR (B: 13-139), 2AAI (B: 8-134), 1HWO (B: 10-135), 1HWP (B: 10-135), 1HWN (B:10-135), 1HWM (B:3-266), 1V6U (A: 312-436), 1ISW (A:312-436), 1ISV (A:312-436), 1ITO (A:312-436), 1V6W (A: 312-436), 1V6X (A: 312-436), 1XYF (A:312-436), 1ISY (A: 312-436), 1ISZ (A:312-436), 1V6V (A:312-436), 1ISX (A:312-436), 1KNM (A:7-131), 1KNL (A:9-133), 1MC9(A:9-133), 1QXM (A: 29-157), 1PUM (B: 140-262), 1M2T (B: 390-510), 1ONK (B: 140-262), 1OQL (B: 140-262), 1PC8 (B: 136-254), 1TFM (B: 136-254), 2MLL (B: 136-254), 1CE7 (B:136-254), 2AAI (B: 138-261), 1ABR (B: 143-266), 1HWO (B: 138-262), 1HWP (B: 138-262), 1HWM (B: 138-262), 1HWN (B: 139-263), 1FWU (A: 3-123), 1DQG (A: 4-124), 1DQO (A: 4-124), 1FWV (A: 3-123).

The RICIN domain is a good example how homologous sequences belonging to different subfamilies of a domain specialized on binding different ligands and thus, on various functions. The functional sites are no longer conserved throughout the whole family and they could not be inferred from homologous sequences. However, functional sites can be predicted, if the information is taken from the most closely related sequences.

Alternating locations of functional sites are especially prominent in DNA binding domains, like the C2H2-zinc finder and homeodomains (LUSCOMBE and THORNTON 2002). Another example is the high mobility group (HMG) domain, which is shown in complex

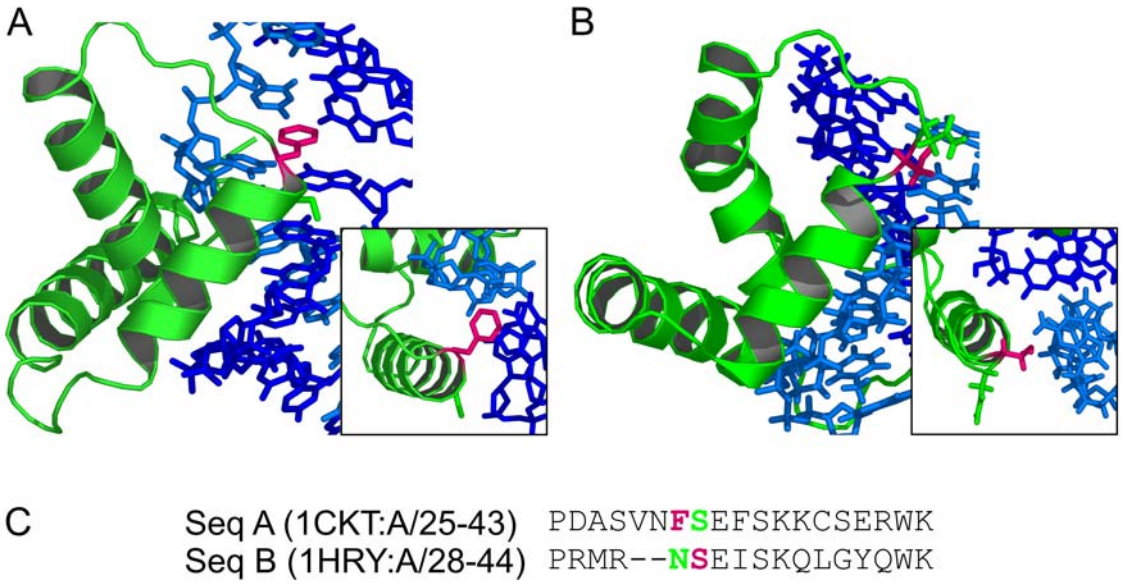


Figure 13: Variable location of interacting amino acid residues in the HMG domain

Sequence specific interaction by the high mobility group (SMART: HMG) domain (green) is achieved by an amino acid side chain (pink) pointing into the DNA double helix (blue). The interaction is achieved by a phenylalanine in figure 13a (OHNDORF *et al.* 1999) or by a serine residue in figure 13b (WERNER *et al.* 1995). The sequence alignment (figure 13c) reveals that these two interacting residues are not located at corresponding position.

with its target DNA in figure 13. Here, an amino acid side chain pointing into the DNA helix recognizes the DNA bases of the target sequence. This key position is located in the loop connecting the first and second alpha helix of the domain. The contact is carried out by a serine residue in figure 13a (pdb id: 1hry; WERNER *et al.* 1995). In contrast, a phenylalanine establishes the contact in figure 13b (pdb id: 1ckt; OHNDORF *et al.* 1999) and the serine residue corresponding to the structure shown in 3a is pointing away from the DNA helix. In the sequence alignment, these two key positions are located adjacent to each other. Many other domains show this variability in the location of interacting amino acid residues and profit from the flexibility to fine-tune substrate specific binding sites based on the same structurally conserved protein fold.

6.3.2 Amino acid conservation of interacting sites

Having observed great differences in the location of interacting sites within conserved domains, the question arises how these sites behave with regard to their amino acid conservation. It is generally believed that interacting sites are more highly conserved than non-interacting solvent-assessable sites. However, about one third of all interactions

in our analysis were achieved by backbone atoms only, so that the kind of amino acid has no direct effect on the interaction. For the remaining two thirds, which account for side chain interactions, we calculated a score for the conservation of amino acid similar to the interaction score ConsInt. We next compared the relative frequency distribution of the amino acid conservation score of interacting sites with the one of non-interacting sites, which are composed of amino acids located in the core of the domain as well as non-interacting surface residues. The distribution of interacting sites is slightly shifted to higher amino acid conservation scores (figure 14). This shift is clearly visible and is statistically significant (Kolmogorov-Smirnoff test: p-value<2,2x10-16), although the data have not been restricted to surface residues. Our finding that interacting sites are more highly conserved, on average, is in accordance with the results of other groups who reported that domain-domain interfaces are better conserved than the rest of the surface residues (LITTLER and HUBBARD 2005).

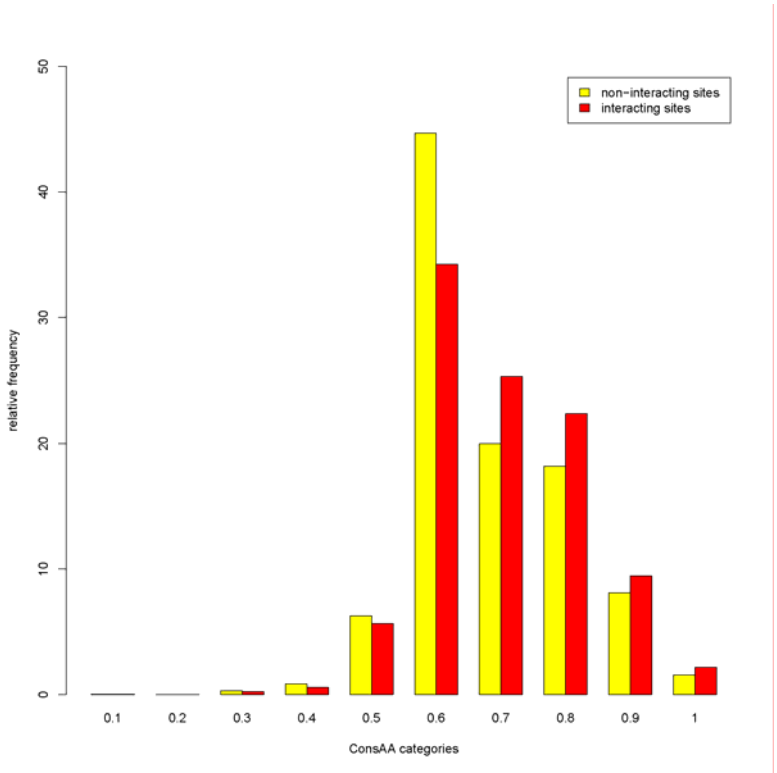


Figure 14: Amino acid conservation of interacting and non-interacting sites.

Non-interacting sites (yellow) are slightly more highly conserved than interacting sites (red) as shown by the shift to higher amino acid conservation of interacting sites.

Strikingly, there are many interacting sites with very low amino acid conservation scores. A possible explanation could be that very few residues of the interaction interface are important for a stable interaction and conserved in their amino acid. It has been shown that only a subset of interface residues contribute a crucial part to the binding energy, while residues around these so-called hot spots are less conserved (CLACKSON and WELLS 1995; KESKIN *et al.* 2005). Another explanation could be the increased specificity for various ligands. The data set contains orthologus sequences, which might be conserved in function and substrate specificity and paralogous sequences, which might have accumulated mutations throughout evolution and adopted new substrate specificity. The zinc finger domain, for example, interacts mainly through an aromatic residue with the nucleic acid, but specificity is provided by a nearby position, which can vary in its amino acid and is also interacting with the bases of the nucleic acid (figure 15). Hence, the variety of amino acids at functional sites could be advantageous to recognize numerous different ligands by the same domain family.

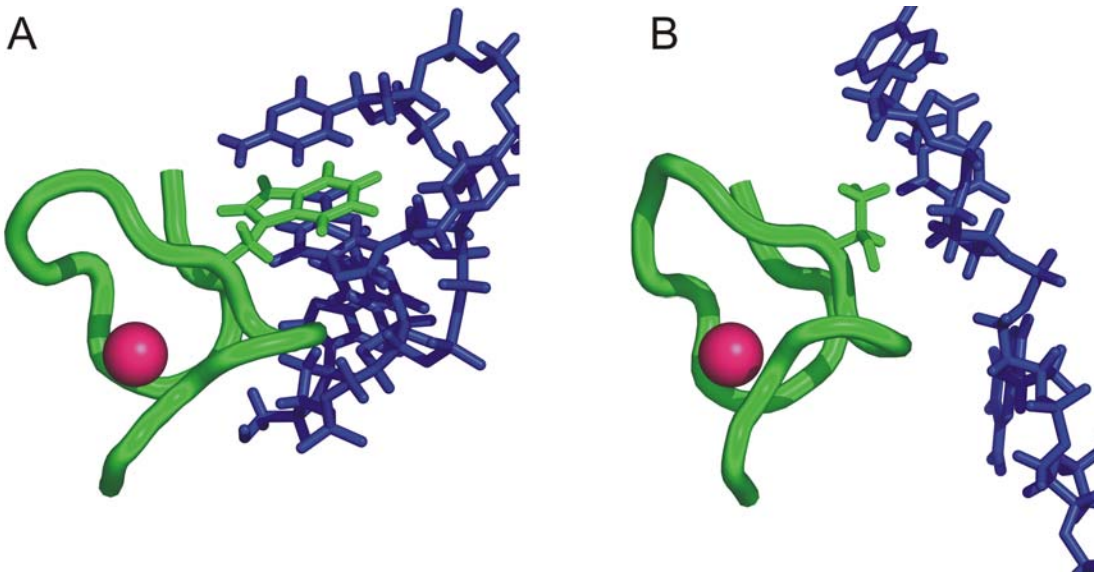


Figure 15: Substrate specific interaction by varying the type of amino acid.

Substrate specificity in the zinc finger domain (SMART: ZnF_C2H2) is ensured by various amino acids that interact with the bases of the DNA. The protein domain is highlighted in green, the DNA chain in orange and the zinc atom in red.

6.3.3 Correlation of interaction and amino acid conservation

In order to test whether often-used interacting sites coincide with highly conserved sites we plotted the interaction score against the amino acid conservation score (figure 16). Since our dataset includes various ligands, which might have different preferences in terms of the amino acid conservation or interaction, we divided the data by the type of ligand. Groups analyzed correspond to peptides, nucleotides, ions and all ligands, including those, which could not be classified into one of the prior groups. Statistical analysis detected a significant positive correlation between the interaction score and its amino acid conservation in all observed groups (see figure 16 for details). The trend to higher interaction scores with increasing amino acid conservation is clearly visible if the data are divided into groups and the median calculated for each group. The median values of ion binding sites are shifted to higher amino acid conservation scores compared to the other three ligand groups, indicating that ion binding sites are more highly conserved than sites binding peptides, nucleotides or other small molecules. The highest median interaction scores are found in the group of nucleotide ligands, consequently nucleotide binding sites preferentially use the equivalent positions in homologous sequences. Interestingly, sites interacting with peptides are not less conserved in their amino acids but are more flexible in the location of interaction compared to nucleotides. An unexpected finding is the great variance of amino acid conservation scores for high interaction scores, especially in the nucleotide and peptide group. This indicates that these interacting sites are very flexible in the type of amino acid and specialized to complement the ligand and to increase specificity of the protein-ligand interaction.

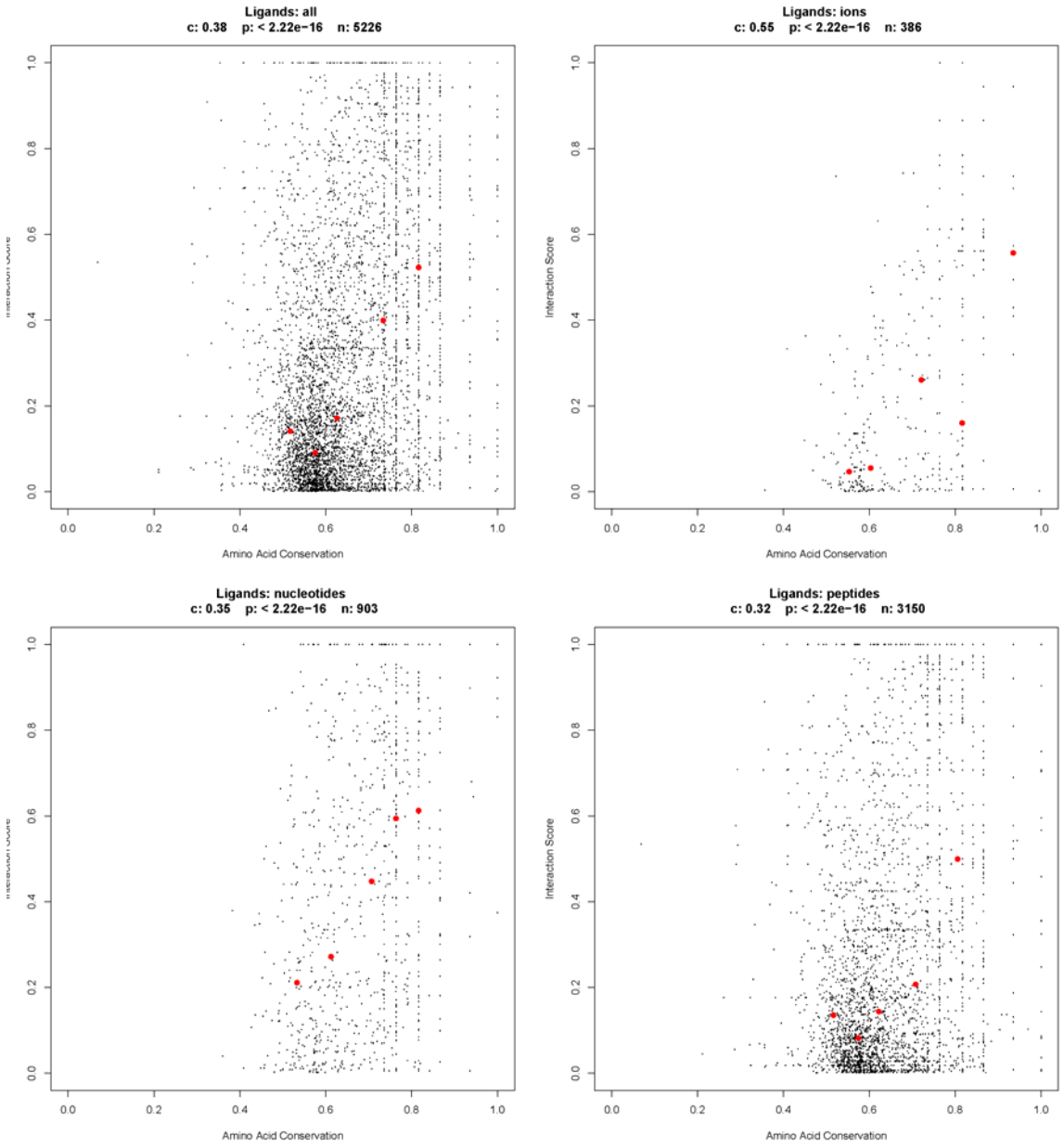


Figure 16: Correlation of interaction scores and amino acid conservation

For better visualization of the correlation, the data was divided into five groups corresponding to the 0-20% quantile, 20-40% quantile, etc. of the amino acid conservation scores and then the median interaction score and median amino acid conservation score was calculated for each group and plotted with red dots. The correlation coefficient, p-value and population are indicated for each ligand above the graphs. The correlation coefficient was calculated according to Pearson's method under the null-hypothesis of no correlation (c=0).

6.4 Conclusions

Our analysis reveals that functional sites can be highly variable in their amino acid conservation and very flexible in using various locations in the protein domain. The properties of functional sites are dependent on the protein family and can vary from highly conserved, as observed in enzymes involved in DNA replication, to protein families that are highly variable with various amino acids at various locations, as for example immunoglobulins or carbohydrate-binding domains. Similar results were obtained by other groups. Pachenko and co-workers analyzed 86 domains from the CDD database and report that functional sites of homologous sequences can greatly differ in their physicochemical properties and their location in the three-dimensional structure (PANCHENKO *et al.* 2004). Variability in functional sites was also described by Devos and Valencia. By comparing the conservation of binding sites in structural alignments, they found high conservation in diverged sequences contrasting highly similar sequences with different interacting sites (DEVOS and VALENCIA 2000). Our findings present valuable information for the improvement of methods to predict functional sites. In most of these methods, the prediction is based on a set of homologous sequences. This approach results only in reliable predictions if the investigated protein family is conserved in most of the functional sites. Approaches to improve the accuracy of prediction have been made recently by using orthologous proteins with presumably the same ligand specificity (MIRNY and GELFAND 2002) or by sub-typing protein families (HANNENHALLI and RUSSELL 2000). With the growing number of sequences within protein families, trends that consider variability of functional sites and use subgrouping aided by experimental information might become widely accepted in the future and promise to be successful in the prediction of functional sites.

Table 10: Interacting sites in protein domains

For each domain, the number of domains with ligand interactions extracted from PDB (# PDB), the number of interacting sites corresponding to HMM match states (int sites), the number of interacting sites corresponding to HMM insert states (loop sites), the percentage of insert state interacting sites from the total number of interacting sites (% loop) and the number of interacting sites in the domain consensus (domain int sites) are given. The last column considers conserved positions in the domain consensus sequence, while all other columns count sites in sequences belonging to the domain family. The number of conserved interacting sites (domain int sites) can be 0 despite plenty of interacting sites in family sequences if the site-specific interaction score does not yield a positive value due to identical sequences. Abbreviations of domain names are according to SMART (LETUNIC *et al.* 2004).

Domain	# in PDB	total sites	loop sites	% loop	domain int sites
14_3_3	6	117	0	0,0	20
35EXOc	34	434	0	0,0	20
AAA	93	961	523	35,2	57
AAI	13	162	73	31,1	19
Aamy	102	2359	314	11,7	166
Aamy_C	31	192	0	0,0	4
ACTIN	31	1380	1	0,1	66
ALBUMIN	19	234	0	0,0	40
alkPPc	33	2284	0	0,0	71
Amb_all	4	18	0	0,0	3
Ami_2	4	39	2	4,9	0
ANK	54	230	0	0,0	14
ANX	32	144	0	0,0	16
ARF	15	552	0	0,0	38
ARM	151	682	13	1,9	0
B_lectin	8	190	0	0,0	31
B41	6	54	0	0,0	2
BHL	11	236	0	0,0	23
BIR	17	195	0	0,0	18
BowB	5	69	0	0,0	10
BRCT	13	68	12	15,0	6
BRLZ	24	282	0	0,0	36
BROMO	4	67	1	1,5	13
BTB	4	16	6	27,3	0
C2	5	21	23	52,3	5
CA	7	97	4	4,0	15
CASc	24	941	3	0,3	94
CBD_IV	5	107	1	0,9	5
CCP	20	208	9	4,1	20
CH	6	38	7	15,6	1
CHROMO	6	88	0	0,0	16
ChtBD1	17	69	0	0,0	6
ChtBD3	16	170	0	0,0	6
CLECT	64	591	163	21,6	38
cNMP	24	322	0	0,0	16
CSP	5	32	1	3,0	0
CUB	4	28	4	12,5	3
CYCc	17	351	5	1,4	14
CYCLIN	69	993	4	0,4	36
DEXDc	12	82	15	15,5	4
DSRM	4	64	0	0,0	14
DWB	4	85	0	0,0	18
EFh	235	2076	0	0,0	27
EGF	38	257	145	36,1	18

Table 10 (continued)

Domain	# in PDB	total sites	loop sites	% loop	domain int sites
EGF_CA	26	218	17	7,2	9
ENDO3c	33	680	0	0,0	40
ETS	13	244	0	0,0	24
FA58C	5	44	2	4,3	6
FBG	18	249	10	3,9	15
FBOX	4	64	0	0,0	14
FES	12	93	0	0,0	9
FGF	31	458	2	0,4	43
FHA	12	96	8	7,7	7
FN3	20	85	92	52,0	10
FU	12	58	3	4,9	10
FYVE	4	35	0	0,0	13
G_alpha	17	617	0	0,0	52
GAL4	12	168	4	2,3	23
GEL	27	413	5	1,2	21
GGL	9	333	0	0,0	31
GHA	4	139	0	0,0	0
GHB	4	147	0	0,0	3
GLA	20	256	1	0,4	28
GLECT	20	196	0	0,0	13
Glyco_10	30	557	6	1,1	35
Glyco_18	34	916	59	6,1	96
GuKc	10	163	0	0,0	25
H2A	22	1523	1	0,1	65
H2B	22	1523	0	0,0	64
H3	23	1491	0	0,0	73
H4	23	972	64	6,2	43
HATPase_c	30	556	32	5,4	45
HDc	16	101	211	67,6	12
HhH1	115	885	0	0,0	5
HintN	4	28	9	24,3	4
HLH	10	107	0	0,0	11
HMG	11	263	1	0,4	33
HNHc	11	139	0	0,0	17
HOLI	85	1173	68	5,5	51
HOX	32	519	6	1,1	37
HTH_CRP	13	155	0	0,0	11
HTH_DTXR	10	98	0	0,0	12
HTH_LACI	30	402	0	0,0	13
HTH_XRE	8	108	3	2,7	16
HX	17	78	2	2,5	5
IG	33	148	183	55,3	20
IGc1	89	634	3	0,5	51
Igc2	31	256	60	19,0	31
Igv	485	2689	1307	32,7	51
IL2	7	77	0	0,0	6
IL6	4	47	0	0,0	9
IlGF	162	2450	0	0,0	37
Int_alpha	10	37	18	32,7	6
IPT	27	157	0	0,0	1
IQ	17	306	0	0,0	19
JmjC	6	140	0	0,0	1
KAZAL	34	428	0	0,0	23
KH	10	97	12	11,0	13

Table 10 (continued)

Domain	# in PDB	total sites	loop sites	% loop	domain int sites
KISc	20	518	12	2,3	62
Knot1	4	12	1	7,7	0
KOW	18	145	0	0,0	0
KR	12	150	0	0,0	15
KU	50	814	43	5,0	24
Ku78	4	343	0	0,0	51
L27	4	112	1	0,9	32
LamG	11	144	10	6,5	2
LDLa	7	58	1	1,7	5
LMWPc	13	140	0	0,0	12
LRR_TYP	15	25	0	0,0	4
LRRCT	5	16	15	48,4	3
LRRNT	5	20	0	0,0	3
LY	11	36	3	7,7	5
LYZ1	105	1567	0	0,0	66
MADS	7	133	0	0,0	21
MATH	14	217	17	7,3	19
MBT	9	79	0	0,0	8
ML	4	16	3	15,8	1
MYSc	24	417	6	1,4	21
NDK	10	99	0	0,0	11
NGF	7	129	0	0,0	22
NH	4	59	0	0,0	7
NTR	5	105	0	0,0	21
PA2c	22	322	0	0,0	25
PAC	7	59	0	0,0	6
PAS	8	184	6	3,2	45
PAX	4	163	0	0,0	28
PBD	5	97	0	0,0	24
PbH1	21	42	13	23,6	6
PBPb	54	629	81	11,4	27
PBPe	48	638	0	0,0	24
PDGF	8	98	0	0,0	11
PDZ	24	270	22	7,5	17
Pept_C1	25	770	1	0,1	103
PGRP	4	39	0	0,0	1
PH	12	91	67	42,4	9
PhBP	8	67	0	0,0	6
POL3Bc	6	170	0	0,0	7
POLAc	41	730	0	0,0	23
POLBc	8	171	6	3,4	24
POLXc	99	2019	0	0,0	37
PP2Ac	9	183	0	0,0	29
PQQ	26	119	28	19,0	13
PreSET	4	37	1	2,6	8
PROF	8	87	0	0,0	6
PRP	4	56	0	0,0	16
PTB	8	173	2	1,1	22
PTI	10	102	0	0,0	12
PTPc	52	811	9	1,1	30
PUA	6	40	0	0,0	4
Pumilio	24	100	0	0,0	5
RAB	22	571	0	0,0	32
rADc	4	58	0	0,0	0

Table 10 (continued)

Domain	# in PDB	total sites	loop sites	% loop	domain int sites
RAN	11	400	0	0,0	10
RanBD	4	143	0	0,0	30
RAS	24	717	0	0,0	52
RasGEF	4	200	1	0,5	0
RasGEFN	4	90	0	0,0	0
REC	33	336	14	4,0	37
RGS	5	113	3	2,6	13
RHO	29	973	0	0,0	86
RHOD	6	48	4	7,7	2
RhoGAP	5	105	15	12,5	19
RhoGEF	6	150	0	0,0	23
RICIN	47	980	28	2,8	53
RING	4	39	25	39,1	13
RL11	10	197	0	0,0	20
RNAse_Pc	71	1046	0	0,0	57
RPOL8c	5	148	0	0,0	0
RPOL9	7	117	0	0,0	0
RPOLA_N	10	536	3	0,6	58
RPOLD	12	572	0	0,0	37
RRM	49	843	25	2,9	47
S_TK_X	28	78	0	0,0	7
S_TKc	159	3698	682	15,6	117
S1	5	44	3	6,4	1
S4	18	253	3	1,2	7
SANT	13	180	1	0,6	19
SCY	12	114	0	0,0	19
Sec7	5	148	0	0,0	27
SEL1	4	11	0	0,0	0
SERPIN	43	1391	46	3,2	133
SET	8	59	51	46,4	6
SH2	85	1126	24	2,1	42
SH3	47	577	18	3,0	31
Skp1	4	134	0	0,0	13
Sm	11	167	3	1,8	33
SPEC	5	46	0	0,0	2
START	4	82	0	0,0	23
STI	6	131	3	2,2	9
t_SNARE	17	571	0	0,0	56
TGc	8	66	1	1,5	2
TGFB	9	124	0	0,0	20
THN	6	18	0	0,0	2
TNF	10	175	3	1,7	18
TNFR	15	84	0	0,0	17
TOP1Ac	11	112	0	0,0	10
TOP1Bc	9	61	0	0,0	7
TOP2c	5	245	0	0,0	50
TOPEUc	9	345	0	0,0	43
TOPRIM	13	67	1	1,5	7
TPR	15	69	0	0,0	14
TR_FER	33	329	1	0,3	13
TR_THY	27	191	0	0,0	15
TRASH	14	242	0	0,0	0
Tryp_SPc	107	3534	567	13,8	138
TSPc	5	197	4	2,0	0
TyrKc	31	642	11	1,7	46

Table 10 (continued)

Domain	# in PDB	total sites	loop sites	% loop	domain int sites
UBCc	9	139	12	7,9	24
UBQ	17	275	27	8,9	38
UIM	4	55	0	0,0	18
VHS	6	92	0	0,0	12
VWA	31	322	72	18,3	53
WD40	81	552	86	13,5	22
WH1	4	44	0	0,0	9
WW	10	97	0	0,0	12
Zn_pept	7	109	1	0,9	12
ZnF_C2C2	5	69	0	0,0	0
ZnF_C2H2	69	549	70	11,3	16
ZnF_C2HC	7	83	0	0,0	12
ZnF_C4	23	448	0	0,0	28
ZnF_GATA	7	109	0	0,0	13
ZnF_TAZ	4	164	0	0,0	44
ZnMc	65	1352	23	1,7	53

6.5 Methods

6.5.1 Data set

The analysis is based on the October version of PDB (27969 structures). All protein sequences extracted from PDB files were scanned against all SMART domains (667) using hmmsearch from the HMMER package[14] (version 2.3.2). Profile HMMs were retrieved from the SMART family alignments and score thresholds were used as assigned by SMART for each individual domain (SCHULTZ *et al.* 1998). The search resulted in 8747 protein structures containing at least one of 480 SMART domains. For those structures containing a protein-ligand interaction, we calculated the distance for each atom of all protein compounds to each atom of all ligands. We considered amino acids as interacting if the distance of any atom of the amino acid to any atom of the ligand was smaller than 4 Angstrom, which is a very conservative threshold and is in the range of the two oxygen atoms in a hydrogen bond. Interactions to water molecules were neglected, as well as interactions between identical chains, because homodimers can present artefacts arising from the crystallization process. We are aware of losing information about naturally occurring homodimers. If more than one model of the protein structure exists, for example

[14] http://hmmer.wustl.edu (HMMER: sequence analysis using profile hidden Markov models)

in the case of NMR data, all models were taken into account and a position was treated as interacting if more than 50% of all models harbour an interaction at this position. For each domain family, a multiple sequence alignment was generated from the sequences identified in the PDB scan for SMART domains. The alignments were created with hmmalign (HMMER package) according to the SMART profile HMMs. Distance trees were created with PROTDIST and FITCH, both from the Phylip package (FELSENSTEIN 1989).

A problem with using the protein data bank is the overrepresentation of some proteins, while others are completely absent. To deal with the biased nature of the database, we used sequence weights, correlated to the evolutionary distance between the sequences. The distance is small for similar sequences, so that these proteins are weighted with a negligible small factor. Although many proteins are redundant in our dataset it is advantageous to consider all proteins because they can be bound to different ligands or the complex crystallized under different conditions. The most important interacting sites should clearly stand out in the large scale analysis.

Ligands interacting with protein domains were, wherever possible, classified into groups of peptides, nucleotides or ions. The group of ions was restricted to small ions and excluded typical buffer anions. The groups of peptides and nucleotides also included modified molecules that are functionally alike. We also analysed all ligands together, including carbohydrates, buffer ions and other small molecules.

6.5.2 Calculation of scores

For each alignment position consistent with the HMM consensus sequence, we calculated a position-specific interaction conservation score (ConsInt) to describe how well a position is conserved in ligand interactions within the domain family.

$$ConsInt(i) = \sqrt{\frac{\sum_{a}^{N}\sum_{b}^{N} Dist(seq_a, seq_b) \times Int_i(seq_a(i), seq_b(i))}{\sum_{a}^{N}\sum_{b}^{N} Dist(seq_a, seq_b)}}$$

$$Int_i = \begin{cases} 1, \text{if seq a and b interact at pos i} \\ 0, \text{otherwise} \end{cases}$$

where N is the number of sequences in the alignment, Dist(seqa,seqb) is the phylogenetic distance between sequence a and b obtained from the phylogenetic tree and Inti takes a value of 1 if sequence a and b both interact at position i, and 0 otherwise. The score ranges from 0 to 1 and is 1 if all sequences interact at the position of interest. To obtain a score greater than 0 for a certain position at least two non-identical sequences have to interact at this position. In this way, interactions to non-physiological ligands like artificially synthesized peptides should be lost, and important interacting positions should have noticeably higher scores. The score takes into account the phylogenetic distances between the sequences so that highly similar sequences are weighted more weakly, in contrast to interacting sites in divergent sequences, which are weighted more strongly.

Similar to the interaction score, a position specific amino acid conservation score (ConsAA) was calculated.

$$ConsAA(i) = \sqrt{\frac{\sum_{a}^{N}\sum_{b}^{N} Dist(seq_a, seq_b) \times Subst(seq_a(i), seq_b(i))}{\sum_{a}^{N}\sum_{b}^{N} Dist(seq_a, seq_b)}}$$

Here, Substi(a,b) measures the similarity between the amino acids at position i in sequences a and b. It is based on the VTML 160 substitution matrix (MULLER *et al.* 2002). The conservation score was normalized between 0 (low aa conservation) and 1 (high aa conservation) ranging from the lowest to the highest score of the amino acid substitution matrix. It is important to note that our amino acid conservation scores do not describe protein families as found in the SMART database, but only the data set used in our analysis, so that the amino acid conservation score can be compared with the interaction conservation score for each position in the family alignment.

The software suite R was used for the statistical analyses (R DEVELOPMENT CORE TEAM 2004). The concentration of data points on vertical lines found for higher amino acid conservation scores in the scatter plots of figure 16 are an effect of the discrete values of the amino acid substitution matrix and the few substitutions at conserved sites.

7 Concluding Discussion

The ultimate goal of the projects presented in this thesis was to investigate the properties of protein domains as well as their evolution in order to understand their function and to improve the prediction of function of so far uncharacterised protein sequences. Two different approaches were undertaken to pursue this aim. On one hand, in-depth analyses were carried out, which investigated individual protein families. For these families we gained insights into their evolution and were able to propose functional important sites (chapter 2, 3 and 4). On the other hand, large-scale analyses were used to get a broader picture of the general properties of protein domains (chapter 5 and 6). However, the computer-aided performance of large-scale analyses often includes to sacrify the accuracy of the analysis. While in-depth analysis can identify family specific details, the large-scale analysis can compare groups of protein families in regard of special features and allow a rapid coverage of all protein families described in the databases, providing a more general description of the protein families.

In this chapter, I will discuss the results of the individual analyses presented in this thesis in a larger framework and will highlight new developments in the field achieved by my colleagues and other groups. In addition, I will show how the findings of my own work contribute to the development of the area of protein function annotation.

7.1 Zooming on the O-linked N-acetyl-β-D-glucosaminidase family

Focusing on protein domains important in signal transduction pathways, the first in-depth study involved the N-acetyl-β-D-glucosaminidase (GlcNAcase) family, which was little understood in its catalytic mechanism. Moreover, the enzymatic function had not been mapped to any particular part of the protein sequence, containing two separate domains. To gain insight into the evolution of this enzyme family, homologous sequences were searched with PSI-BLAST. This iterative search method is advantageous for diverged families, but also bears the risk to include false positive hits in the generation of the

position specific scoring matrix. To avoid gliding into a different protein family, the alignment of each new hit identified by PSI-BLAST was visually inspected. The result of the sequence search was the most important part of the analysis, on which all other conclusions are based. Thus, the evolutionary relationship detected between the GlcNAcases and the GCN5-related acetyltransferases had to be verified by other methods. We obtained threefold evidence for a common ancestry: First, intermediate sequence searches link the human GlcNAcase to the acetyltransferases via a putative acetyltransferase from *Streptomyces coelicolor*. Second, the secondary structure prediction for the GlcNAcase sequences is corresponding to the experimentally determined three-dimensional structures of acetyltransferases. Third, the GlcNAcase sequences share a motif with an acetyltransferase family member, which is involved in the biosynthesis of UDP-GlcNAc and thus able to bind N-acetylglucosamine-6-phosphate (GlcNAc6P).

The alignment of GlcNAcase family members, which was built from the PSI-BLAST search hits, also allowed to scan the family for positions conserved in their type of amino acids and to predict essential functional sites. In general, only very few positions in a family alignment are absolutely conserved. These positions usually correspond to catalytic sites or important structural sites. Thus, the prediction of catalytic sites is relatively straightforward by searching for sites that are well conserved throughout all family members. However, it cannot be excluded that well conserved residues are playing a structural role and are fundamental for the correct folding of the protein. Similarly to catalytic sites, substrate or cofactor binding sites are well conserved. Very often these regions are described as protein motifs. For example, a striking difference between the GlcNAcases and the remote homologous acetyltransferases is the absence of a Q/RxxGxG motif (WOLF *et al.* 1998) needed for binding of the cofactor acetyl-CoA. Since the conservation of an amino acid alone does not present reliable evidence for the presence of catalytic sites or substrate/cofactor binding positions, it is beneficial to use additional information, such as experimentally characterized homologous sequences. By using the well-studied remote homologous acetyltransferases sequences, information was transferred onto the uncharacterised GlcNAcase family in regions of high amino acid conservation between the diverged family members. The acetyltransferases exhibited the location of the catalytic center, the substrate binding and cofactor binding residues. Interestingly, the motif for substrate binding was conserved between GlcNAcase sequences and a family member of acetyltransferases (GNA1 acetyltransferase), which is specific for N-

acetylglucosamine. Assuming that the location of the catalytic center was maintained, we were able to propose highly conserved residues presumably playing an essential role in the hydrolysation of GlcNAc modified proteins.

The classification of GlcNAcase sequences as family members of GCN5-related acetyltransferases also implied the disclosure of a new trend in the evolution of acetyltransferases. In general, evolutionary processes can change substrate affinity as well as the underlying catalytic mechanism to generate new enzymes. Examples for both mechanisms in the evolution of new enzymes are abundant. Biochemical pathways like histidine biosynthesis (FANI *et al.* 1994), tryptophane synthesis (WILMANNS *et al.* 1991), or methionine biosynthesis (BELFAIZA *et al.* 1986) contain many homologous enzymes, which indicates that one enzyme with a particular substrate affinity gave rise to another enzyme of the same pathway by changing the catalytic reaction performed on the substrate. Horowitz proposed that pathways evolved backwards, where a functional enzyme is being used as starting material to develop a previous reaction in the pathway leading to a compound that can bind to both enzymes (HOROWITZ and NETZENBERG 1965). However, homologous enzymes of a pathway are not always found in adjacent reactions. It is also possible that enzymes with different catalytic mechanisms were recruited from other pathways and then enzymes catalyzing similar reactions in different pathways would be homologous (JENSEN 1976). In the case of the GCN5-related acetyltransferase family, both mechanisms of enzyme evolution have occurred and led to the generation of new enzymes. The O-GlcNAcases are so far the first family members that do not transfer an acetyl group from the cofactor acetyl-CoA to various substrates. Instead, they have developed a new catalytic mechanism, starting from their ability to bind glucosamine. All other acetyltransferase family members described in the literature have changed their substrate affinity and maintained the catalytic mechanism as well as cofactor binding ability while diverging from the common ancestor.

7.2 Chasing functional sites

Similar to the GlcNAcases, another domain involved in signal transduction pathways was studied with the aim to identify functional important sites. The CHASE domain, an extracellular domain of a sensory receptor as part of the two component

signalling pathway in bacteria and lower eukaryotes was little characterized and positions involved in the binding of cytokinin in the plant CHASE domain were of great interest. The prediction of functional important sites for plant CHASE domains was possible due to the functional divergence within the CHASE family, which can bind to physicochemical different ligands, such as cytokinin-like adenine derivatives in plants and peptide ligands as assumed for bacteria. By comparing evolutionary site rates between the functional different subgroups of the CHASE family, putative functional residues were predicted (chapter 3.3.1). Four of the five predicted functional residues were experimentally verified by substitutions of these residues to alanine, which led to loss of the ligand-binding ability.

When the study of the CHASE domain was performed, its structure was still unknown. Recently, Pas and colleagues suggested a three-dimensional model for the CHASE domain and also predicted positions, which are presumably interacting with the cytokinin ligand (PAS *et al.* 2004). Their approach is based on the similarity to distant related structures of sensory domains, a fumarate sensor from *E. coli* and a citrate sensor from *Klebsiella pneumoniae*, which both activate the bacterial two-component signalling system. Both structures adopt a PAS-domain-like fold, which might be phylogenetically related to the CHASE domain. Pas and colleagues predicted several ligand-binding positions by molecular modelling and by docking various ligands into a surface cleft. However, none of the suggested positions is in accordance with our own predictions or with the results of our analysis. More than half of the positions that are presumably ligand-binding and conserved among plant sequences according to Pas and colleagues are not conserved in the type of amino acid according to the multiple sequence alignment of our own analysis. This might be caused by the fact that the Pas analysis used a very small number of CHASE sequences from plants and that their analysis is biased towards the *Arabidopsis thaliana* receptors. Unfortunately, the exact number of sequences is not revealed in the Pas report. To gain additional information on the diversity of the CHASE domain among plants, we scanned Genbank's EST database and obtained 14 non-identical CHASE domain sequences from plants, which comprise sequences from rice, maize, millet, wheat, tomato and rose alongside *Arabidopsis thaliana* and provide an adequate basis for studying patterns of amino acid conservation.

The discrepancy between our own and the Pas analysis affects not only the pattern of conserved amino acids in the plant CHASE sequences, but also the set of ligand-binding

residues. This shows how strong prediction methods are depending on the quality of the data set and also on the quality of the multiple sequence alignment. The cytokinin-binding assay carried out as part of the analysis provides reliable evidence for the accuracy of the prediction of important ligand-binding sites in plant CHASE domains.

7.3 Loss and gain in the phosphatase family

The detailed investigation of the protein tyrosine phosphatase (PTP) family gave rise to a new classification of this family and led to the disclosure of new functions of the phosphatase fold. The analysis was motivated by a report of an inactive phosphatase in the literature (CUI *et al.* 1998). The computational detection of an inactive enzyme due to substitutions at catalytic positions would have appeared unreal. Especially considering the numerous sequencing errors and falsely predicted genes giving rise to proteins lacking whole exons, or possibly lacking catalytic positions, the presence of inactive enzymes seems questionable. However, the first report was based on experimental studies proving that the described phosphatase does not contain a catalytic activity and that it functions as phosphatase antagonist on the cellular level. This led to the question whether additional inactive phosphatases exist in metazoan genomes. In this thesis, the family of protein tyrosine phosphatases was studied in detail. By scanning the catalytic positions known from the literature for substitutions of catalytic essential amino acids, numerous additional inactive phosphatases were identified (see chapter 4.4.1 or table 4). This step was automatically carried out using HMMs generated from the SMART family alignment. The catalytic regions of proteins are in general highly conserved and align perfectly to an HMM. However, the performance of the HMM in identifying the correct catalytic position was confirmed by visually scanning the multiple sequence alignment generated with HMMalign of the same software package. This demonstrated that HMMs are successful in finding sequences, in which the catalytic positions have been substituted. Since inactive phosphatase sequence are less conserved in the catalytic region, a shift in the alignment of the protein sequence to the HMM cannot be excluded. Hence, the type of amino acid at substituted catalytic positions has to be treated with care.

The phylogenetic tree generated from the multiple sequence alignment revealed a cluster of inactive phosphatase domains, of which all correspond to the second domain of

receptor type phosphatases. Strikingly, this group splits into two further subgroups, one with conservative substitutions at catalytic positions (D2A) and one with random substitutions (D2B). The location of these inactive phosphatase domains adjacent to catalytically active phosphatase domains shows that these domains had been redundant in the protein and explains why these domains have accumulated mutations in the progress of evolution. This genetic procedure seems to be similar to the fate of duplicated genes, which can freely mutate and acquire new functions. Similarly, at the domain level, relaxation of selective constraint can result in complete loss of the domain or in the acquirement of advantageous mutations leading to a new function. The conservation of inactive domains between human, mouse and the pufferfish argues against the first possibility and demonstrates that the inactive domains are functionally important.

In order to get insight into the function of these inactive phosphatase domains, we pursued the strategy to identify functional sites, hoping that these would give us a hint on the cellular role of inactive phosphatases. In contrast to the GlcNAcase enzymes, which possess highly conserved catalytic and substrate binding residues and allow the easy identification of those, the phosphatases demanded a new method to study patterns of higher conserved amino acids. Functional sites that are neither involved in catalysis nor in binding of common substrates or cofactors are less conserved and do not stick out from multiple sequence alignments.

Taking advantage of the fact that the phosphatase family can be divided into groups of active and inactive domains with, obviously, different functions, we compared rates of evolution of single sites between the functionally different subgroups (chapter 4.4.2). Comparison of evolutionary rates has been used as measurement for functional differences in other protein families. The metazoan Myb family can be divided into two subfamilies based on evolutionary constraints of an acidic region. Slowly evolving acidic regions correlate with the protein's ability to function as transcriptional activator (SIMON *et al.* 2002). In the Pax family, which is involved in brain development, the evolutionary rates correlate inversely with the importance of the encoded protein as inferred from gene knockout studies. The Pax2 gene, which is exposing the smallest rate of evolution, thus the most conserved gene, is known to cause the most severe brain phenotype compared to other Pax family members (ABUROMIA *et al.* 2003).

In case of the phosphatase family, the comparison of evolutionary site rates detected several positions that are either evolving more slowly in the group of inactive enzymes or, in contrast, are evolving at higher rates. Positions with small rates of evolution concentrate around the former catalytic cleft in the D2B subgroup. This provides further evidence that this domain has completely lost its catalytic activity. Surprisingly, the subgroup presenting conservative substitutions at catalytic positions showed a completely different behaviour in this region. Only two positions in the D2A subgroup were evolving faster than in the compared group of active enzymes and most of the residues around the catalytic cleft seem to be under purifying selection. Complementary to this observation, there is experimental proof that the D2A domain is still able to specifically bind phosphorylated tyrosine residues. Although the catalytic amino acids have been substituted, conservation of the former catalytic cleft enables the binding of substrate. The cellular function of the D2A phosphatase domain can be inferred from the location adjacent to active phosphatase and the ability to bind phosphorylated tyrosine. The domain might be responsible for substrate specificity of the adjacent domain by contributing to the recognition of proteins that are multiple times phosphorylated.

Mapping of residues that are more conserved in inactive domains than in active phosphatases onto the three-dimensional structure revealed a second striking feature of the inactive domains. Both subgroups, D2A and D2B contain a cluster of slowly evolving residues, which could function as interaction interface (chapter 4.4.4). It is known that many phosphatases form homo- and heterodimers (BLANCHETOT *et al.* 2002). The interaction might involve this cluster. Simultaneously, the presence of an interaction interface suggests a function for both inactive phosphatase domains, that is the regulation of enzymatic activity by the enzymatically inactive adjacent phosphatase domain.

The method to study evolutionary rates of amino acid substitutions has been successfully applied in this thesis, and it has resulted in the suggestion of new functions for the inactive phosphatase domains. The inactive phosphatase domains have not only lost a function, that is to catalyse the hydrolysation of phospho-tyrosine, but also gained a new function. In the evolution of new pathways, the phosphatases can be compared to the GlcNAcases. The GlcNAcase catalytic mechanism was changed to enrich signalling pathways. Similarly, the phosphatase domain's fate included the loss of catalytic function and the adoption of a new beneficial function to contribute to signalling pathways.

7.4 New modules for regulatory tasks

The analysis of the phosphatase family has led to the question whether catalytically inactive domains are restricted to this family or whether they are also found in other enzyme families. The detailed and manually verified analysis of the phosphatases was successful in identifying inactive family members with an automatic HMM approach, so that this method was suitable for a large-scale analysis. The detection of substitutions at catalytic positions of enzymatic SMART domains presented in chapter 5 was a semi-automatic approach. The first step involved the manual extraction of catalytic positions from the primary literature, while the second step was the automatic scan of various peptide sets from sequenced metazoan genomes for enzymatic domains and the automatic inspection of substitutions at the prior extracted catalytic positions.

Numerous inactive enzymatic domains were found in many different domain families (chapter 5.2.1). As the large amount came as surprise, it was necessary to prove that the sequences are not an erroneous product of the various databases. Evidence that these catalytically inactive proteins exist in metazoan genomes was achieved by investigating whether an inactive domain corresponds to an orthologous inactive domain in a related species. Conservation of this feature in evolution would suggest that inactive domains are functional important and not subject to genetic drift. Comparison of orthologous genes in human, mouse and rat or *Drosophila* and *Anopheles*, respectively, showed that inactive domains are conserved across species (chapter 5.2.4). For the majority of inactive genes, an inactive domain was confirmed to be encoded in the orthologous gene. Few exceptions, in which the orthologous gene did not encode an inactive domain, are very likely caused by incomplete orthologous genes.

Knowing that inactive domains are an important part of cellular processes, their presumable function was investigated by classifying the domains in regard to their cellular localization. Interestingly, the group of nuclear domains contained hardly any inactive domains. Nuclear domains play such a fundamental role in cellular processes, that inactivating substitutions cannot be tolerated, apparently. In contrast, the group of signalling domains contained numerous inactive domains. Their function is very likely in regulating their active counterparts, as suggested for the protein tyrosine phosphatases. Examples from the literature suggest that enzymatically inactive domains regulate the enzymatic activity of functional homologues. Superoxide dismutase (SOD) is in sequence

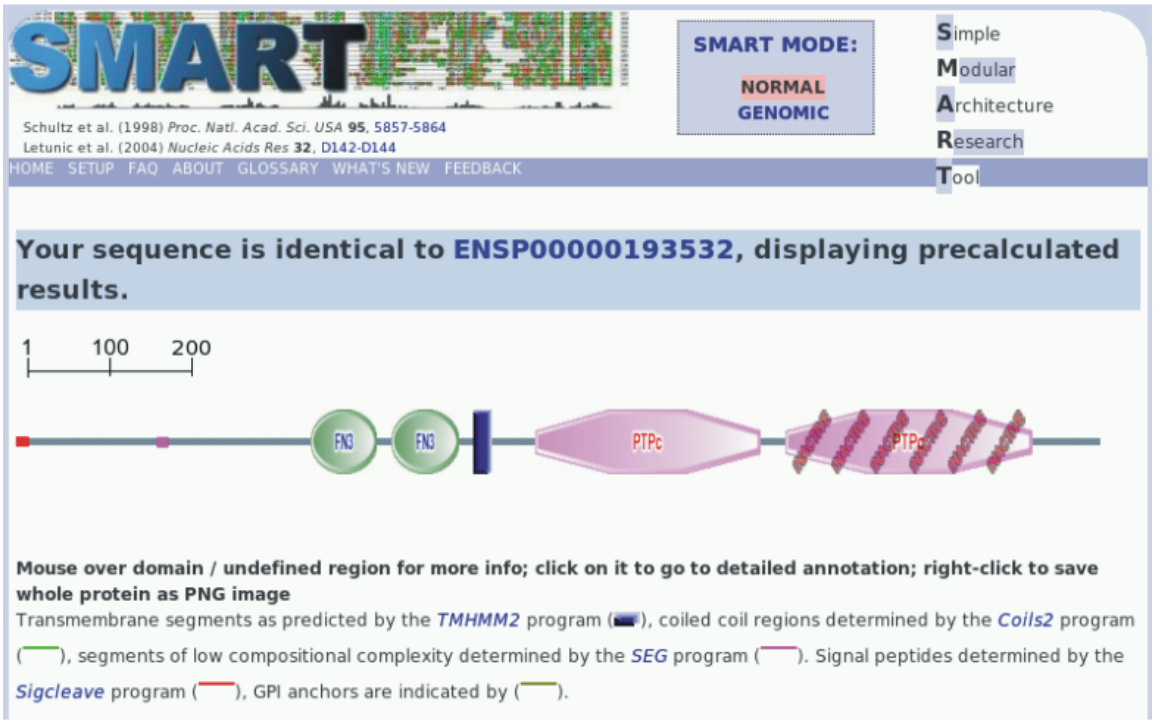
and structure highly similar to the catalytically inactive SOD copper chaperone and promotes its stable folding. Similarly, the Phospholipase A2 (PLA2) is regulated by the PLA2 inhibitor, which covers the catalytic center of the enzyme by forming a heterodimer and prevents substrate binding (BARTLETT *et al.* 2003). There are also a few cases of inactive enzymes among extracellular domains, such as the inactive lysozyme domain lactalbumin, which functions as regulatory subunit of lactose synthase.

The large number of inactive domains among signalling enzymes strengthens the assumption of a role in regulatory processes. To enlarge the amount of regulatory domains the cellular protein repertoire was used and slightly adjusted to perform new functions. The adaptation of existing modules is more efficient than de novo genesis of protein domains. Evolutionarywise, this is a very economic and fast way to increase the protein network in the cell.

The date of the invention of inactive enzymatic domains has also been estimated in this analysis. Inactive domains are present in all of the investigated metazoan genomes, but seem to be absent in yeast. This suggests the development of inactive domains to have occurred simultaneously to the emergence of metazoan. A functional complex regulatory network was the prerequisite for the development of multi-cellular organism, in which the cells can communicate and determine differentiation processes. Thus, the higher level of complexity emerging from new regulatory modules was an important invention leading to the onset of metazoan evolution: „It's not a bug – it's a feature".

Recently, an inactive enzyme-homologue was described in plants. The ADP-glucose pyrophosphorylase exists as heterotetrameric enzyme in plants, comprised of two homologous subunits. Ballicora and co-workers have shown that the inactive subunit diverged from the active enzyme and acquired differential modulatory properties. They resurrected the enzymatic activity of the inactive enzyme homologue by substituting two essential amino acid residues (BALLICORA *et al.* 2005). The ancient inactivation of catalytic domains suggests that this is a general evolutionary mechanism and has occurred wherever increased complexity and new regulatory functions were demanded. This mechanism has occurred multiple times in different protein families and also in different evolutionary lineages. It has even been suggested that many transcription regulators arose from ancient enzymes (ARAVIND and KOONIN 1998; GRISHIN 2001).

A



B

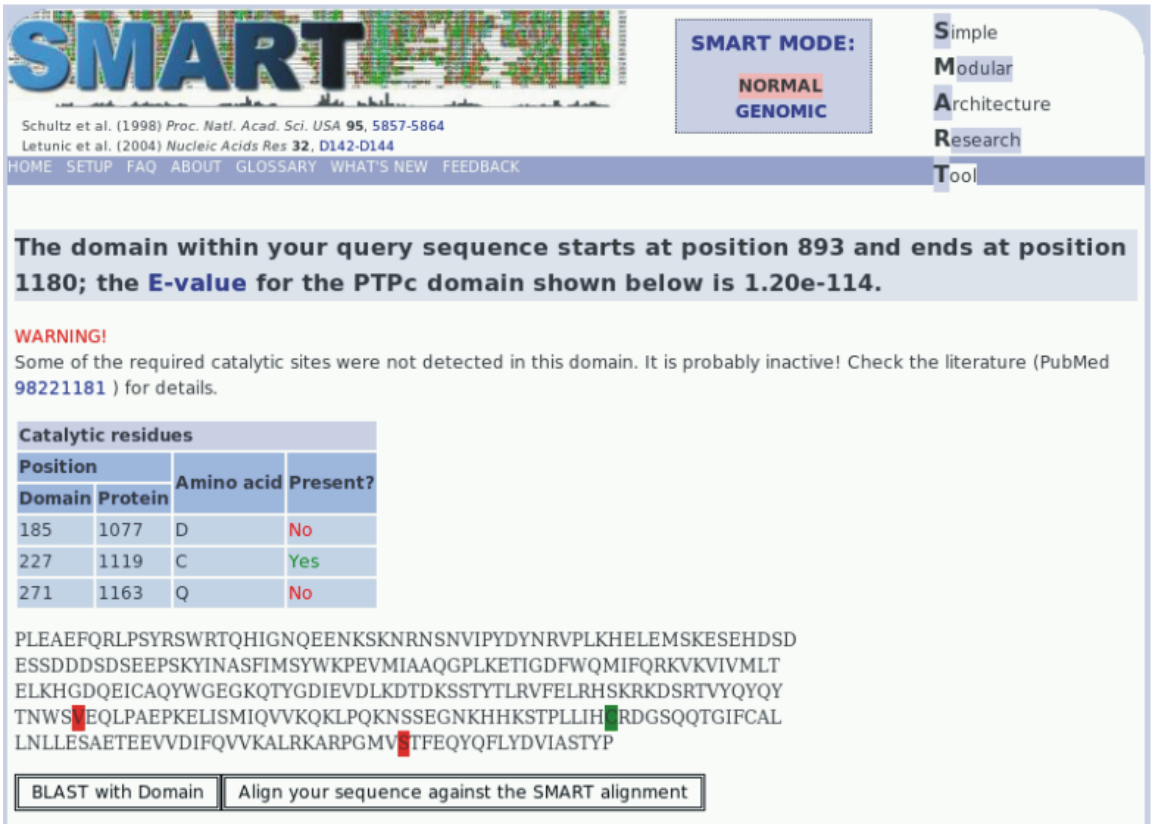


Figure 17: SMART output of a sequence analysis of the leukocyte common antigen precursor (CD45). (A) The sequence contains two protein tyrosine phosphatase (PTPc) domains, of which the carboxyterminal domain is catalytically inactive. (B) Detailed output of the inactive PTPc domain. Catalytic positions are listed in a table and highlighted in the domain sequence.

The dataset of catalytic amino acids and their corresponding position collected as part of the study allows now to scan new protein sequences for the presence of catalytic amino acids and to classify the proteins as active or inactive enzymes. In order to provide the dataset to the scientific community, it was integrated into the SMART database. Uncharacterized sequences can be submitted to the SMART online tool. If an enzymatic domain is detected, the sequence is subsequently searched for the presence of catalytic amino acids at the defined catalytic positions. Figure 17a shows the SMART output of a search with the leukocyte common antigen precursor (CD45; ENSEMBL id: ENSP00000193532).

The second PTPc domain is marked as inactive domain. By clicking on the domain icon, more detailed information becomes available in a new window (figure 17b). All catalytic positions are listed with regard to their location within the domain and within the protein. For each position the presence of the catalytic amino acid is indicated. In addition catalytic positions are coloured green if the catalytic amino acid is present and red if the sequence carries a substitution turning the domain into an inactive one.

7.5 The nature of functional sites

The surprising weak conservation and variability of amino acids at catalytic sites observed in enzymatic domains motivated us to investigate this trend more closely. The data set was enlarged and now comprised functional sites in general. The properties of functional sites were analysed in a fully automatic approach. Functional sites were extracted from experimentally derived structures and were compared in regard of their location within the protein domain family and their conservation of amino acids. For each domain family described in the SMART database, a multiple sequence alignment was generated with information on functional sites. The analysis revealed that functional sites can behave very differently in terms of their amino acid conservation and structural location. These findings are in accordance with other reports. Devos and co-workers studied pairwise structural alignments and extracted interacting sites similar to the approach used in this work. They found that functional sites can be quiet different despite a high level of sequence similarity and vice versa distantly related proteins can have a conserved binding site (DEVOS and VALENCIA 2000).

The collection of the data set of interacting sites collected in this analysis is fundamental for understanding the properties of functional sites in different protein domain families and it can now be used for the prediction of functional sites. In the following, two approaches are discussed how to use this data set for the characterisation of new protein sequences.

The first approach to predict functional sites is based on hidden Markov Models and has already been implemented by Friedrich and co-workers who have been motivated by the availability of the interaction data generated during the analysis described in chapter 6. In this approach the classical HMM is extended by a new state (figure 18, by courtesy of Torben Friedrich), or rather, the classical match state is split into an interacting match state (M_i) and a non-interacting match state (M_{ni}). The HMM was trained with the sequence data along with information on interacting sites obtained from the analysis presented in chapter 6 of this thesis. While the HMM is being trained, emission and transition probabilities are estimated from the sequence and interaction data. In order to predict interaction sites for a novel protein sequence, the sequence is aligned to the HMM and the most probable path between the different states is searched. The interaction profile HMM has been tested with the data generated in chapter 6. The outcome of the prediction was strongly depending on the type and ligand preferences of the domain (Friedrich, personal communication). For some domains, the performance was very successful, scoring 0.98 in a ROC (Receiver Operating Characteristic) curve analysis, while it was less reliable for other domains. It has to be further investigated whether the quality of prediction can be improved by using a subfamily based approach.

The second approach to predict functional sites is based on sequence similarity and considers evolutionary distances between the sequences. The information on functional sites is transferred from sequences of the dataset with experimentally determined functional sites. The probability of a functional site in the unknown sequence should correlate with the evolutionary distance. This approach is similar to the calculation of the score describing the conservation of interaction in chapter 6. The first step involves the generation of a multiple sequence alignment and the calculation of a distance tree for the novel sequence together with sequences from the dataset. Pairwise distances can be extracted from that tree. In the following, an interaction score is calculated according to the multiple sequence alignment for each position. The score should be based on the distances

corresponding to sequences that interact at this specific position and should correlate inversely with the phylogenetic distance. In other words, the score should strongly increase for interaction sites of very similar sequences, but marginally increase for diverged sequences. The score should also consider the physico-chemical properties of interacting residues as similar or conserved amino acids are more likely to interact. In addition it should be taken into account, that the conservation of functional sites depending on the protein family and prediction is easy or reliable for protein domain families like nuclear enzymes but rather difficult or unconfident for extracellular protein domains. The underlying assumption of the suggested method is that interacting sites are depending on their evolutionary origin and that closely related sequences are also similar in the usage of interacting sites.

A common drawback of these prediction methods is that a reliable outcome can only be accomplished from orthologous sequences, where conservation of function from the last common ancestor is assumed. The real data contains a mixture of orthologous and paralogous sequences and the major problem is the small number of experimentally determined, and thus verified functional sites for most protein domain families. With the increasing number of sequence data by various genome projects methods predicting

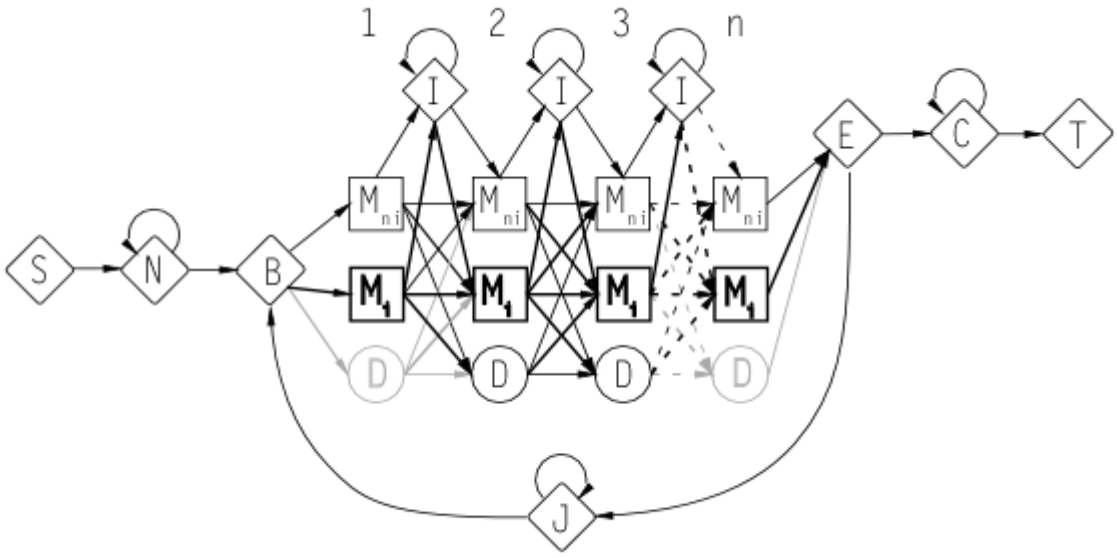


Figure 18: Interaction profile HMM. In addition to the conventional insert and delete states, the HMM contains interacting (M_i) and non-interacting match states (M_{ni}). Flanking regions of the domain are described with the S and N state for the N-terminal part of the protein and the C and T state for the carboxy-terminal part. Repetition of the domain is modelled with the J state. (Courtesy of Torben Friedrich)

functional sites based on orthologous sequences gain in attraction and have been reported recently (MIRNY and GELFAND 2002).

An important difference between the two here presented methods is that the HMM approach puts more weight on local similarities, where the outcome of the prediction for a certain position is dependent on the previous state in the alignment. Features of protein interaction predetermined by the secondary structure in the interaction region are neglected, such as interacting positions every three residues in α-helices. Of course secondary structure elements could be included in the model, but with the increasing complexity of the model, the quality of the prediction might suffer.

In contrast to the interaction profile HMM, the sequence similarity approach described above uses the overall sequence similarity to infer interaction sites. The information is transferred from real biological observations. The prediction is strongly orientated at the closest homologue and might fail if a new interaction region has developed in the novel sequence. At the moment both methods are still limited by the small number of experimentally determined sequences. However, the improvement of the prediction of functional sites will come along with the increasing number of available protein complexes.

A very important result of the analysis of functional sites was the recognition of the necessity to subclassify protein families into groups of functional related sequences. The exponentially growing sequence databases resulted in an increased number of sequences in each protein family. Although it was rather difficult to find a few related sequences a decade ago, the trend has now moved to organize proteins into subfamilies not only according to their function but also according to additional features like ligand preferences. In fact, many approaches towards this direction have been described in the literature. The PANTHER database divides protein families into subfamilies of related function based on a family tree, but it is not freely available for the scientific community (THOMAS *et al.* 2003). Hannenhalli and Russell also classify sub-families using functional differences such as substrate specificity and analyze the positional entropy in the alignment to search for sub-family defining residues (HANNENHALLI and RUSSELL 2000). The FunShift database uses BETE (SJOLANDER 1998), an automatic method to divide families into subfamilies based on relative entropy, to subclassify the protein families from Pfam (ABHIMAN and SONNHAMMER 2005). Following, the subfamilies are compared in regard of evolutionary

constraints, where two types of functional shifts can be differentiated: Rate shifting sites vary in evolutionary site rates. In contrast, conservation shifting sites are conserved in two subfamilies but vary in the type of amino acid. These different behaviours are also referred to as type I and type II functional divergence, respectively.

In the future, methods that include evolutionary and experimental information and consider family specific properties will become very attractive and are promising tools to predict functional sites of the exponentially growing uncharacterised sequences.

7.6 Outlook

During the last years the annotation of protein function developed to extremes. On one side, large-scale analyses like genomic context methods gain significance, but on the other side, in-depth studies become more important to understand fine-tuned functions and interactions of proteins. At the moment we still have to deal with many problems in the field of protein function prediction and it is important to focus on combining large-scale and in-depth studies to obtain a complete picture and on the integration of all the data produced by the ‘omics communities. Various data types in addition to sequences and structures, such as expression data, genomic context, interaction data, time and spatial regulation have to be fused to understand the organization of cellular processes and finally higher order biological system. The next few years will encounter the development of many new methods to integrate new findings from genomics and proteomics into our understanding of nature.

8 Summary

The growing number of uncharacterised sequences in public databases has turned the prediction of protein function into a challenging research field. Traditional annotation methods are often error-prone due to the small subset of proteins with experimentally verified function. Goal of this thesis was to analyse the function and evolution of protein domains in order to understand molecular processes in the cell. The focus was on signalling domains of little understood function, as well as on functional sites of protein domains in general.

Glucosaminidases (GlcNAcases) represent key enzymes in signal transduction pathways. Together with glucosamine transferases, they serve as molecular switches, similar to kinases and phosphatases. Little was known about the molecular function and structure of the GlcNAcases. In this thesis, the GlcNAcases were identified as remote homologues of N-acetyltransferases. By comparing the homologous sequences, I was able to predict functional sites of the GlcNAcase family and to identify the GlcNAcases as the first family member of the acetyltransferase superfamily with a distinct catalytic mechanism, which is not involved in the transfer of acetyl groups.

In a similar approach, the sensor domain of a plant hormone receptor was studied. I was able to predict putative ligand-binding sites by comparing evolutionary constraints in functionally diverged subfamilies. Most of the putative ligand-binding sites have been experimentally confirmed in the meantime.

Due to the importance of enzymes involved in cellular signalling, it seems impossible to find substitutions of catalytic amino acids that turn them catalytically inactive. Nevertheless, by scanning catalytic positions of the protein tyrosine phosphatase families, I found many inactive domains among single domain and tandem domain phosphatases in metazoan proteomes. In addition, I found that inactive phosphatases are conserved throughout evolution, which led to the question about the function of these catalytically inactive phosphatase domains. An analysis of evolutionary site rates of amino acid substitutions revealed a cluster of conserved residues in the apparently redundant domain of tandem phosphatases. This putative regulatory center might be responsible for

the experimentally verified dimerization of the active and inactive domain in order to control the catalytic activity of the active phosphatase domain. Moreover, I detected a subgroup of inactive phosphatases, which presumably functions in substrate recognition, based on different evolutionary site rates within the phosphatase family.

The characterization of these new regulatory modules in the phosphatase family raised the question whether inactivation of enzymes is a more general evolutionary mechanism to enlarge signalling pathways and whether inactive domains are also found in other enzyme families. A large-scale analysis of substitutions at catalytic positions of enzymatic domains was performed in this work. I identified many domains with inactivating substitutions in various enzyme families. Signalling domains harbour a particular high occurrence of catalytically inactive domains indicating that these domains have evolved to modulate existing regulatory pathways. Furthermore, it was shown that inactivation of enzymes by single substitutions happened multiple times independently in evolution.

The surprising variability of amino acids at catalytic positions was decisive for a subsequent analysis of the diversity of functional sites in general. Using functional residues extracted from structural complexes I could show that functional sites of protein domains do not only vary in their type of amino acid but also in their structural location within the domain. In the process of evolution, protein domains have arisen from duplication events and subsequently adapted to new binding partners and developed new functions, which is reflected in the high variability of functional sites. However, great differences exist between domain families. The analysis demonstrated that functional sites of nuclear domains are more conserved than functional sites of extracellular domains. Furthermore, the type of ligand influences the degree of conservation, for example ion binding sites are more conserved than peptide binding sites.

The work presented in this thesis has led to the detection of functional sites in various protein domains involved in signalling pathways and it has resulted in insights into the molecular function of those domains. In addition, properties of functional sites of protein domains were revealed. This knowledge can be used in the future to improve the prediction of protein function and to identify functional sites of proteins.

9 Zusammenfassung

Durch den rasanten Anstieg unbekannter Proteinsequenzen in öffentlichen Datenbanken ist die Vorhersage der Proteinfunktion zu einem herausfordernden Forschungsgebiet geworden. Herkömmliche Annotationsmethoden sind häufig fehlerhaft, da nur einem kleinen Teil der Proteine experimentell eine Funktion zugewiesen werden konnte. Ziel der hier vorliegenden Arbeit war es, die Funktion und Evolution von Proteindomänen in Hinblick auf die molekularen Vorgänge innerhalb der Zelle zu untersuchen. Der Schwerpunkt lag auf Signaldomänen mit unbekannter Funktion und auf funktionell wichtigen Positionen in Domänen.

Glucosaminidasen (GlcNAcasen) spielen eine wichtige Rolle in Signaltransduktionswegen. Zusammen mit den Glucosamintransferasen dienen sie als molekulare Schalter, ähnlich den Kinasen und Phosphatasen, jedoch war sehr wenig über ihre molekulare Funktion, sowie über ihre Struktur bekannt. In dieser Studie wurde die entfernte Verwandtschaft der GlcNAcasen zu den Acetyltransferasen gezeigt. Durch den Vergleich von homologen Sequenzen konnte ich funktionelle Positionen vorhersagen und die GLcNAcasen als erstes Mitglied der Acetyltransferasen-Superfamilie mit einem neuen katalytischen Mechanismus identifizieren, der nicht den Transfer von Acetylgruppen vermittelt.

In einem ähnlichen Ansatz wurde die Sensordomäne eines Hormonrezeptors aus Pflanzen untersucht. Dabei konnte ich durch den Vergleich von evolutiven Zwängen in funktionell unterschiedlichen Subfamilien wahrscheinliche Liganden-bindende Positionen bestimmen. Die meisten dieser Vorhersagen sind in der Zwischenzeit experimentell bestätigt worden.

Aufgrund der entscheidenden Bedeutung von enzymatischen Domänen in Signal-transduktionsprozessen erscheint es unmöglich, Substitutionen von katalytischen Amino-säuren zu finden, die die Domäne inaktivieren würden. Dennoch habe ich in einer Analyse der katalytischen Positionen in der Proteintyrosinphosphatase-Familie viele inaktive Domänen in Einzel- und Tandem-Domänen-Phosphatasen in den Proteomen von Metazoa gefunden. Ich habe zusätzlich beobachtet, dass die inaktiven Domänen in der Evolution

konserviert sind, was die Frage aufwirft, welche Funktion diese katalytisch inaktiven Domänen haben. Eine Analyse der Evolutionsraten von Aminosäuresubstitutionen identifizierte eine Ansammlung von konservierten Positionen in der scheinbar überflüssigen inaktiven Domäne von Tandemphosphatasen. Dieser möglicherweise regulatorische Bereich könnte für die Dimerisierung der aktiven und inaktiven Domäne verantwortlich sein, welche experimentell nachgewiesen wurde, sowie für die Regulation der katalytischen Aktivität der Phosphatasedomäne. Außerdem habe ich durch die unterschiedlichen Evolutionsraten eine Untergruppe der inaktiven Phosphatasen entdeckt, die wahrscheinlich an der Substraterkennung beteiligt ist.

Die Charakterisierung dieser neuen regulatorischen Module in der Phosphatase-Familie führte zu der Frage, ob die Inaktivierung von Enzymen ein allgemeiner Mechanismus in der Evolution ist, um Signaltransduktionswege zu erweitern, und ob es auch in anderen Enzymfamilien inaktive Domänen gibt. Dazu wurde eine umfassende Analyse durchgeführt, um Substitutionen an katalytischen Positionen in enzymatischen Domänen zu untersuchen. Ich habe in vielen Domänen aus unterschiedlichen Enzymfamilien inaktivierende Substitutionen gefunden. Einen besonders hohen Anteil an katalytisch inaktiven Domänen gibt es in Signaldomänen, was zeigt, daß diese Domänen entstanden sind, um existierende regulatorische Netze zu modifizieren. Es konnte ferner gezeigt werden, daß die Inaktivierung von Enzymen durch einzelne Subsitutionen mehrmals unabhängig voneinander in der Evolution stattgefunden hat.

Die überraschende Variabilität von Aminosäuren an katalytischen Positionen war ausschlaggebend für eine anschließende, allgemeinere Analyse von funktionellen Positionen. Mit Hilfe von funktionellen Positionen, die aus strukturellen Komplexen extrahiert wurden, konnte ich zeigen, dass funktionelle Positionen nicht nur in der Aminosäure, sondern auch in ihrer Lokalisation innerhalb der Struktur variieren. Im Laufe der Evolution haben sich Domänen aus Duplikationsprozessen gebildet, sich neuen Bindungspartnern angepasst und neue Funktionen entwickelt, was sich nun in der hohen Variabilität ihrer funktionellen Positionen widerspiegelt. Dennoch gibt es große Unterschiede zwischen Domänenfamilien. Die Analyse hat gezeigt, dass funktionelle Positionen von nuklearen Domänen viel stärker konserviert sind, als jene von extrazellulären Domänen. Außerdem beeinflusst die Art des Liganden den Grad der

Konservierung, so sind z. B. Ionen-bindende Positionen stärker konserviert als Peptid-bindende Positionen.

Die hier vorgestellte Studie beschreibt funktionelle Positionen in verschiedenen an Signaltransduktionswegen beteiligten Proteindomänen und liefert Einblicke in ihre molekulare Funktion. Außerdem wurden Eigenschaften von funktionell wichtigen Positionen aufgezeigt. Diese Erkenntnisse können in Zukunft zur Optimierung der Vorhersage von Proteinfunktionen und zur Identifikation von funktionellen Positionen genutzt werden.

10 References

ABHIMAN S. and SONNHAMMER E.L. 2005. FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res **33:** D197-200.

ABUROMIA R., KHANER O. and SIDOW A. 2003. Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. J Struct Funct Genomics **3:** 45-52.

ADAMS M.D., CELNIKER S.E., HOLT R.A., EVANS C.A., GOCAYNE J.D., AMANATIDES P.G., SCHERER S.E., LI P.W., HOSKINS R.A., GALLE R.F., *et al.* 2000. The genome sequence of Drosophila melanogaster. Science **287:** 2185-2195.

ALEX L.A. and SIMON M.I. 1994. Protein histidine kinases and signal transduction in prokaryotes and eukaryotes. Trends Genet **10:** 133-138.

ALOY P., CEULEMANS H., STARK A. and RUSSELL R.B. 2003. The relationship between sequence and interaction divergence in proteins. J Mol Biol **332:** 989-998.

ALTSCHUL S.F., GISH W., MILLER W., MYERS E.W. and LIPMAN D.J. 1990. Basic local alignment search tool. J Mol Biol **215:** 403-410.

ALTSCHUL S.F., MADDEN T.L., SCHAFFER A.A., ZHANG J., ZHANG Z., MILLER W. and LIPMAN D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

AMITAI G., SHEMESH A., SITBON E., SHKLAR M., NETANELY D., VENGER I. and PIETROKOVSKI S. 2004. Network analysis of protein structures identifies functional residues. J Mol Biol **344:** 1135-1146.

ANANTHARAMAN V. and ARAVIND L. 2001. The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors. Trends Biochem Sci **26:** 579-582.

ANDERSEN J.N., MORTENSEN O.H., PETERS G.H., DRAKE P.G., IVERSEN L.F., OLSEN O.H., JANSEN P.G., ANDERSEN H.S., TONKS N.K. and MOLLER N.P. 2001. Structural and evolutionary relationships among protein tyrosine phosphatase domains. Mol Cell Biol **21:** 7117-7136.

ANDRADE M.A., BROWN N.P., LEROY C., HOERSCH S., DE DARUVAR A., REICH C., FRANCHINI A., TAMAMES J., VALENCIA A., OUZOUNIS C., *et al.* 1999. Automated genome sequence analysis and annotation. Bioinformatics **15:** 391-412.

ANDREEVA A., HOWORTH D., BRENNER S.E., HUBBARD T.J., CHOTHIA C. and MURZIN A.G. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res **32:** D226-229.

APARICIO S., CHAPMAN J., STUPKA E., PUTNAM N., CHIA J.M., DEHAL P., CHRISTOFFELS A., RASH S., HOON S., SMIT A., *et al.* 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297:** 1301-1310.

ARABIDOPSIS GENOME INITIATIVE 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-815.

ARAVIND L. and KOONIN E.V. 1998. Eukaryotic transcription regulators derive from ancient enzymatic domains. Curr Biol **8:** R111-113.

ARMON A., GRAUR D. and BEN-TAL N. 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol **307:** 447-463.

ARNOLD C.S., JOHNSON G.V., COLE R.N., DONG D.L., LEE M. and HART G.W. 1996. The microtubule-associated protein tau is extensively modified with O-linked N-acetylglucosamine. J Biol Chem **271:** 28741-28744.

ASANTE-APPIAH E. and KENNEDY B.P. 2003. Protein tyrosine phosphatases: the quest for negative regulators of insulin action. Am J Physiol Endocrinol Metab **284:** E663-670.

ATTWOOD T.K., FLOWER D.R., LEWIS A.P., MABEY J.E., MORGAN S.R., SCORDIS P., SELLEY J.N. and WRIGHT W. 1999. PRINTS prepares for the new millennium. Nucleic Acids Res **27:** 220-225.

BALDI P., CHAUVIN Y., HUNKAPILLER T. and MCCLURE M.A. 1994. Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci U S A **91:** 1059-1063.

BALLICORA M.A., DUBAY J.R., DEVILLERS C.H. and PREISS J. 2005. Resurrecting the ancestral enzymatic role of a modulatory subunit. J Biol Chem **280:** 10189-10195.

BARTLETT G.J., BORKAKOTI N. and THORNTON J.M. 2003. Catalysing new reactions during evolution: economy of residues and mechanism. J Mol Biol **331:** 829-860.

BATEMAN A., BIRNEY E., CERRUTI L., DURBIN R., ETWILLER L., EDDY S.R., GRIFFITHS-JONES S., HOWE K.L., MARSHALL M. and SONNHAMMER E.L. 2002. The Pfam protein families database. Nucleic Acids Res **30:** 276-280.

BATEMAN A., COIN L., DURBIN R., FINN R.D., HOLLICH V., GRIFFITHS-JONES S., KHANNA A., MARSHALL M., MOXON S., SONNHAMMER E.L., *et al.* 2004. The Pfam protein families database. Nucleic Acids Res **32:** D138-141.

BECKMANN G., HANKE J., BORK P. and REICH J.G. 1998. Merging extracellular domains: fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins. J Mol Biol **275:** 725-730.

BELFAIZA J., PARSOT C., MARTEL A., DE LA TOUR C.B., MARGARITA D., COHEN G.N. and SAINT-GIRONS I. 1986. Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. Proc Natl Acad Sci U S A **83:** 867-871.

BENSON D.A., KARSCH-MIZRACHI I., LIPMAN D.J., OSTELL J. and WHEELER D.L. 2005. GenBank. Nucleic Acids Res **33:** D34-38.

BERMAN H.M., BHAT T.N., BOURNE P.E., FENG Z., GILLILAND G., WEISSIG H. and WESTBROOK J. 2000. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol **7 Suppl:** 957-959.

BERNAL A., EAR U. and KYRPIDES N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res **29:** 126-127.

BHARATHAN G., JANSSEN B.J., KELLOGG E.A. and SINHA N. 1999. Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. Mol Biol Evol **16:** 553-563.

BHATNAGAR R.S., FUTTERER K., FARAZI T.A., KOROLEV S., MURRAY C.L., JACKSON-MACHELSKI E., GOKEL G.W., GORDON J.I. and WAKSMAN G. 1998. Structure of N-myristoyltransferase with bound myristoylCoA and peptide substrate analogs. Nat Struct Biol **5:** 1091-1097.

BHATNAGAR R.S., FUTTERER K., WAKSMAN G. and GORDON J.I. 1999. The structure of myristoyl-CoA:protein N-myristoyltransferase. Biochim Biophys Acta **1441:** 162-172.

BHINGE A., CHAKRABARTI P., UTHANUMALLIAN K., BAJAJ K., CHAKRABORTY K. and VARADARAJAN R. 2004. Accurate detection of protein:ligand binding sites using molecular dynamics simulations. Structure (Camb) **12:** 1989-1999.

BILWES A.M., DEN HERTOG J., HUNTER T. and NOEL J.P. 1996. Structural basis for inhibition of receptor protein-tyrosine phosphatase-alpha by dimerization. Nature **382:** 555-559.

BLAKE C.C., KOENIG D.F., MAIR G.A., NORTH A.C., PHILLIPS D.C. and SARMA V.R. 1965. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. Nature **206:** 757-761.

BLANCHETOT C. and DEN HERTOG J. 2000. Multiple interactions between receptor protein-tyrosine phosphatase (RPTP) alpha and membrane-distal protein-tyrosine phosphatase domains of various RPTPs. J Biol Chem **275:** 12446-12452.

BLANCHETOT C., TERTOOLEN L.G., OVERVOORDE J. and DEN HERTOG J. 2002. Intra- and intermolecular interactions between intracellular domains of receptor protein-tyrosine phosphatases. J Biol Chem **277:** 47263-47269.

BLISKA J.B., CLEMENS J.C., DIXON J.E. and FALKOW S. 1992. The Yersinia tyrosine phosphatase: specificity of a bacterial virulence determinant for phosphoproteins in the J774A.1 macrophage. J Exp Med **176:** 1625-1630.

BLOUIN C., BOUCHER Y. and ROGER A.J. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res **31:** 790-797.

BLOW D.M., BIRKTOFT J.J. and HARTLEY B.S. 1969. Role of a buried acid group in the mechanism of action of chymotrypsin. Nature **221:** 337-340.

BOECKMANN B., BAIROCH A., APWEILER R., BLATTER M.C., ESTREICHER A., GASTEIGER E., MARTIN M.J., MICHOUD K., O'DONOVAN C., PHAN I., *et al.* 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31:** 365-370.

BORDO D. and BORK P. 2002. The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. EMBO Rep **3:** 741-746.

BORK P. and BAIROCH A. 1996. Go hunting in sequence databases but watch out for the traps. Trends Genet **12:** 425-427.

BORK P. and KOONIN E.V. 1998. Predicting functions from protein sequences--where are the bottlenecks? Nat Genet **18:** 313-318.

BRENNER S.E., CHOTHIA C., HUBBARD T.J. and MURZIN A.G. 1996. Understanding protein structure: using scop for fold interpretation. Methods Enzymol **266:** 635-643.

BRU C., COURCELLE E., CARRERE S., BEAUSSE Y., DALMAR S. and KAHN D. 2005. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res **33:** D212-215.

BUCHAN D.W., RISON S.C., BRAY J.E., LEE D., PEARL F., THORNTON J.M. and ORENGO C.A. 2003. Gene3D: structural assignments for the biologist and bioinformaticist alike. Nucleic Acids Res **31:** 469-473.

C. ELEGANS SEQUENCING CONSORTIUM 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science **282:** 2012-2018.

CAMPBELL S.J., GOLD N.D., JACKSON R.M. and WESTHEAD D.R. 2003. Ligand binding: functional site location, similarity and docking. Curr Opin Struct Biol **13:** 389-395.

CASARI G., SANDER C. and VALENCIA A. 1995. A method to predict functional residues in proteins. Nat Struct Biol **2:** 171-178.

CHENG X., COLE R.N., ZAIA J. and HART G.W. 2000. Alternative O-glycosylation/O-phosphorylation of the murine estrogen receptor beta. Biochemistry **39:** 11609-11620.

CHOTHIA C. 1991. Asymmetry in protein structures. Ciba Found Symp **162:** 36-49; discussion 49-57.

CHOTHIA C. 1992. Proteins. One thousand families for the molecular biologist. Nature **357:** 543-544.

CHOU T.Y., HART G.W. and DANG C.V. 1995. c-Myc is glycosylated at threonine 58, a known phosphorylation site and a mutational hot spot in lymphomas. J Biol Chem **270:** 18961-18965.

CLACKSON T. and WELLS J.A. 1995. A hot spot of binding energy in a hormone-receptor interface. Science **267:** 383-386.

COMTESSE N., MALDENER E. and MEESE E. 2001. Identification of a nuclear variant of MGEA5, a cytoplasmic hyaluronidase and a beta-N-acetylglucosaminidase. Biochem Biophys Res Commun **283:** 634-640.

COPLEY R.R., RUSSELL R.B. and PONTING C.P. 2001. Sialidase-like Asp-boxes: sequence-similar structures within different protein folds. Protein Sci **10:** 285-292.

COTTER D.A., DUNBAR A.J., BUCONJIC S.D. and WHELDRAKE J.F. 1999. Ammonium phosphate in sori of Dictyostelium discoideum promotes spore dormancy through stimulation of the osmosensor ACG. Microbiology **145 (Pt 8):** 1891-1901.

CUFF J.A., CLAMP M.E., SIDDIQUI A.S., FINLAY M. and BARTON G.J. 1998. JPred: a consensus secondary structure prediction server. Bioinformatics **14:** 892-893.

CUI L., YU W.P., DEAIZPURUA H.J., SCHMIDLI R.S. and PALLEN C.J. 1996. Cloning and characterization of islet cell antigen-related protein-tyrosine phosphatase (PTP), a novel receptor-like PTP and autoantigen in insulin-dependent diabetes. J Biol Chem **271:** 24817-24823.

CUI X., DE VIVO I., SLANY R., MIYAMOTO A., FIRESTEIN R. and CLEARY M.L. 1998. Association of SET domain and myotubularin-related proteins modulates growth control. Nat Genet **18:** 331-337.

DAHIA P.L. 2000. PTEN, a unique tumor suppressor gene. Endocr Relat Cancer **7:** 115-129.

DANDEKAR T., SNEL B., HUYNEN M. and BORK P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci **23:** 324-328.

DAS S. and SMITH T.F. 2000. Identifying nature's protein Lego set. Adv Protein Chem **54:** 159-183.

DE VIVO I., CUI X., DOMEN J. and CLEARY M.L. 1998. Growth stimulation of primary B cell precursors by the anti-phosphatase Sbf1. Proc Natl Acad Sci U S A **95:** 9471-9476.

DEHAL P., SATOU Y., CAMPBELL R.K., CHAPMAN J., DEGNAN B., DE TOMASO A., DAVIDSON B., DI GREGORIO A., GELPKE M., GOODSTEIN D.M., *et al.* 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science **298:** 2157-2167.

DESHPANDE N., ADDESS K.J., BLUHM W.F., MERINO-OTT J.C., TOWNSEND-MERINO W., ZHANG Q., KNEZEVICH C., XIE L., CHEN L., FENG Z., *et al.* 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res **33 Database Issue:** D233-237.

DEVEDJIEV Y., POPOV A., ATANASOV B. and BARTUNIK H.D. 1997. X-ray structure at 1.76 A resolution of a polypeptide phospholipase A2 inhibitor. J Mol Biol **266:** 160-172.

DEVOS D. and VALENCIA A. 2000. Practical limits of function prediction. Proteins **41:** 98-107.

DOERKS T., BAIROCH A. and BORK P. 1998. Protein annotation: detective work for function prediction. Trends Genet **14:** 248-250.

DURAND D. 2003. Vertebrate evolution: doubling and shuffling with a full deck. Trends Genet **19:** 2-5.

DYDA F., KLEIN D.C. and HICKMAN A.B. 2000. GCN5-related N-acetyltransferases: a structural overview. Annu Rev Biophys Biomol Struct **29:** 81-103.

EDDY S.R. 1996. Hidden Markov models. Curr Opin Struct Biol **6:** 361-365.

EDDY S.R. 1998. Profile hidden Markov models. Bioinformatics **14:** 755-763.

ENRIGHT A.J., ILIOPOULOS I., KYRPIDES N.C. and OUZOUNIS C.A. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature **402:** 86-90.

FANI R., LIO P., CHIARELLI I. and BAZZICALUPO M. 1994. The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes. J Mol Evol **38:** 489-495.

FAUMAN E.B., YUVANIYAMA C., SCHUBERT H.L., STUCKEY J.A. and SAPER M.A. 1996. The X-ray crystal structures of Yersinia tyrosine phosphatase with bound tungstate and nitrate. Mechanistic implications. J Biol Chem **271:** 18780-18788.

FELBERG J. and JOHNSON P. 2000. Stable interdomain interaction within the cytoplasmic domain of CD45 increases enzyme stability. Biochem Biophys Res Commun **271:** 292-298.

FELSENSTEIN J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5:** 164-166.

FLEISCHMANN R.D., ADAMS M.D., WHITE O., CLAYTON R.A., KIRKNESS E.F., KERLAVAGE A.R., BULT C.J., TOMB J.F., DOUGHERTY B.A., MERRICK J.M., *et al.* 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:** 496-512.

FRASER C.M., GOCAYNE J.D., WHITE O., ADAMS M.D., CLAYTON R.A., FLEISCHMANN R.D., BULT C.J., KERLAVAGE A.R., SUTTON G., KELLEY J.M., *et al.* 1995. The minimal gene complement of Mycoplasma genitalium. Science **270:** 397-403.

GABALDON T. and HUYNEN M.A. 2004. Prediction of protein function and pathways in the genome era. Cell Mol Life Sci **61:** 930-944.

GALPERIN M.Y. and KOONIN E.V. 1998. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol **1:** 55-67.

GALPERIN M.Y. and KOONIN E.V. 2000. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol **18:** 609-613.

GALTIER N., GOUY M. and GAUTIER C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci **12:** 543-548.

GAO Y., WELLS L., COMER F.I., PARKER G.J. and HART G.W. 2001. Dynamic O-glycosylation of nuclear and cytosolic proteins: cloning and characterization of a neutral, cytosolic beta-N-acetylglucosaminidase from human brain. J Biol Chem **276:** 9838-9845.

GERSTEIN M. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des **3:** 497-512.

GILBERT W. 1978. Why genes in pieces? Nature **271:** 501.

GLUSMAN G., YANAI I., RUBIN I. and LANCET D. 2001. The complete human olfactory subgenome. Genome Res **11:** 685-702.

GOFFEAU A., BARRELL B.G., BUSSEY H., DAVIS R.W., DUJON B., FELDMANN H., GALIBERT F., HOHEISEL J.D., JACQ C., JOHNSTON M.*, et al.* 1996. Life with 6000 genes. Science **274:** 546, 563-547.

GOODSTADT L. and PONTING C.P. 2001. CHROMA: consensus-based colouring of multiple alignments for publication. Bioinformatics **17:** 845-846.

GOUGH J., KARPLUS K., HUGHEY R. and CHOTHIA C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol **313:** 903-919.

GRISHIN N.V. 2001. Mh1 domain of Smad is a degraded homing endonuclease. J Mol Biol **307:** 31-37.

GU X. and VANDER VELDEN K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics **18:** 500-501.

HAFT D.H., SELENGUT J.D. and WHITE O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res **31:** 371-373.

HANNENHALLI S.S. and RUSSELL R.B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol **303:** 61-76.

HANOVER J.A. 2001. Glycan-dependent signaling: O-linked N-acetylglucosamine. Faseb J **15:** 1865-1876.

HECKEL D., COMTESSE N., BRASS N., BLIN N., ZANG K.D. and MEESE E. 1998. Novel immunogenic antigen homologous to hyaluronidase in meningioma. Hum Mol Genet **7:** 1859-1872.

HENIKOFF J.G. and HENIKOFF S. 1996. Blocks database and its applications. Methods Enzymol **266:** 88-105.

HEYL A. and SCHMULLING T. 2003. Cytokinin signal perception and transduction. Curr Opin Plant Biol **6:** 480-488.

HOROWITZ N.H. and NETZENBERG R.L. 1965. Biochemical Aspects of Genetics. Annu Rev Biochem **34:** 527-564.

HUGHEY R. and KROGH A. 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci **12:** 95-107.

HULO N., SIGRIST C.J., LE SAUX V., LANGENDIJK-GENEVAUX P.S., BORDOLI L., GATTIKER A., DE CASTRO E., BUCHER P. and BAIROCH A. 2004. Recent improvements to the PROSITE database. Nucleic Acids Res **32:** D134-137.

HUNTER T. 1998. Anti-phosphatases take the stage. Nat Genet **18:** 303-305.

HUYNEN M.A. and BORK P. 1998. Measuring genome evolution. Proc Natl Acad Sci U S A **95:** 5849-5856.

HUYNEN M.A. and SNEL B. 2000. Gene and context: integrative approaches to genome analysis. Adv Protein Chem **54:** 345-379.

INOUE T., HIGUCHI M., HASHIMOTO Y., SEKI M., KOBAYASHI M., KATO T., TABATA S., SHINOZAKI K. and KAKIMOTO T. 2001. Identification of CRE1 as a cytokinin receptor from Arabidopsis. Nature **409:** 1060-1063.

ISAKOFF S.J., CARDOZO T., ANDREEV J., LI Z., FERGUSON K.M., ABAGYAN R., LEMMON M.A., ARONHEIM A. and SKOLNIK E.Y. 1998. Identification and analysis of PH domain-containing targets of phosphatidylinositol 3-kinase using a novel in vivo assay in yeast. Embo J **17:** 5374-5387.

ISLAM S.A., LUO J. and STERNBERG M.J. 1995. Identification and analysis of domains in proteins. Protein Eng **8:** 513-525.

JEFFERY C.J. 2003. Moonlighting proteins: old proteins learning new tricks. Trends Genet **19:** 415-417.

JENSEN R.A. 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol **30:** 409-425.

JOHNSON P., OSTERGAARD H.L., WASDEN C. and TROWBRIDGE I.S. 1992. Mutational analysis of CD45. A leukocyte-specific protein tyrosine phosphatase. J Biol Chem **267:** 8035-8041.

JONES S. and THORNTON J.M. 1997. Analysis of protein-protein interaction sites using surface patches. J Mol Biol **272:** 121-132.

JONES S. and THORNTON J.M. 2004. Searching for functional sites in protein structures. Curr Opin Chem Biol **8:** 3-7.

JUSTEMENT L.B. 2001. The role of the protein tyrosine phosphatase CD45 in regulation of B lymphocyte activation. Int Rev Immunol **20:** 713-738.

KALHAMMER G., BAHLER M., SCHMITZ F., JOCKEL J. and BLOCK C. 1997. Ras-binding domains: predicting function versus folding. FEBS Lett **414:** 599-602.

KAMBAYASHI Y., TAKAHASHI K., BARDHAN S. and INAGAMI T. 1995. Cloning and expression of protein tyrosine phosphatase-like protein derived from a rat pheochromocytoma cell line. Biochem J **306:** 331-335.

KANZ C., ALDEBERT P., ALTHORPE N., BAKER W., BALDWIN A., BATES K., BROWNE P., VAN DEN BROEK A., CASTRO M., COCHRANE G., *et al.* 2005. The EMBL Nucleotide Sequence Database. Nucleic Acids Res **33:** D29-33.

KARTHA G., BELLO J. and HARKER D. 1967. Tertiary structure of ribonuclease. Nature **213:** 862-865.

KESKIN O., MA B. and NUSSINOV R. 2005. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol **345:** 1281-1294.

KIMBER M.S., NACHMAN J., CUNNINGHAM A.M., GISH G.D., PAWSON T. and PAI E.F. 2000. Structural basis for specificity switching of the Src SH2 domain. Mol Cell **5:** 1043-1049.

KINOSHITA K. and NAKAMURA H. 2005. Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci **14:** 711-718.

KITAYAMA H., SUGIMOTO Y., MATSUZAKI T., IKAWA Y. and NODA M. 1989. A ras-related gene with transformation suppressor activity. Cell **56:** 77-84.

KOLSTO A.B. 1997. Dynamic bacterial genome organization. Mol Microbiol **24:** 241-248.

KROGH A., BROWN M., MIAN I.S., SJOLANDER K. and HAUSSLER D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol **235:** 1501-1531.

LAMB A.L., TORRES A.S., O'HALLORAN T.V. and ROSENZWEIG A.C. 2000. Heterodimer formation between superoxide dismutase and its copper chaperone. Biochemistry **39:** 14720-14727.

LANDER E.S., LINTON L.M., BIRREN B., NUSBAUM C., ZODY M.C., BALDWIN J., DEVON K., DEWAR K., DOYLE M., FITZHUGH W., *et al.* 2001. Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LANDGRAF R., FISCHER D. and EISENBERG D. 1999. Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng **12:** 943-951.

LANDGRAF R., XENARIOS I. and EISENBERG D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol **307:** 1487-1502.

LARSEN T.A., OLSON A.J. and GOODSELL D.S. 1998. Morphology of protein-protein interfaces. Structure **6:** 421-427.

LASKOWSKI R.A., LUSCOMBE N.M., SWINDELLS M.B. and THORNTON J.M. 1996. Protein clefts in molecular recognition and function. Protein Sci **5:** 2438-2452.

LAURIE A.T. and JACKSON R.M. 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics.

LEE J.Y. and SPICER A.P. 2000. Hyaluronan: a multifunctional, megaDalton, stealth molecule. Curr Opin Cell Biol **12:** 581-586.

LEONARD C.J., ARAVIND L. and KOONIN E.V. 1998. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res **8:** 1038-1047.

LETUNIC I., COPLEY R.R., SCHMIDT S., CICCARELLI F.D., DOERKS T., SCHULTZ J., PONTING C.P. and BORK P. 2004. SMART 4.0: towards genomic data integration. Nucleic Acids Res **32:** D142-144.

LETUNIC I., GOODSTADT L., DICKENS N.J., DOERKS T., SCHULTZ J., MOTT R., CICCARELLI F., COPLEY R.R., PONTING C.P. and BORK P. 2002. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res **30:** 242-244.

LICHTARGE O., BOURNE H.R. and COHEN F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol **257:** 342-358.

LICHTARGE O. and SOWA M.E. 2002. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol **12:** 21-27.

LIJNZAAD P. and ARGOS P. 1997. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins **28:** 333-343.

LIM K.L., KOLATKAR P.R., NG K.P., NG C.H. and PALLEN C.J. 1998. Interconversion of the kinetic identities of the tandem catalytic domains of receptor-like protein-tyrosine phosphatase PTPalpha by two point mutations is synergistic and substrate-dependent. J Biol Chem **273:** 28986-28993.

LITTLER S.J. and HUBBARD S.J. 2005. Conservation of orientation and sequence in protein domain--domain interactions. J Mol Biol **345:** 1265-1279.

LIVINGSTONE C.D. and BARTON G.J. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci **9:** 745-756.

LO CONTE L., CHOTHIA C. and JANIN J. 1999. The atomic structure of protein-protein recognition sites. J Mol Biol **285:** 2177-2198.

LUPAS A.N., PONTING C.P. and RUSSELL R.B. 2001. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol **134:** 191-203.

LUSCOMBE N.M. and THORNTON J.M. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol **320:** 991-1009.

LYNCH M. and CONERY J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155.

MADABUSHI S., YAO H., MARSH M., KRISTENSEN D.M., PHILIPPI A., SOWA M.E. and LICHTARGE O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol **316:** 139-154.

MADEJ T., GIBRAT J.F. and BRYANT S.H. 1995. Threading a database of protein cores. Proteins **23:** 356-369.

MAEHAMA T. and DIXON J.E. 1998. The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. J Biol Chem **273:** 13375-13378.

MAHONEN A.P., BONKE M., KAUPPINEN L., RIIKONEN M., BENFEY P.N. and HELARIUTTA Y. 2000. A novel two-component hybrid molecule regulates vascular morphogenesis of the Arabidopsis root. Genes Dev **14:** 2938-2943.

MAJETI R., BILWES A.M., NOEL J.P., HUNTER T. and WEISS A. 1998. Dimerization-induced inhibition of receptor protein tyrosine phosphatase function through an inhibitory wedge. Science **279:** 88-91.

MARCHLER-BAUER A., ANDERSON J.B., CHERUKURI P.F., DEWEESE-SCOTT C., GEER L.Y., GWADZ M., HE S., HURWITZ D.I., JACKSON J.D., KE Z.*, et al.* 2005. CDD: a

Conserved Domain Database for protein classification. Nucleic Acids Res **33:** D192-196.

MARCHLER-BAUER A., PANCHENKO A.R., ARIEL N. and BRYANT S.H. 2002. Comparison of sequence and structure alignments for protein domains. Proteins **48:** 439-446.

MARCOTTE E.M. 2000. Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol **10:** 359-365.

MARCOTTE E.M., PELLEGRINI M., NG H.L., RICE D.W., YEATES T.O. and EISENBERG D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285:** 751-753.

MI H., LAZAREVA-ULITSKY B., LOO R., KEJARIWAL A., VANDERGRIFF J., RABKIN S., GUO N., MURUGANUJAN A., DOREMIEUX O., CAMPBELL M.J., *et al.* 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res **33:** D284-288.

MIGHELL A.J., SMITH N.R., ROBINSON P.A. and MARKHAM A.F. 2000. Vertebrate pseudogenes. FEBS Lett **468:** 109-114.

MIRNY L.A. and GELFAND M.S. 2002. Using orthologous and paralogous proteins to identify specificity determining residues. Genome Biol **3:** PREPRINT0002.

MOUGEL C. and ZHULIN I.B. 2001. CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants. Trends Biochem Sci **26:** 582-584.

MULDER N.J., APWEILER R., ATTWOOD T.K., BAIROCH A., BATEMAN A., BINNS D., BRADLEY P., BORK P., BUCHER P., CERUTTI L., *et al.* 2005. InterPro, progress and status in 2005. Nucleic Acids Res **33:** D201-205.

MULLER T., SPANG R. and VINGRON M. 2002. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol **19:** 8-13.

MURZIN A.G., BRENNER S.E., HUBBARD T. and CHOTHIA C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247:** 536-540.

MUSHEGIAN A.R. and KOONIN E.V. 1996. Gene order is not conserved in bacterial evolution. Trends Genet **12:** 289-290.

NEI M. and GOJOBORI T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3:** 418-426.

NEUWALD A.F. and LANDSMAN D. 1997. GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. Trends Biochem Sci **22:** 154-155.

NIELSEN R. and YANG Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929-936.

NOBES C.D., LAURITZEN I., MATTEI M.G., PARIS S., HALL A. and CHARDIN P. 1998. A new member of the Rho family, Rnd1, promotes disassembly of actin filament structures and loss of cell adhesion. J Cell Biol **141:** 187-197.

OHNDORF U.M., ROULD M.A., HE Q., PABO C.O. and LIPPARD S.J. 1999. Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. Nature **399:** 708-712.

ORENGO C.A., JONES D.T. and THORNTON J.M. 1994. Protein superfamilies and domain superfolds. Nature **372:** 631-634.

ORENGO C.A., MICHIE A.D., JONES S., JONES D.T., SWINDELLS M.B. and THORNTON J.M. 1997. CATH--a hierarchic classification of protein domain structures. Structure **5:** 1093-1108.

PABO C.O. and SAUER R.T. 1992. Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem **61:** 1053-1095.

PANCHENKO A.R., KONDRASHOV F. and BRYANT S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci **13:** 884-892.

PANNIFER A.D., FLINT A.J., TONKS N.K. and BARFORD D. 1998. Visualization of the cysteinyl-phosphate intermediate of a protein-tyrosine phosphatase by x-ray crystallography. J Biol Chem **273:** 10454-10462.

PARK J., KARPLUS K., BARRETT C., HUGHEY R., HAUSSLER D., HUBBARD T. and CHOTHIA C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol **284:** 1201-1210.

PARKINSON J.S. and KOFOID E.C. 1992. Communication modules in bacterial signaling proteins. Annu Rev Genet **26:** 71-112.

PAS J., VON GROTTHUSS M., WYRWICZ L.S., RYCHLEWSKI L. and BARCISZEWSKI J. 2004. Structure prediction, evolution and ligand interaction of CHASE domain. FEBS Lett **576:** 287-290.

PASCAL J.M., DAY P.J., MONZINGO A.F., ERNST S.R., ROBERTUS J.D., IGLESIAS R., PEREZ Y., FERRERAS J.M., CITORES L. and GIRBES T. 2001. 2.8-A crystal structure of a nontoxic type-II ribosome-inactivating protein, ebulin l. Proteins **43:** 319-326.

PATTHY L. 1987. Intron-dependent evolution: preferred types of exons and introns. FEBS Lett **214:** 1-7.

PEARL F., TODD A., SILLITOE I., DIBLEY M., REDFERN O., LEWIS T., BENNETT C., MARSDEN R., GRANT A., LEE D., *et al.* 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res **33:** D247-251.

PEARSON W.R. and LIPMAN D.J. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **85:** 2444-2448.

PELLEGRINI M., MARCOTTE E.M., THOMPSON M.J., EISENBERG D. and YEATES T.O. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96:** 4285-4288.

PENEFF C., MENGIN-LECREULX D. and BOURNE Y. 2001. The crystal structures of Apo and complexed Saccharomyces cerevisiae GNA1 shed light on the catalytic mechanism of an amino-sugar N-acetyltransferase. J Biol Chem **276:** 16328-16334.

PFUHL M. and PASTORE A. 1995. Tertiary structure of an immunoglobulin-like domain from the giant muscle protein titin: a new member of the I set. Structure **3:** 391-401.

PILS B. and SCHULTZ J. 2004. Evolution of the multifunctional protein tyrosine phosphatase family. Mol Biol Evol **21:** 625-631.

PONTING C.P. 1997. Evidence for PDZ domains in bacteria, yeast, and plants. Protein Sci **6:** 464-468.

PONTING C.P. 2001. Issues in predicting protein function from sequence. Brief Bioinform **2:** 19-29.

PONTING C.P., ARAVIND L., SCHULTZ J., BORK P. and KOONIN E.V. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. J Mol Biol **289:** 729-745.

PONTING C.P. and RUSSELL R.B. 2000. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. J Mol Biol **302:** 1041-1047.

PONTING C.P., SCHULTZ J., COPLEY R.R., ANDRADE M.A. and BORK P. 2000. Evolution of domain families. Adv Protein Chem **54:** 185-244.

PUIUS Y.A., ZHAO Y., SULLIVAN M., LAWRENCE D.S., ALMO S.C. and ZHANG Z.Y. 1997. Identification of a second aryl phosphate-binding site in protein-tyrosine

phosphatase 1B: a paradigm for inhibitor design. Proc Natl Acad Sci U S A **94:** 13420-13425.

PUPKO T., BELL R.E., MAYROSE I., GLASER F. and BEN-TAL N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics **18 Suppl 1:** S71-77.

QASBA P.K. and SAFAYA S.K. 1984. Similarity of the nucleotide sequences of rat alpha-lactalbumin and chicken lysozyme genes. Nature **308:** 377-380.

R DEVELOPMENT CORE TEAM 2004. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

REDFERN O., GRANT A., MAIBAUM M. and ORENGO C. 2005. Survey of current protein family databases and their application in comparative, structural and functional genomics. J Chromatogr B Analyt Technol Biomed Life Sci **815:** 97-107.

REKHA N., MACHADO S.M., NARAYANAN C., KRUPA A. and SRINIVASAN N. 2005. Interaction interfaces of protein domains are not topologically equivalent across families within superfamilies: Implications for metabolic and signaling pathways. Proteins **58:** 339-353.

ROST B., LIU J., NAIR R., WRZESZCZYNSKI K.O. and OFRAN Y. 2003. Automatic prediction of protein function. Cell Mol Life Sci **60:** 2637-2650.

ROUQUIER S., TAVIAUX S., TRASK B.J., BRAND-ARPON V., VAN DEN ENGH G., DEMAILLE J. and GIORGI D. 1998. Distribution of olfactory receptor genes in the human genome. Nat Genet **18:** 243-250.

RUTENBER E., READY M. and ROBERTUS J.D. 1987. Structure and evolution of ricin B chain. Nature **326:** 624-626.

SCHAFFER A.A., WOLF Y.I., PONTING C.P., KOONIN E.V., ARAVIND L. and ALTSCHUL S.F. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics **15:** 1000-1011.

SCHULTZ J., MILPETZ F., BORK P. and PONTING C.P. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A **95:** 5857-5864.

SHAPIRO L. and SCHERER P.E. 1998. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. Curr Biol **8:** 335-338.

SHAW P., FREEMAN J., BOVEY R. and IGGO R. 1996. Regulation of specific DNA binding by p53: evidence for a role for O-glycosylation and charged residues at the carboxy-terminus. Oncogene **12:** 921-930.

SHEEN J. 2002. Phosphorelay and transcription control in cytokinin signal transduction. Science **296:** 1650-1652.

SIDDIQUI A.S. and BARTON G.J. 1995. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci **4:** 872-884.

SIDOW A. 1996. Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev **6:** 715-722.

SIMINOVITCH K.A., LAMHONWAH A.M., SOMANI A.K., CARDIFF R. and MILLS G.B. 1999. Involvement of the SHP-1 tyrosine phosphatase in regulating B lymphocyte antigen receptor signaling, proliferation and transformation. Curr Top Microbiol Immunol **246:** 291-297; discussion 298.

SIMON A.L., STONE E.A. and SIDOW A. 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc Natl Acad Sci U S A **99:** 2912-2917.

SJOLANDER K. 1998. Phylogenetic inference in protein superfamilies: analysis of SH2 domains. Proc Int Conf Intell Syst Mol Biol **6:** 165-174.

SJOLANDER K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics **20:** 170-179.

SMITH T.F. and ZHANG X. 1997. The challenges of genome sequence annotation or "the devil is in the details". Nat Biotechnol **15:** 1222-1223.

SONNHAMMER E.L. and KOONIN E.V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet **18:** 619-620.

SOWDHAMINI R. and BLUNDELL T.L. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. Protein Sci **4:** 506-520.

SPICHAL L., RAKOVA N.Y., RIEFLER M., MIZUNO T., ROMANOV G.A., STRNAD M. and SCHMULLING T. 2004. Two cytokinin receptors of Arabidopsis thaliana, CRE1/AHK4 and AHK3, differ in their ligand specificity in a bacterial assay. Plant Cell Physiol **45:** 1299-1305.

STARK A., SHKUMATOV A. and RUSSELL R.B. 2004. Finding functional sites in structural genomics proteins. Structure (Camb) **12:** 1405-1412.

STREULI M., KRUEGER N.X., THAI T., TANG M. and SAITO H. 1990. Distinct functional roles of the two intracellular phosphatase like domains of the receptor-linked protein tyrosine phosphatases LCA and LAR. Embo J **9:** 2399-2407.

STRIMMER K. and VON HAESELER A. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol **13:** 964-969.

STUCKEY J.A., SCHUBERT H.L., FAUMAN E.B., ZHANG Z.Y., DIXON J.E. and SAPER M.A. 1994. Crystal structure of Yersinia protein tyrosine phosphatase at 2.5 A and the complex with tungstate. Nature **370:** 571-575.

SUN Y.J., CHOU C.C., CHEN W.S., WU R.T., MENG M. and HSIAO C.D. 1999. The crystal structure of a multifunctional protein: phosphoglucose isomerase/autocrine motility factor/neuroleukin. Proc Natl Acad Sci U S A **96:** 5412-5417.

SUZUKI T., MIWA K., ISHIKAWA K., YAMADA H., AIBA H. and MIZUNO T. 2001. The Arabidopsis sensor His-kinase, AHk4, can respond to cytokinins. Plant Cell Physiol **42:** 107-113.

SUZUKI Y. and GOJOBORI T. 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol **16:** 1315-1328.

SWINDELLS M.B. 1995. A procedure for detecting structural domains in proteins. Protein Sci **4:** 103-112.

TATENO Y., SAITOU N., OKUBO K., SUGAWARA H. and GOJOBORI T. 2005. DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Res **33:** D25-28.

TATUSOV R.L., GALPERIN M.Y., NATALE D.A. and KOONIN E.V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28:** 33-36.

TEICHMANN S.A., PARK J. and CHOTHIA C. 1998. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. Proc Natl Acad Sci U S A **95:** 14658-14663.

THOMAS P.D., CAMPBELL M.J., KEJARIWAL A., MI H., KARLAK B., DAVERMAN R., DIEMER K., MURUGANUJAN A. and NARECHANIA A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res **13:** 2129-2141.

THOMPSON J.D., GIBSON T.J., PLEWNIAK F., JEANMOUGIN F. and HIGGINS D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25:** 4876-4882.

THOMPSON J.D., HIGGINS D.G. and GIBSON T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

THORNTON J.M., TODD A.E., MILBURN D., BORKAKOTI N. and ORENGO C.A. 2000. From structure to function: approaches and limitations. Nat Struct Biol **7 Suppl:** 991-994.

TODD A.E., ORENGO C.A. and THORNTON J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol **307:** 1113-1143.

VALDAR W.S. and THORNTON J.M. 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins **42:** 108-124.

VALENCIA A. 2005. Automatic annotation of protein function. Curr Opin Struct Biol **15:** 267-274.

VALENCIA A. and PAZOS F. 2002. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol **12:** 368-373.

VAN HUIJSDUIJNEN R.H., BOMBRUN A. and SWINNEN D. 2002. Selecting protein tyrosine phosphatases as drug targets. Drug Discov Today **7:** 1013-1019.

VENTER J.C., ADAMS M.D., MYERS E.W., LI P.W., MURAL R.J., SUTTON G.G., SMITH H.O., YANDELL M., EVANS C.A., HOLT R.A., *et al.* 2001. The sequence of the human genome. Science **291:** 1304-1351.

VILLAFRANCA J.E. and ROBERTUS J.D. 1981. Ricin B chain is a product of gene duplication. J Biol Chem **256:** 554-556.

VINGRON M. and WATERMAN M.S. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol **235:** 1-12.

VON MERING C., JENSEN L.J., SNEL B., HOOPER S.D., KRUPP M., FOGLIERINI M., JOUFFRE N., HUYNEN M.A. and BORK P. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res **33:** D433-437.

VOSSELLER K., WELLS L., LANE M.D. and HART G.W. 2002. Elevated nucleocytoplasmic glycosylation by O-GlcNAc results in insulin resistance associated with defects in Akt activation in 3T3-L1 adipocytes. Proc Natl Acad Sci U S A **99:** 5313-5318.

WANG N., SODERBOM F., ANJARD C., SHAULSKY G. and LOOMIS W.F. 1999. SDF-2 induction of terminal differentiation in Dictyostelium discoideum is mediated by the membrane-spanning sensor kinase DhkA. Mol Cell Biol **19:** 4750-4756.

WANG Y., ANDERSON J.B., CHEN J., GEER L.Y., HE S., HURWITZ D.I., LIEBERT C.A., MADEJ T., MARCHLER G.H., MARCHLER-BAUER A., *et al.* 2002. MMDB: Entrez's 3D-structure database. Nucleic Acids Res **30:** 249-252.

WANG Y. and PALLEN C.J. 1991. The receptor-like protein tyrosine phosphatase HPTP alpha has two active catalytic domains with distinct substrate specificities. Embo J **10:** 3231-3237.

WATERSTON R.H., LINDBLAD-TOH K., BIRNEY E., ROGERS J., ABRIL J.F., AGARWAL P., AGARWALA R., AINSCOUGH R., ALEXANDERSSON M., AN P., *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

WELLS L., GAO Y., MAHONEY J.A., VOSSELLER K., CHEN C., ROSEN A. and HART G.W. 2002. Dynamic O-glycosylation of nuclear and cytosolic proteins: further characterization of the nucleocytoplasmic beta-N-acetylglucosaminidase, O-GlcNAcase. J Biol Chem **277:** 1755-1761.

WELLS L., VOSSELLER K. and HART G.W. 2001. Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc. Science **291:** 2376-2378.

WERNER M.H., HUTH J.R., GRONENBORN A.M. and CLORE G.M. 1995. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. Cell **81:** 705-714.

WHISSTOCK J.C. and LESK A.M. 2003. Prediction of protein function from protein sequence and structure. Q Rev Biophys **36:** 307-340.

WILMANNS M., HYDE C.C., DAVIES D.R., KIRSCHNER K. and JANSONIUS J.N. 1991. Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. Biochemistry **30:** 9161-9169.

WISHART M.J., DENU J.M., WILLIAMS J.A. and DIXON J.E. 1995. A single mutation converts a novel phosphotyrosine binding domain into a dual-specificity phosphatase. J Biol Chem **270:** 26782-26785.

WOLF E., VASSILEV A., MAKINO Y., SALI A., NAKATANI Y. and BURLEY S.K. 1998. Crystal structure of a GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-acetyltransferase. Cell **94:** 439-449.

WOLFE K.H. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet **2:** 333-341.

WRABL J.O. and GRISHIN N.V. 2001. Homology between O-linked GlcNAc transferases and proteins of the glycogen phosphorylase superfamily. J Mol Biol **314:** 365-374.

WRIGHT C.S., ALDEN R.A. and KRAUT J. 1969. Structure of subtilisin BPN' at 2.5 angstrom resolution. Nature **221:** 235-242.

WU C., SUN M., LIU L. and ZHOU G.W. 2003. The function of the protein tyrosine phosphatase SHP-1 in cancer. Gene **306:** 1-12.

WU C.H., NIKOLSKAYA A., HUANG H., YEH L.S., NATALE D.A., VINAYAKA C.R., HU Z.Z., MAZUMDER R., KUMAR S., KOURTESIS P.*, et al.* 2004. PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res **32:** D112-114.

WU G., FISER A., TER KUILE B., SALI A. and MULLER M. 1999. Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci U S A **96:** 6285-6290.

YAKUNIN A.F., YEE A.A., SAVCHENKO A., EDWARDS A.M. and ARROWSMITH C.H. 2004. Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol **8:** 42-48.

YAMADA H., SUZUKI T., TERADA K., TAKEI K., ISHIKAWA K., MIWA K., YAMASHINO T. and MIZUNO T. 2001. The Arabidopsis AHK4 histidine kinase is a cytokinin-binding receptor that transduces cytokinin signals across the membrane. Plant Cell Physiol **42:** 1017-1023.

YANG X.L., SCHIMMEL P. and EWALT K.L. 2004. Relationship of two human tRNA synthetases used in cell signaling. Trends Biochem Sci **29:** 250-256.

YANG Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39:** 306-314.

YANG Z. 2001. Adaptive molecular evolution, pp. 327-350 in *Handbook of statistical genetics*, edited by BALDING D., BISHOP M. and CANNINGS C. Wiley, London.

YANG Z. and BIELAWSKI J.P. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution **15:** 496-503.

YAP K.L., AMES J.B., SWINDELLS M.B. and IKURA M. 1999. Diversity of conformational states and changes within the EF-hand protein superfamily. Proteins **37:** 499-507.

YUVANIYAMA J., DENU J.M., DIXON J.E. and SAPER M.A. 1996. Crystal structure of the dual specificity protein phosphatase VHR. Science **272:** 1328-1331.

ZHANG J., ROSENBERG H.F. and NEI M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A **95:** 3708-3713.

ZHANG X. and FIRESTEIN S. 2002. The olfactory receptor gene superfamily of the mouse. Nat Neurosci **5:** 124-133.

Contributions

The work presented in the individual chapters of this thesis has been partially produced in collaborations. Contributions by other people are indicated here:

Chapter 2:

Prediction of structure and functional residues for O-GlcNAcase, a divergent homologue of acetyltransferases
Schultz J, Pils B
FEBS Lett. 529:179-82.

The analysis of the O-GlcNAcase domain was conducted by Birgit Pils, the analysis of the hyaluronidase domain by Jörg Schultz, who also guided this project. The manuscript was written by both authors.

Chapter 3:

Prediction of Cytokinin-binding sites in the CHASE domain
Birgit Pils carried out the computational analysis. The experiments were performed by Nicola Nielsen[*]. The project was conducted by Alexander Heyl[*].

Chapter 4:

Evolution of the multifunctional protein tyrosine phosphatase family
Pils B, Schultz J
Mol Biol Evol. 21:625-31.

Birgit Pils carried out all steps of the analysis and prepared a draft of the manuscript. Jörg Schultz supervised the project.

Chapter 5:

Inactive enzyme-homologues find new function in regulatory processes
Pils B, Schultz J
J Mol Biol. 340:399-404.

Birgit Pils collected the information on catalytic sites and performed the large-scale analysis. She also drafted the manuscript. Jörg Schultz supervised the project.

Chapter 6:

Variation in structural location and amino acid conservation of functional sites in protein domain families
Pils B, Copley RR[§], Schultz J
BMC Bioinformatics. In revision

The data set was generated by Birgit Pils together with Richard Copley. Birgit Pils carried out the analysis and prepared a draft of the manuscript. The statistical analysis was assisted by Wolfgang Huber[#]. Jörg Schultz supervised the project.

---

[*] Freie Universität Berlin, Institut für Biologie, Angewandte Genetik, Albrecht-Thaer-Weg 6, 14195 Berlin
[§] Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, OX3 7BN Oxford, UK
[#] EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Curriculum Vitae

Birgit Pils

Education

Nov 2003 – Apr 2005	PhD student in the group of Dr. Jörg Schultz at the Department of Bioinformatics, University of Würzburg
May 2002 - Oct 2003	PhD student in the group of Dr. Jörg Schultz at the Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin
Sep 2000 – Feb 2002	Diploma student and subsequently post-graduate researcher at the Institute for Genetics at Stanford University, USA, supervised by Dr. Uta Francke and Dr. Hans-Joachim Lipps Title of diploma thesis: „Genomic Organization of PWCR1 small nucleolar RNAs and Detection of a Host Gene in the Imprinted Prader-Willi Syndrome Region“
Oct 1998 – Sep 2001	Studies of Biochemistry at the University of Witten/Herdecke Degree: Diploma in Biochemistry (Sep 2001)
Oct 1995 – Sep 1998	Studies of Biochemistry at the University of Bochum
June 1995	Abitur (A-levels equivalent), Main-Taunus Schule, Hofheim am Taunus.
Aug 1991 – June 1992	Exchange student at Lordsburgh High School, Lordsburgh, New Mexico, USA
Aug 1984 – Jun 1991	Elisabethenschule, Hofheim am Taunus.

Academic Activities

July/August 2004	Research fellow of the Japan Society for the Promotion of Science (JSPS) visiting the group of Dr. Hiroyuki Toh, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
July – Sep 1999	Internship at the Institute of Cell and Molecular Biology, University of Edinburgh, UK, Dr. Millicent Masters
July – Sep 1997	Internship at the Institute for Biochemistry, Escola Paulista de Medicina, Sao Paulo, Brazil, Dr. Claudio Sampaio
1997 – 2000	Treasurer for IAESTE (International Association for the exchange of students for technical experience) at the local committee of Bochum (1999) and social event manager for foreign exchange students

List of Publications

Publications associated with this thesis:

Letunic I, Copley RR, Doerks T, Ciccarelli F, **Pils B**, Pinkert S, Schultz J, Bork P. 2005. SMART 5.0: Improvements of Functional Annotation and Insights into Domain Evolution. In preparation.

Friedrich T, **Pils B**, Dandekar T, Schultz J, Müller T. 2005. Interaction Profile Hidden Markov Model – A Method for Interaction Sites Prediction. In preparation.

Pils B, Copley RR, Schultz J. 2005. Variation in structural location and amino acid conservation of functional sites in protein domain families. BMC Bioinformatics. Accepted.

Pils B, Schultz J. 2004. Inactive enzyme-homologues find new function in regulatory processes. J Mol Biol. 340:399-404.

Pils B, Schultz J. 2004. Evolution of the multi-functional protein tyrosine phosphatase family. Mol Biol Evol. 21:625-31.

Schultz J, **Pils B**. 2002. Prediction of structure and functional residues for O-GlcNAcase, a divergent homologue of acetyltransferases. FEBS Lett. 529:179-82.

Prior Publications:

Gallagher RC, **Pils B,** Albalwi M, Francke U. 2002. Evidence for the role of PWCR1/HBII-85 C/D box small nucleolar RNAs in Prader-Willi syndrome. Am J Hum Genet. 71:669-78.

Conference Contributions

The Nature of Functional Sites.
Pils B, Copley RR, Schultz J.
Human Genome Meeting, Kyoto, Japan, April 18-21, 2005.
(Oral Presentation)

Large Scale Analysis of Substitutions at Enzyme's Catalytic Sites in Metazoan Genomes.
Pils B, Schultz J.
Human Genome Meeting, Berlin, April 4-7, 2004.
(Poster Presentation)
Award: 2nd Poster Prize from the Nature Publishing Group

Functional divergence in the protein tyrosine phosphatase family.
Pils B, Schultz J.
German Conference on Bioinformatics, Munich, October 12-14, 2003.
(Poster Presentation)