Thesis

The human proteome is shaped by evolution and interactions

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität
Würzburg

vorgelegt von

Stefan Pinkert

aus Dresden

Würzburg 2008

Eingereicht am: .....

Mitglieder der Promotionskommission:

Vorsitzender: .....

Gutachter : Prof. Dr. Jörg Schultz

Gutachter: Prof. Dr. Georg Krohne

Tag des Promotionskolloquiums: .....

Doktorurkunde ausgehändigt am: .....

I want to thank

Prof. Dr. J. Schultz for the interesting discussions the wonderful work environment und his neverending patience.
Prof. Dr. Krohne for the time to read and evaluate my thesis.

The bioinformatics group for scientific and social support.

Dr. Birgit Pils for proof-reading this thesis.

My family which always believed in me.
Special thanks to my girlfriend Jessica and her family for endless motivation and support.

Dedicated to the beloved ones who are gone.

Contents

List of Figures

List of Tables

Chapter 1

General introduction

1.1 Evolution

The question "Where do we come from" has puzzled mankind almost since its existence. One of the first scientists addressing this question and achieving a major break-through was Charles Darwin with his theory *On the Origin of Species* in the mid 19th century. With this work,he created the base for the theory of evolution. Darwin deciphered the underlying rules of development of life on earth. He stated that individuals most suited to the environment are more likely to survive and to reproduce. So their inheritable traits stay in future generations. After interminable time variations accumulate and finally lead to new species. Species are defined as groups of individuals who are capable of interbreeding and whose offspring is also fertile. For example, a donkey and a horse can interbreed but their offspring, the mule is not fertile. Another way to distinguish species is based on molecular markers. The relationship between bacteria, archaea and eukaryotes based on 16S-rDNA from mitochondria was proposed by (Woese and Fox 1977). Ideally have the used DNA or RNA sequences differences in all species of the group of interest but are not to different to be compared. A detailed review by Halanych 2004 about the different molecular marker like internal transcribed spacer (ITS) and ITS2 and the resulting tree is used in this work. The evolution of new better suited species means that other less adapted species living in the same environment become extinct. The path of evolution itself is endless but many paths of species have ended in the interim. So if we are looking at all species living today, we only see a small part of the puzzle. Some clades, groups of taxa, which share a common feature, are gone and others have succeeded, at least for some time.
It is important to keep in mind that evolution itself is a passive process. But

nevertheless as every biologist knows *Nothing in Biology Makes Sense Except in the Light of Evolution.* These famous words from Dobzansky 1964 define still the importance of evolution in biology. Every life on earth we see today, including viruses, as prime example is the result of evolution. This can be reduced to one species after another but this definition misses the fact that evolution rhymes (Dawkins 2004). But how can we catch this all forming but not directly observable force? Let us begin at the source of all changes in organisms.
Alterations in the DNA sometimes lead to new alleles. In some rare cases those are an advantage for the individual and it can reproduce more often. If, in the even more unlikely case, the advantageous allele is in or enters the germline, the offspring can reproduce more often too. Some of those changes are on the cellular level and others influence the morphology of the whole organisms. We can directly compare species based on their appearance. This was done in the past by looking for morphological features and group species according to them. Linné developed a hierarchical system and conventions for naming species. His ranked taxonomy is divided into (Domain)Kingdom, Phylum,Class,Order,Genus,Species. The ultimate goal is a taxonomy containing all species in one big tree of life.
Given we have different species. How to group them? One possibility is comparing molecular data. If all those species share one protein we can compare the amino acid sequences. We group the two sequences together with the smallest number of different amino acid at all positions and add than one by one the other sequences having the smallest differences to the resulting sequence. But if all sequences have different amino acids on the same position which are more closely related? One could create a table with all amino acids, as column the original amino acids and the changed ones as rows. In general, we can think of three parameters. Chances based on the DNA sequence triplett based code translation, the difference in chemical features of the amino acids or calculated probabilities based on real sequences. On solution is the (P)oint (A)ccepted (M)utations matrix (Dayhoff et al. 1978). The original set developed by Dayhoff contained 71 families of closely related proteins with 1572 observed mutations. Today, PAM1 to PAM250 matrices are in use. The number stands for the number of point mutations per 100 amino acids. All PAM matrices are calculated by multiplying PAM1 with itself. Another substitution matrix called BLOSUM,BLOcks of amino acid SUbstitution Matrix (Henikoff and Henikoff 1992) was developed in 1992. It is based on the BLOCKS Database. The advantage of BLOSUM is that all matrices are calculated on real alignments and not extrapolated like PAM matrices greater than PAM1. The BLOSUM number is equivalent to the sequence identity of clustered sequences for the alignment. This means comparing closely related sequences is best done with

low PAM matrices or high BLOSUM. But how reliable is our tree? Perhaps two sequences are nearly identical and could change the tree if one amino acid is mutated. A technique called bootstrapping will help us. Columns of the multiple sequence alignment of our sequences are randomly sampled. The number of columns stays the same but some randomly chosen columns are deleted and others are used multiple times. The next step is to calculate trees for each of those new alignments. All resulting trees are then combined to a consensus tree (compare figure 1.2).The percentage at the inner nodes reflects the frequency of the occurrence of the node combinations in the sampled trees. High values mean robust pairs of nodes. Finally, some definitions are needed to explaine phylogenetic trees and gene/protein relations as shown in figure 1.2.

Horizontal Gene Transfer is a common genetic mechanism in *Bacteria*. They exchange genetic material by conjugation or trough *Bacteriophages*. The frequency of this mechanism in multicellular organism is under discussion e.g. by de la Cruz and Davies 2000 and Stanhope et al. 2001. **Gene Loss** describes the fact that a gene is not part of a genome any longer. **Convergent Evolution** denotes the phenomenon that in the evolution of two distant related species functionally similar but distinct related features like antifreeze proteins or wings originate.

Homologous genes share a common ancestor. The subtypes of homology are orthology, paralogy, and xenology. A more general definition of homology was made by Fitch 2000, *Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character.* Orthology and paralogy differ in that one proceeds from speciation and the other from gene duplication, but either evolutionary course of divergence has the same potential for acquisition of new properties (Jensen 2001). Homologous genes in one species with an history involving an lateral(compare chapter 1.1) gene transfer are **xenologous**. How are homologues computationally identified?

The Blast (McGinnis and Madden 2004; Altschul et al. 1990) tool can search even large sequence databases in a short time to identify homologues of the query sequence. The underlying algorithm is based on local alignments. The gap penalties as well as costs and the substitution matrix like BLOSUM62 are adjustable. The BLOSUM matrix number is used to better adjust the changes in the amino acid between the related sequences, closely related sequences are best found with an high numbered matrix like BLOSUM80, gap penalties and cost should also be set higher than for more distantly releted sequences. For each hit, found sequence, are two values calculated, the S-score and the E-value. The S-score reflects the similarity of the query to the sequence shown. And the E-value gives the probability that due to chance, another alignment in the database has an greater S-score. The E-value is dependent on the size of

the database.

1.2 Protein

A chain of amino acids connected by peptide bonds is the base of all proteins. Up to twenty different amino acids are mostly used. Additionally two uncommon amino acids are sometimes included. The biggest proteins are consisting of up to 27,000 amino acids. The peptide bond of amino acids is different based on the two involved amino acids. Their side chains define the angles Θ(C_αN bond) and Ψ(C_αCO bond).These different angles lead to the forming of local of secondary structures like α-*helices*(Θ between –40° and –100°, Ψ between –40° and –65°) or β-*sheets*(Ψ between –80° and –120°, Ψ between 120° and 170°). Amino acids can be classified in different ways, for example based on their different sidechains, which can be aliphatic and even aromatic. Other features are for example the in size or pH-value. On important feature is their hydrophilic character, which is determined by the polarity of the side chain. Proteins as part of cells are mostly in an aqueous environment. Therefore the protein is folded so that hydrophobic parts of the amino acid chain are in the core and hydrophilic parts on the surface. In membrane bound proteins are hydrophobic parts bound to the lipid bilayer. The spatial organisation of the secondary structures of a protein is called the tertiary structure. Finally, the interaction of multiple polymer chains is the quaternary structure of proteins.
The classification of proteins is important to find similarities between allready known proteins and to transfer their features to new proteins exhibiting the same similarities. How are proteins classified?
In general, we can differentiate between structural versus functional and curated versus automated (Ouzounis et al. 2003) classifications. Structural classifications compare similarity on the level of primary or tertiary structures. Classifications of the amino acid sequence are mainly focused on domain, motifs and protein families. Protein families share parts of their sequences which are more conserved through evolution. The fixed and variable elements of these regions can often be described with motifs. PROSITE (Hulo et al. 2008) as one of the oldest database for motifs is well supported by the community and contains many hand curated motifs. Multiple motifs are combined to protein fingerprints in the PRINTS(23,24) database. The PFAM, Protein FAMilies database (Bateman et al. 2004) and the SMART, Simple Modular Architecture Research Tool (Schultz et al. 1998) are focused on protein domains. They both use hidden markov models to detect domains. Smart includes mainly small signalling,

nucleous or extracellular domains which are hard to detect but widespread in proteins. Pfam contains a broad spectrum of over ten thousand domains most of them part of Pfam A, which is the manually curated set complemented by the on automatic alignments based set Pfam B. Another important databases is the COG, Clusters of Orthologous Groups (Tatusov et al. 2003). It is based on sequences of orthologous protein encoding genes. Each COG consists of proteins or groups of paralogues found in at least three different genomes.
The three most known tertiary structure classification databases are SCOP, CATH and FSSP. One disadvantage shared by all tools based on tertiary structures is their reliance on solved protein structures. Those are stored in PDB, the Protein Resource Database (Berman et al. 2000).
SCOP, the Structural Classification Of Proteins database (Murzin et al. 1995)(compare table 1.2) exists since 1995. Scop uses visual inspection and manual curation to classify in three levels. *Families* are based on a clear evolutionary relationship and in most cases proteins within families have more than 30 percent pairwise residue identity. The proteins in a *superfamily* probably have a common evolutionary origin suggested by their structural and functional features, but their sequence identities are low. *Folds* share the same arrangement of major secondary structures which are connected in the same topological way. Cath clusters proteins in four levels Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). The classification is a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis (Orengo et al. 1997; Cuff et al. 2008). FSSP is known under two names Fold classification based on Structure-Structure alignment of Proteins and Families of Structurally Similar Protein. The alignments are computed by the automatic Distance matrix ALIgnment (Dali) server. New PDB entries are automaticaly scanned and the individual chains divedid into two sets, a representative set and a set of sequence homologs with more than 25% sequence identity for each representative chain(Holm and Sander 1994).
Functional classifications can be based on enzymatic reactions, functional roles or cellular localisation. One of the oldest schemes is based on hierarchical classification of the EC, Enzyme Commission. Six main types of reactions are further subdivided in three levels. Proteins classified with the same four level based number can catalyze the same reaction in different species and could therefore be homologous. One more general approach was developed by the (G)ene (O)ntology Consortium in 2000 (Ashburner et al. 2000a). Proteins are annotated in three independent classes and each of them is further subdivided in form of a directed acyclic graph(compare figure 1.4).

- Cellular compartment: extracellular, cytosolic, ribosome, etc.
- Biological Function: cellular physiological process, signal transduction, pyrimidine metabolic process, etc.
- Molecular function: catalytic activity, transporter activity, binding, etc.

The GO classification of proteins is achieved with manual and automated methods. In each annotation source and evidence are noted. Proteins can be annotated with zero or more GO terms. KEGG, the Kyoto Encyclopedia of Genes and Genomes (Ogata et al. 1999) contains amongst other information many biological pathways. The data is manually extracted from the literature and added to reference pathways. These pathways can be filtered for the subset of proteins found in different species. An important part in the unterstanding of proteins their function is the question how new proteins arise. How do proteins emerge?
One source are gene duplications. With the cell having two copies of the same gene the evolutionary pressure is reduced. One of the copies can accumulate mutations. This leads to pseudo-genes most cases. But sometimes the mutations result in a changed protein function beneficial for the organism. This is called neofunctionalisation. In single cell species the resulting new gene is automatically part of the population. In multicellular species the new gene has to be part of the germline to be passed on. An whole genome duplication would dramatically increases the chance for new genes and thus proteins. (Sidow 1996 proposed 2 whole genome duplication in the early evolution of vertabrates. This hypothesis of (2R) rounds of genome duplications is still under discussion (Hokamp et al. 2003). The large group of mammalien olfactory genes is a good example for many gene duplications which are followed by neofunctionalisation and pseudo genes. These genes very different rates of pseudogenes 20 percent in mouse in contrast to 60 percent in human (

A first approach to predict protein function started with transferring function from a close experimentally described homolog. Homologous sequences share(not necessarily) similar functions. Sequence similarity indicates functional similarity. For example enzymes could share the same mechanism and the same same substrate. With less sequence similarity they can still share the same mechanism but not the substrate. The major protein databases are listed in table 1.1. But one has to keep in mind that even in this well maintained databases are not error free. Artamonova et al. 2005 analysed the high quality standard and manually curated SwissProt/UniProt database for errors and found an rate of 33%–45%. Errors are typically caused by under annotation and trivial misannotations. This is a small problem looking at single sequences but a bigger

Table 1.1 Primary DNA/Protein Sequence Databases

Database	WWW-Address	Descriptions
EMBL	www.ebi.ac.uk/embl/	European Molecular Biology Laboratory nucleotide sequence database at EBI, Hinxton, UK
GenBank	/www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html	DNA Genome Sequence Database at National Center for Biotechnology information, NCBI, Bethesda, MD, USA
DDBJ	www.ddbj.nig.ac.jp/	DNA Data Bank Japan at CIB , Mishima, Japan
SWISS-PROT/TrEMBL	www.expasy.ch/	Protein Sequence Database (Swiss Institute of Bioinformatics, SIB, Geneva, CH)
PIR-PSD	pir.georgetown.edu/pirwww/search/-textpsd.shtml	PIR-International Protein Sequence Database, annotated protein database by PIR, MIPS and JIPID at NBRF, Georgetown University, USA
SRS	srs.ebi.ac.uk/	Sequence Retrieval System

The Primary DNA/Protein Sequence Databases are synchronized once a day.

one for large scale analyses.

In Bacteria exists groups of genes which are regulated by the same operon. These genes are in most cases part of the same pathway and they transcribed to one polycistronic mRNA (JACOB and MONOD 1961). The fusion of two orthologs of genes into one polypeptide in an organism indicates an functional interaction of the two corresponding proteins.

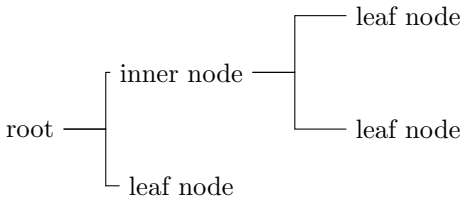


Figure 1.1 A rooted tree has a root node, inner nodes and leaf nodes. Leaf nodes are the e.g. sequences or species being compared.

Table 1.2 Statistics of Scop

Class	Number of		
	Folds	Superfamilies	Families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

The number of entries where taken on July 23, 2008.

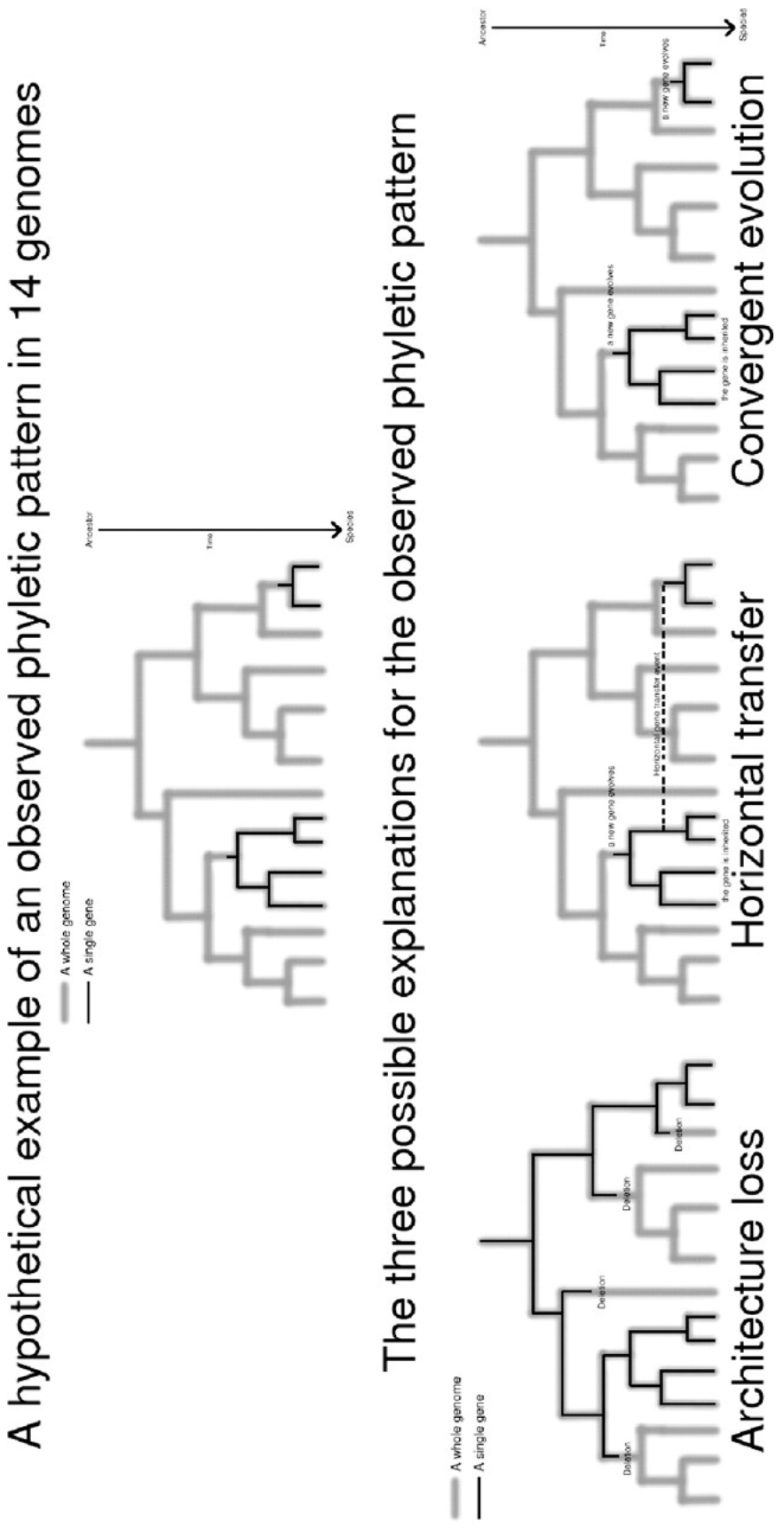


Figure 1.2 Three mechanisms can explain the occurrence of a gene in different clades, taken from Gough 2005.

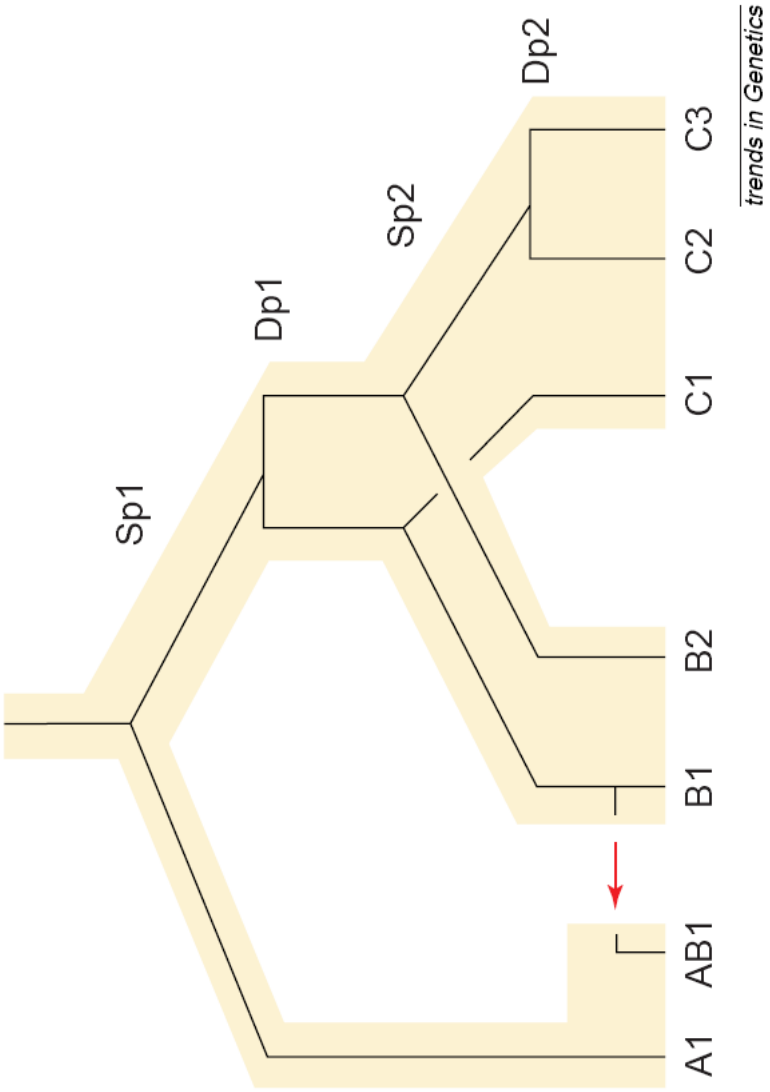


Figure 1.3 The concept of orthologs and paralogs and xenalogs, taken from Fitch 2000. The lines show one possible evolution of a gene from an common ancestor into three populations A, B, C. A speciation event (Sp1, Sp2) is shown by an inverse Y, genes which have their common ancestor at this point are orthologous. Gene duplications (Dp1, Dp2) occur on horizontal bars and indicate paralogous genes. The red arrow indicates xenology, the resulting gene AB1 is xenologous to all other genes. C2 and C3 are orthologous to B2 but paralogous to each other.

Figure 1.4 This picture was created by the Gene Ontologizer (Bauer et al. 2008). Colours represent the significance and the numbers the count of proteins annotated with the specific GO term.

1.3 Domain

The molecular function of proteins is often carried out by structural independent parts called domains. Other features of proteins are transmembrane regions, which are essential to bind membranes, unstructured loops connect for connecting domains and finally the signal peptide region at the N-terminus of the sequence. Proteins consist in most cases of one or more domains. Domains are conserved evolutionary parts of proteins that often correspond with functional units (Bornberg-Bauer et al. 2005). Domains can be defined in three ways.

- Structural, a domain is based on motifs folded compact and local Richardson 1981).
- Functional, domains are defined as the smallest part which is needed to carry out a function (Bork 1991).
- Evolutionary, domains are genetically mobile as described by Bork et al. 1997; Schultz et al. 2000.

Domains are typically 100 to 250 amino acids long. In general, their fold is stable sometimes stabilized by ions and in other cases by the aggregation to multimers e.g β-propeller. The combination of many small domains can lead to very big proteins. Most proteins consist of more than one domain especially eukaryotic proteins. It is estimated that 2/3 of prokaryotic and 80 percent of eukaryotic proteins are of multidomain character. The sum of the functions of each domain in a protein reflects its functions.
Ignoring the domain based architecture of proteins could even lead to falsely annotated proteins (compare figure 1.5). Although the rate at which new sequences that could probably contain new domains is accelerated by full genome sequencing projects, the number of newly found domains is decreasing(Copley et al. 2002). One possible explanation could be that the number of different folds is limited(Chothia 1992; Orengo et al. 1994). Most domains exist since the Metazoa or earlier in evolution (Pal and Guda 2006)
Domains can be more or less easily found in proteins sequences. The first version of the Simple Modular Architecture Tool (SMART) (Schultz et al. 1998) has gone online in 1998 with computermodels for 86 signalling domains. Ten years later SMART is now able to predict 752 different domains and in addition all domains stored in the Pfam database (Sonnhammer et al. 1997). This project started in 1996 and since the beginning it is divided into PfamA and PfamB. Part A is like Smart based on manually curated alignments, while Part B contains only automatically aligned sequences. The most sophisticated method to

detect domains are Hidden Markov Models (HMM). The underlying principle of HMMs is the following: First, homologous sequences of the domain of interest are gathered with tools like blast. Second, an multiple alignment of those sequences is created. Automated alignments are normally improved by manual curation. In the last step the HMM is trained with the multiple sequence alignment (compare figure 1.6). A classical HMM can store three different states for each position of the multiple sequence alignment, which are insert, delete and match state. The latter one contains probabilities for each amino acid. Each state has an probability of being followed by one of the three states. In an classical HMM the actual state depends only on the state before. The trained HMM is able to detect variants of the domain not found in a blast search, thus it is more sensitive. Newer HMMs incorporate additional information about protein ligand interaction sites (Friedrich et al. 2006).
How do domains arise in evolution?
One theory on the origin of domains is based on small polypeptides. Those short fragments combined to multimers are able to carry out a function. Then fusion of several short polypeptides into one sequence could lead to domains. The internal duplication of those small polypeptides could be another mechanisms for the genesis of new domains.
Special points in evolution like the origin of multi-cellularity gave rise to a great numbers of new domains in this case especially extracellular domains. Another point we have to keep in mind is that our computational prediction is not perfect. So less ancient domains could have undetected distant orthologues. This was shown by Ponting and Russell 2000 for the precursor of a cytokine thought to have arisen in chordata, which has actin binding homologues in Funghi, formerly unknown.
Like proteins, domains, are freed from evolutionary pressure after an gene/genome duplication. This could result in new domains or new suptypes of domains. The linear order of domains (compare 1.3) from N-terminus to the C-terminus can be translated into the domain architecture for this protein e.g. in figure 1.8.

1.4 Large scale

Large scale genomics exist since 1995. The growing number of sequenced genomes is boosted by experimental techniques and bioinformatics. Many metazoan model organisms (*Mus musculus, Rattus norvegigus, Drosophila melanogaster, Saccharomyces cerevisiae, Caenorhabditis elegans*) are sequenced by now. In addition are evolutionary interesting species like *Ciona intestinalis* and *Ornithorhynchus anatinus* have been sequenced. The GOLD, Genomes OnLine

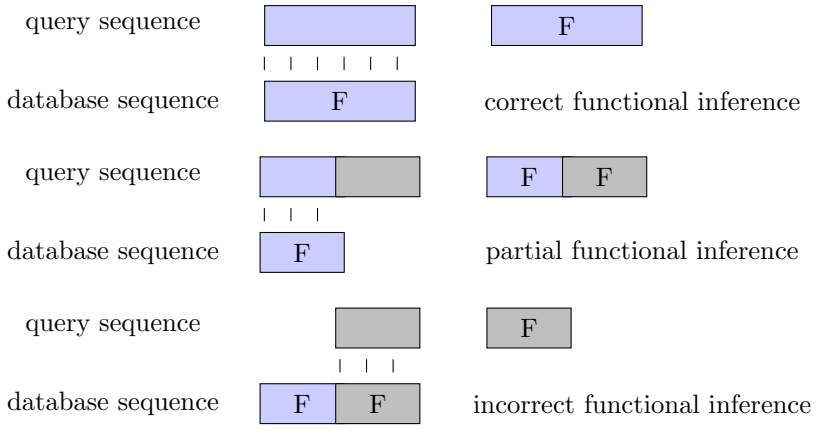


Figure 1.5 The correct annotated sequence is first compared to an full length hit and the function is correctly inferred.
The second example shows an query sequence which aligns in part to the full database sequence. The function is assigned to the hole search sequence. The new sequence and function is stored in the database.
A third sequence is aligned to the former unaligned part of the second sequence and the function is incorrect assigned.(modified from Ponting and Russell 2000)

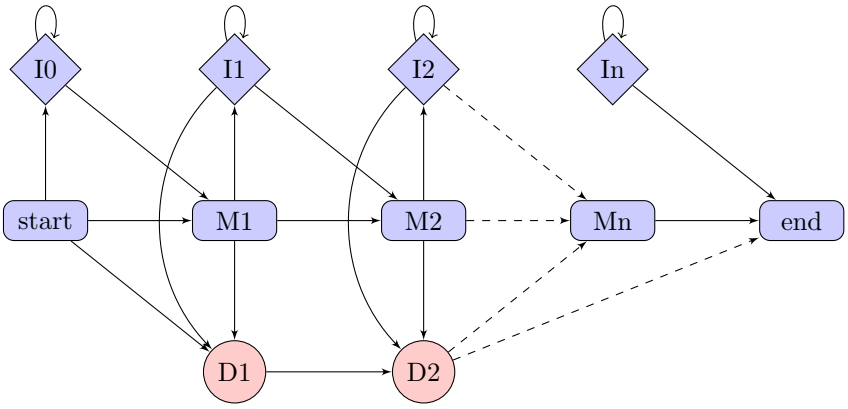


Figure 1.6 An Hidden Markov model stores the positions of an multiple alignment in three different states; match (rectangle), insert (diamond) and delete (circle). Transition probabilities between the different states are reflected by the arrows. In an match state is additionally the amino acid distribution for this position stored.

Database (Kyrpides 1999) lists not less than 4060 genome sequencing projects of which 860 are already finished. Interesting is the ratio of ongoing archeal, bacterial and eukaryotic genomes of 99:1963:1007 in contrast to 54:711:95 (september 2008) already sequenced genomes. The rate of ongoing eukaryotic genomes has increased 10 fold in contrast to around 2 fold for archeal, bacterial genome sequencing projects in contrast to allready finished projects.
The human genome, a big step in the sequencing of full genomes was accomplished by a race between Lander et al. 2001 and Venter et al. 2001. Many new techniques were developed and so the sequencing of new genomes is now cheaper and faster. The human genome consists of more than 3 billion base pairs. This amount of data is far too much to be analysed without computational prefiltering. One step is the prediction of genes. In 1997 the first high quality *ab initio* gene prediction programs like GENEFINDER identified up to 80% of the exons of vertebrate genes exactly (Burge and Karlin 1997). Most *ab initio* algorithms tend to overpredict genes and to miss small exons (Burset and Guigó 1996). Today most genes are also verified by experiments. The Ensemble database (ENSEMBL) uses the Genescan algorithm.
The ENCODE project (Birney et al. 2007) focused on one percent of the genome and made a detailed analysis. Some major outcomes are.

- Non-coding transcripts intercalate with protein-coding genes more often than expected.
- The number of transcription start sites (TSS) is ten fold higher than the number of protein coding genes.
- Around TSSs, the regulatory information is symmetrical and overall in the genome distributed as cluster.
- Histone structure and replication are more correlated than expected.

The hap map project (Consortium 2003) founded in 2003 as one outcome of the sequencing of the human genome focuses on single nucleotide polymorphisms (SNP). SNPs are correlated with large stable areas of chromosomes, they are therefore used to identify the different versions of them. Full human SNP Microarrays can be used to predict all alleles the genome contains. The advantage in contrast to a full genome chip lies in the number of information. Roughly 3 million SNPs (Frazer et al. 2007) mean an reduction factor of thousand regarding the analyzed information. In order to verify the SNPs a new project was founded. This next step is the sequencing of thousand human genomes, a combined effort by the Wellcome Trust Sanger Institute, Beijing Genomics Institute Shenzhen and the Human Genome Research Institute (Consortium).

They will provide the raw sequence data as the project continuous.
But not only genomics made huge steps proteomics gained speed too.
But there are critical points in the annotation process from gene to protein.
The prediction of gene structures is a complicated process. Especially open reading frames with multiple start or stop codons are not always predicted correctly. This could lead to truncation or elongation of the predicted genes especially in intron rich eukaryotes. Alternative splicing complicates the transfer of function from homologous sequences. What can we do if we find only partial best match for sequences?
We could check if a domain lies there. Domains are functional parts of proteins and called "Lego Set" of nature (Das and Smith 2000). Functional inference is done from a comparable small experimental dataset possible 5% (Valencia 2005) and is extrapolated. One way to increase the quality would be to transfer the function only based on orthologous sequences. Orthologues arise from a speciation event in contrast to paralogues, which arise from an intra genome duplication and are therefore free from selective pressure. This fact could lead to differences in functions or selectivity and in the expression pattern, too. Paralogues accumulate often mutations which lead to changed substrate specifity.
Most genes have more than one orthologue in other species. In the case of several gene duplications, we have to deal with many-to-many relations. One possible solution is to divide the orthologs in two different types, in- and outparalogues as proposed by Sonnhammer and Koonin 2002 (compare figure 1.9). The outparalogues are better suited for the prediction of protein function. Another way to decrease the rate of false positives for the detection of orthologs and paralogues lies in the consideration of the genomic context e.g. the conservation of gene order.
In general, 30-50% of the newly sequenced genes do not have a homologue (Stein et al. 2003). Additional factors, we have to keep in mind when comparing proteins, are temporal and spatial expression patterns and gene neighbourhood. The interaction partner could also be different for homologues. Further differences could lie in the 3d structure of the folded protein structure and even in the phenotype of a knockout mutant.
Most of the proteins are multi functional. Even one structural fold can be multifunctional. A special sort of proteins called moonlight proteins can carry out multiple functions depending on environmental factors (Jeffery 2003). Other proteins have different functions inside and outside the cell. Those "exceptions" are only accessible in a small scale approach.
Similar to the human genome project, structural fold projects have been started. These projects try to establish high-throughput 3d structure analysis. The RSGI, Riken Structural Genomics/Proteomics Initiative focuses on the full pro-

teome of *Thermus thermophilus* to fully understand the biological process. The conserved processes are then further examined in model organisms like *Arabidopsis thaliana* and *Mus musculus*. An important factor is that comparing protein structure can also be used to predict protein function (Thornton et al. 2000). Global similarities indicate biological tasks while local similarities can be structural motifs. Even structures with no similarities to known proteins can be analysed. Enzymes for instance have catalytic clefts with functional sites. A good starting point for further analysis.

Interactomics is based on three different experimental techniques/classes. First **Yeast two hybrid**(Y2H), carried out in yeast, where two target proteins are expressed, of which each is bound to one domain of the Gal4 transcription factor. If both proteins interact the two domains BD and AD of the Gal4 transcription factor are able to activate the reporter gene. This method is cheap and can be easily used for large scale screening. Two disadvantages are the high rate of false positives and proteins folded in the cytosol can not be detected (Van Criekinge and Beyaert 1999).

In vitro systems for discovering protein-protein interactions include are affinity chromatography, coimmunoprecipitation antibodys and newer approaches such as protein chip arrays (Howell et al. 2006).

In vivo systems to study, protein-protein interactions are typically based on immobilized antibodies which bind to an epitope of the protein of interest. After carefully washing all non interacting proteins away the complexes are eluted and analysed by mass spectrometry (Vasilescu et al. 2004). This is the most trusted method.

Mathivanan et al. 2006 compared different databases which contain protein-protein-interactions. They differ greatly in their number of binary non-redundant ppis for human, ranging from 101 in PDZBase, 346 in MIPS, 1067 in DIP over 5960 in Reactome, 6621 in Bind to 10244 in IntAct 11,367 in MINT and 36,617 in HPRD. Most interactions are stored in the (H)uman (P)rotein (R)esource (D)atabase (Mishra et al. 2006) which contains information about most known human proteins (compare table 1.3). These information are manually extracted from the literature. Each information e.g. subcellular localisation about a protein is linked to the original publication on pubmed. The (M)olecular (INT)eraction database (Zanzoni et al. 2002) focuses on mammalian interactions. For each interaction a confidence score is given based on the experiment and the number of interactions. Interactors can be examined graphically and additional information can be derived from OMIM, Online Mendelian Inheritance in Man (McKusick-Nathans Institute of Genetic Medicine and National Center for Biotechnology Information) is available. The IntAct database (Hermjakob et al. 2004) contains data from human and several other species. It

Table 1.3 Statistics of HPRD

Reference	Count
Protein Entries	25,661
Protein-Protein Interactions	38,167
Domains	455
PTMs	16,972
PubMed Links	270,466

The number of entries where taken on July 23, 2008.

provides special tools like the prediction of the best bait for protein pull-down experiments or HierarchView showing the interaction network as a two dimensional graph with highlighted nodes annotated with a given GO term. Data about protein complexes is stored in HPRD too. It is important that ppi data and complex data are strictly separated, which is required by the fact that we do not know which proteins in a complex really interact with each other. Protein interactions in a complex can be transferred into ppi data in different ways (compare figure 1.10). Assuming that all proteins in a complex interact with each other has the advantage that we do not miss interactions at the cost of a potential high rate of false positives. One way to solve this problem is the translation of the crystallographic structure into a graph as it has been done by Levy et al. 2006. The disadvantage is that not for every complex is and will be a crystallographic structure available. Therefore it is safer to analyse binary protein-protein interaction and complex data separately.

Genomic information can be used to identify interacting proteins. Neighbouring genes could be translated into functionally interacting proteins or proteins of the same cellular compartment or a functional pathway. A better indicator is gene fusion. Homologues of those fused genes have a high chance of being related and an even higher chance if the genes are orthologues. Physical interaction is very likely for genes conserved as pairs or cluster within a genome. Proteins originating form operons in prokaryotes are regulated together and part of one pathway. Eukaryotic homologues of proteins part of an operon have a higher chance of being part of the same pathway. Cluster of conserved genes indicate participation in same complex. The String database (Snel et al. 2000) provides information on proteins and their interactions based, among other things on genomic context and (conserved) coexpression.

1.5 Motivation

The work is splitted into three parts which reflect three different levels from protein to the interactom (compare figure 1.11). Domains are the functional parts building most proteins. How and when did the different domains arise in the evolution of species? Step by step or in bulk at the beginning of life. Domain architectures are built of domains. Do new domains lead to new domain architectures? Or does each step in evolution reshuffle the domains. What about the composition of domain architectures. Do they grow? Is their complexity changing? All those question are addressed in the chapter *History of domain architectures*.

Most proteins are linked into a network of interactions with other proteins. With ongoing evolution new proteins appear. Are they tightly integrated into the network? Do they form their own networks?

Protein complexes are part of many important pathways. Are those complexes comparable composited to the protein network? Do complexes have similarities to each other? I will deal with these questions in chapter *Evolutionary modules and evolving complexes*.

We learned about the evolution of proteins and their interactions. The next layer is the look at the whole network. Is there a pattern? Can we automatically identify groups of proteins with the same biological features, based only on the interactions and non interactions they share? The answers are given in chapter *Protein Interaction Networks - More than Mere Modules*.

Figure 1.7 Pictogram of the SH3 domain as generated by SMART.

Figure 1.8 Pictogram of the ANK-ANK-ANK-ANK-ANK-ANK-SH3-PDZ domain architecture as generated by SMART.

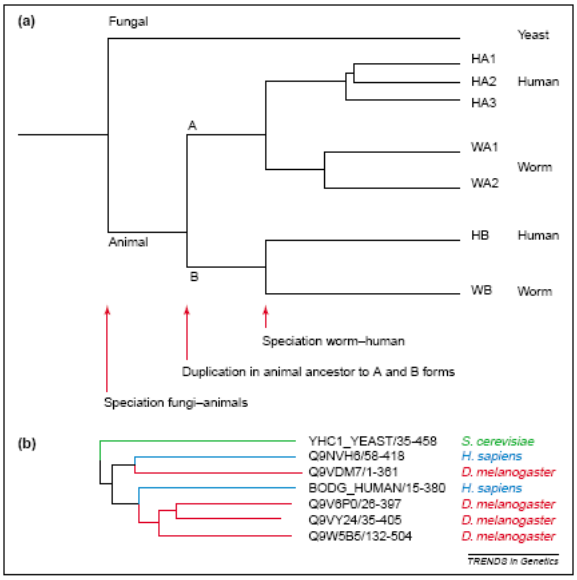


Figure 1.9 *The definition of inparalogs and outparalogs. (a) Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the humanworm split, into genes A and B. After the humanworm split, the A form was in turn duplicated independently in the human and worm lineages. In this scenario, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence inparalogs to each other when comparing human to worm. By contrast, the genes HB and HA* are outparalogs when comparing human with worm..However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animalyeast split pre-dates the HA*HB duplication. (b) Real-life example of inparalogs: ?-butyrobetaine hydroxylases. The points of speciation and duplication are easily identifiable. The alignment is a subset of Pfam:PF03322 and the tree was generated by neighbor-joining in Belvu. All nodes have a bootstrap support exceeding 95%.* (Picture and description from Sonnhammer and Koonin 2002)

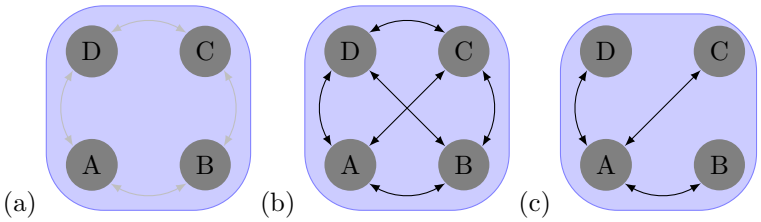


Figure 1.10 (a) The three proteins B,C and D forming a complex with target protein A are identified.
(b) The protein-protein interactions based on the matrix model.
(c) The protein-protein interactions based on the spokes model.

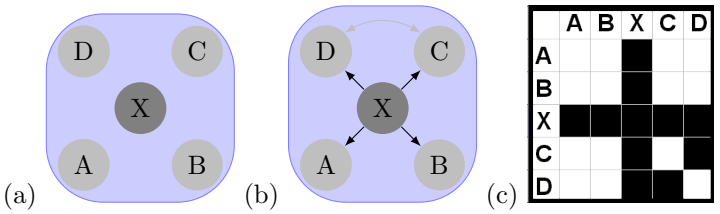


Figure 1.11 (a) Zoomed in, looking at proteins.
(b) One step back, analysing protein-protein interactions.
(c) The full picture, interaction based cluster in the proteome.

Chapter 2

History of domain architectures

2.1 Introduction

Although completing a genome sequencing project is a milestone for analysing an organism, it is only the first step of a long journey. The analysis of the proteome, the entire complement of proteins of an organism, is a bigger challenge. Helpful for the analysis of proteins is the fact that proteins consist of structural conserved and independently foldable parts, the domains. A protein can consist of several domains, which are connected by less conserved regions. A further feature of some domains is their genetic mobility. This leads to the question how important domains are in the evolution of proteins.
The best method today to detect known domains in protein sequences are hidden markov models (Eddy 1996). These are stored together with information regarding the domains in databases like SMART (Letunic et al. 2004) and Pfam (Bateman et al. 2004) and also in the meta-database InterPro (Mulder et al. 2003). Hidden markov models are trained with multiple sequence alignments of known family members of the respective domain. They incorporate position specific probabilities of inserts, deletes and the probability for each amino acid. They offer a more sensitive detection of domains within protein sequences as it is the case with blast. With hidden markov models, a large amount of protein sequences can automatically be scanned for domains. The rate of newly discovered domains is dropping, leading to the assumption that most of the existing domains are found (Copley et al. 2002). The next step after describing and cataloguing the domains is the large scale analysis of genomes, each of them

brings new protein families (Kunin et al. 2003).
Known protein families can be efficiently detected in large scale (Enright et al. 2002). But even between closely related organisms like *C. elegans* and *C. briggsae* only sixty percent of all protein coding genes are homolog (Stein et al. 2003). Domains are an elegant possibility to describe newly found proteins.
This lead to new fields like the distribution of domains in proteins. The work of Wuchty 2001 shows the network spanned by the combination of two domains which has the features of a scale-free network. Mott et al. 2002 predicted the localisation of domains and therefore proteins in cellular compartments based on a similar network. Apic et al. 2001 focused on the phylogenetic background of the domains in contrast to the works before. He found kingdom specific protein families based on ubiquitary domains of the three kingdoms. Domain recombination leads to kingdom specific proteins and could therefore be a factor in their evolution. In contrast to this, Lander et al. 2001 focused on basal events species. They compared the number of domain architectures in Saccharomyces cerevisiae, *Drosophila melanogaster* and *Caenorhabditis elegans* with *Homo sapiens*. In this work, the domain architecture was defined as the linear order of domains in a protein as predicted by the SMART database. One result was the higher number of different domain architectures in human than in other eukaryotic genomes. Most of the analysed domains participate in protein-protein interactions. This leads to the question if these networks have a higher complexity than in other organisms. Another hint that the reshuffling of domains is a driving factor for evolution is the fact that most known domains exist since the offspring of *Metazoa* or longer. Pal and Guda 2006 analysed the evolutionary distribution of 88,025 domains in the human proteome found with their subtraction method. They subtracted along the way from the origin of all life to human split into six steps domains found in e.g. Bacteria from those found in human. More than 92% are found in Metazoa or earlier in evolution. Itoh et al. 2007 did go one step further and analysed domain combinations and found that animal-specific domains are more often connected than other domains. The occurrence of domain combinations as supra-domains was analysed by Vogel et al. 2004, who found some 1400 overrepresented combinations of two or three domains. Lin et al. 2006 presented a web server to compare proteins on the level of domain architectures based on Pfam A.
This leads to the following questions. How has evolution acted on proteins between the very basic events that Apic et al. 2001 described and "now" as analysed by Lander et al. 2001?
Is there a general pattern how domains end up in different domain architectures?

2.2 Materials and methods

Domain architectures are defined as the linear order of domains in a protein (Letunic et al. 2006). All proteins and their corresponding domain architectures from the SMART database (Letunic et al. 2006) were kindly provided by Ivica Letunic. The estimation of the last common ancestor of one domain architecture is based on a tree. This tree requires to contain all species for which proteins should be analysed. Phylogenetic trees, have the highest probability of reflecting the correct relationships of species. But they are far from containing all species. Therefore was the taxonomic information collected by NCBI Taxonomy (NCBI-Taxonomy) was used as a basic tree. New molecular developments (Halanych 2004) especially at the base of the evolutionary tree were incorporated to further refine the constructed tree.

2.2.1 Convert protein to domain architecture

2.2.1.1 Overview

All proteins from the smart database were converted into their corresponding domain architectures following the scheme in figure 2.1.

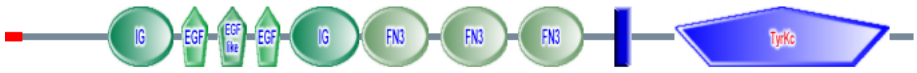
2.2.1.2 Smart domains

The Smart database (Letunic et al. 2006) focuses mainly on signalling, nuclear and extracellular domains, but incorporates all Pfam domains in addition. In case of overlapping domain predictions from Smart and Pfam, Smart domains were preferred.

2.2.1.3 Excluded domains

The domains listed in table 3.1 had to be excluded from domain architectures in this analysis for different reasons. SIGNAL, TM, COIL present additional information about intrinsic features of proteins. These are not domains but signal peptides, trans-membrane regions and coiled-coiled regions. The HMMs of the DM...-domains were all generated in an automated large-scale analysis of *Drosophila melanogaster* (Ponting et al. 2001) and are therefore mainly found in that organism so far.

Figure 2.1 Steps to generate domain architectures

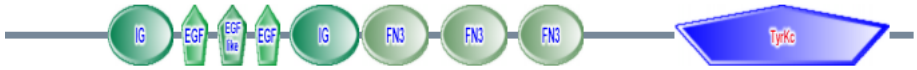


(a) Step 1: The domain architecture of the *tyrosine kinase receptor 1* from *Mus musculus* contains the following domains SIGNAL–IG–EGF–EGF_like–IG–FN3–FN3–FN3–TM–TyrKc

↓

Intrinsic features and DM...-domains where deleted (compare chapter 2.2.1.3).

↓

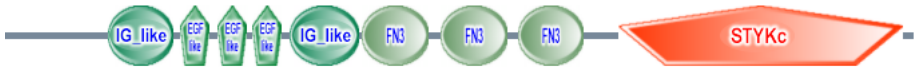


(b) Step 2: The Signal and Transmembrane parts are deleted from the domain architecture(compare table 3.1).

↓

Regrouping of domain supfamilies (compare chapter 2.2.1.4).

↓

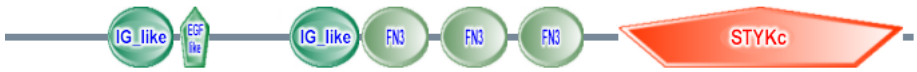


(c) Step 3: EGF, IGF and TyrKc are regrouped into their super families (compare table 2.2).

↓

Direct repeats of selected domains are contracted to one appearance (compare chapter 2.2.1.5).

↓



(d) Step 4: Three direct repeats of EGF_like were reduced to one. The final domain architecture used in the analysis IG_like–EGF_like–IG_like–FN3–FN3–FN3–STyKc

Table 2.1 Excluded Smart domains

SIGNAL
TM
COIL
DM3
DM4_12
DM5
DM6
DM7
DM8
DM9
DM10
DM11
DM13
DM14
DM15
DM16

Domains which are excluded for different reasons (compare chapter 2.2.1.3).

2.2.1.4 Domain super families

Some domains are further categorized in sub families. These cannot always be identified exactly. They were therefore replaced by their super families (compare table 2.2). Another regrouping had to be done with the large family of zinc-finger, some of those domains can not be exact distinguished when found together in one protein. So they were artificially grouped to ZNF_Gen.

2.2.1.5 One time counted domain architectures

Some domains have very short and divergent sequences (Andrade et al. 2000). The prediction of the exact number of repeated domains is error-prone and therefore the number of direct repeats of those domains is only counted once.

2.2.1.6 Pfam domains

Some domains in the Pfam database (Finn et al. 2006) are grouped in clans like the domain-super-families in Smart (compare chapter 2.2.1.4). To reduce the noise level these domains where replaced by their clan name (compare table 8.1).

Table 2.2 Domain subtypes combined in domain-super-families

Superfamily	domain subtypes
PTPc_DSPc	PTPc_DSPc, DSPc, PTPc, PTPc_motif
C2	C2, PI3K_C2
small_GTPase	small_GTPase, RAB, RAN, RAS, RHO, ARF, SAR
STYKc	STYKc, S_TKc, TyrKc
HTH	HTH_ARAC, HTH_ARSR, HTH_ASNC, HTH_CRP, HTH_DEOR, HTH_DTXR, HTH_GNTR, HTH_ICLR, HTH_LACI, HTH_LACI, HTH_LUXR, HTH_MARR, HTH_MERR, HTH_XRE
EGF_like	EGF_like, EGF, EGF_CA, EGF_Lam
IG_like	IG_like, IG, IGc1, IGc2, IGv
LRR	LRR, LRR_BAC, LRR_CC, LRR_RI, LRR_SD22, LRR_TYP
ZnF_Gen	ZnF_BED, ZnF_C2H2, ZnF_C3H1, ZnF_U1

Domain super families on the left and their included domains on the right.

Table 2.3 One time counted domain repeats

ANK
ARM
EFh
SPEC
WD40
RRM
EGF_like
IG_like
ZnF_...

All of those domains are only counted once for each number of direct repeat.

2.2.2 Generation of species tree

Taxonomic information represented by a tree was downloaded from NCBI Taxonomy (Wheeler et al. 2006) and combined with data from Halanych 2004. The taxonomic information from NCBI-Taxonomy includes all species with known sequences. The database has a low resolution for species relations especially at the level of cellular organisms. Therefore I combined it with the data presented by Halanych 2004 to increase the resolution as shown in table 2.4 and figures 2.2 and 2.3. All computations and analyses were realized on the final tree.

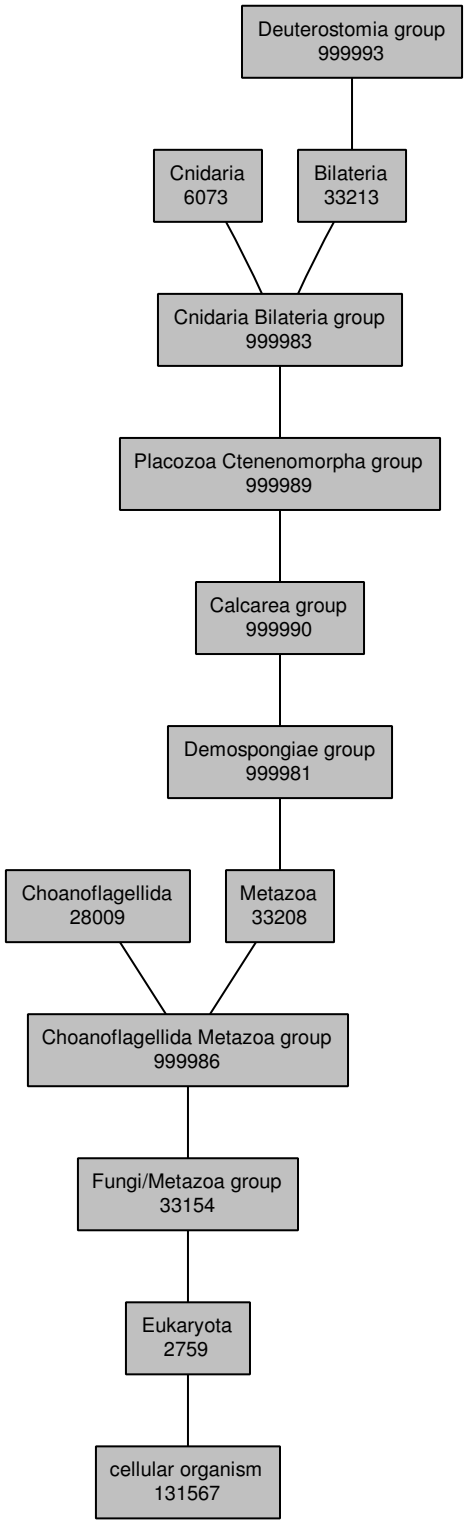


Figure 2.2 The modified basal part of NCBI-Taxonomy Part 1
From *cellular organisms* until *Deuterostomia group*.

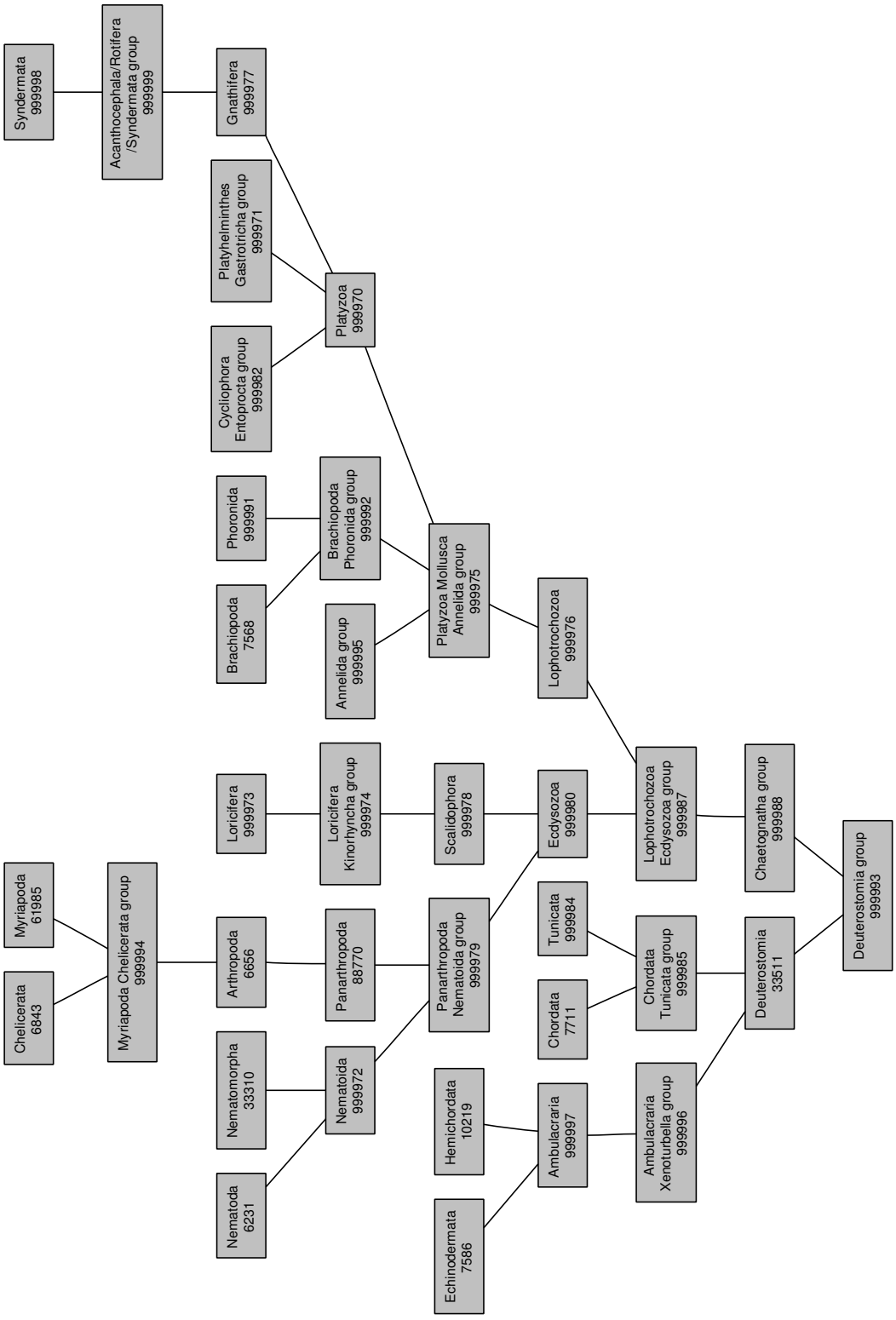


Figure 2.3 The modified basal part of NCBI-Taxonomy Part 2
From *Deuterostomia group* until diverse taxa.

Table 2.4 Modified Taxonomy

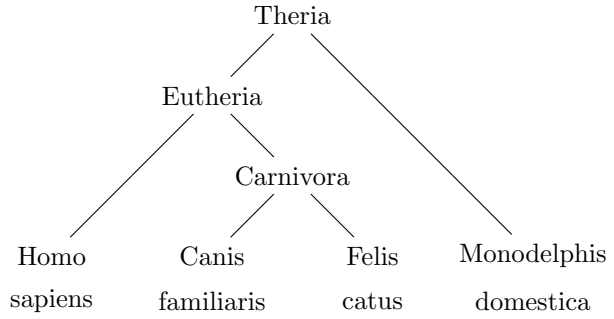
Taxid	Parent taxid	Taxon name
999999	999977	Acanthocephala/Rotifera/Syndermata group
999997	999996	Ambulacraria
999996	33511	Ambulacraria Xenoturbella group
999995	999975	Annelida group
6656	88770	Arthropoda
33213	999983	Bilateria
7568	999992	Brachiopoda
999992	999975	Brachiopoda Phoronida group
999990	999981	Calcarea group
131567	1	cellular organism
999988	999993	Chaetognatha group
6843	999994	Chelicerata
28009	999986	Choanoflagellida
999986	33154	Choanoflagellida Metazoa group
7711	999985	Chordata
999985	33511	Chordata Tunicata group
6073	999983	Cnidaria
999983	999989	Cnidaria Bilateria group
999982	999970	Cycliophora Entoprocta group
999981	33208	Demospongiae group
33511	999993	Deuterostomia
999993	33213	Deuterostomia group
999980	999987	Ecdysozoa
7586	999997	Echinodermata
2759	131567	Eukaryota
33154	2759	Fungi/Metazoa group
999977	999970	Gnathifera
10219	999997	Hemichordata
999976	999987	Lophotrochozoa
999987	999988	Lophotrochozoa Ecdysozoa group
999973	999974	Loricifera
999974	999978	Loricifera Kinorhyncha group
33208	999986	Metazoa
61985	999994	Myriapoda
999994	6656	Myriapoda Chelicerata group
6231	999972	Nematoda
999972	999979	Nematoida
33310	999972	Nematomorpha
88770	999979	Panarthropoda
999979	999980	Panarthropoda Nematoida group
999991	999992	Phoronida
999989	999990	Placozoa Ctenenomorpha group
999971	999970	Platyhelminthes Gastrotricha group
999970	999975	Platyzoa
999975	999976	Platyzoa Mollusca Annelida group
999978	999980	Scalidophora
999998	999999	Syndermata
999984	999985	Tunicata

The base for the modified tree of life was the taxonomy provided by NCBI-Taxonomy (Wheeler et al. 2000) this treelike structure was then modified with data computed by Halanych 2004. I created new ids from 999970 to 999999 and gave inner nodes names like *Choanoflagellida Metazoa group* for the parent node of *Choanoflagellida* and *Metazoa*.

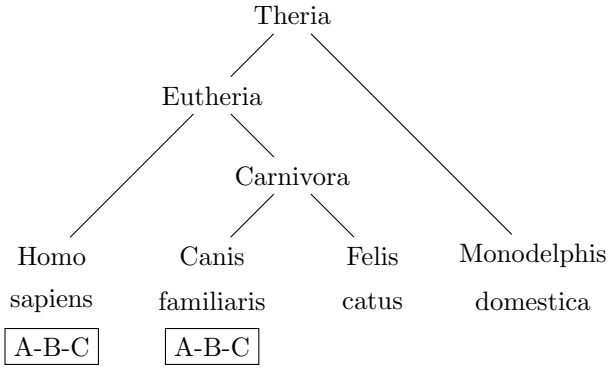
2.2.3 Last common ancestor

The last common ancestor for each domain architecture and therefore homologous protein was predicted by combining information from the domain architectures (compare chapter 2.2.1.1) with the taxonomic data (compare chapter 2.2.2) as described in the following steps.

1. A tree like structure with node parent pairs was built. This can be visualised as tree. *Theria* is the last common ancestor for the species *Homo sapiens*, *Canis familiaris*, *Felis catus* and *Monodelphis domestica* in this example tree. The tree generated from the information of NCBI Taxonomy contained roughly 300000 nodes.

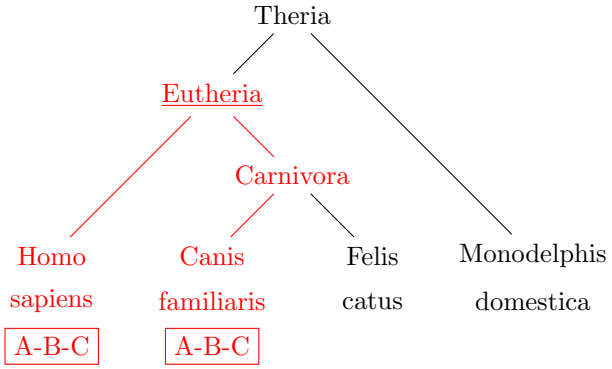


2. Proteins with the domain architecture A–B–C are found in human and dog.

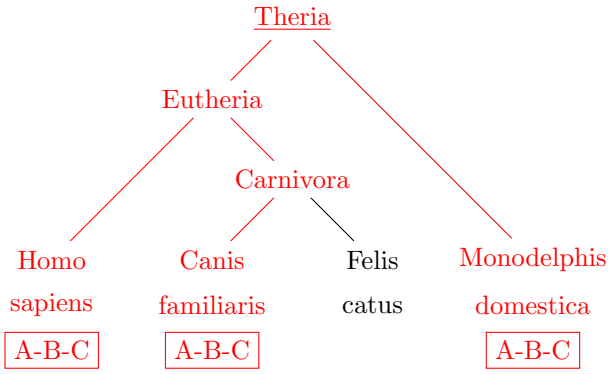


3. The last common ancestor for those proteins has arisen in *Eutheria*. Computationally, I solved this problem by comparing two lists of taxa from the root to these species containing those domain architectures.
   Theria>Eutheria>Homo>sapiens
   Theria>Eutheria>Carnivora>Canis families
   This shows that *Eutheria* is the last taxa they have in common. But *Felis cattus* does not have this domain architecture. This is most likely

explained by gene loss (compare chapter 1.1) in contrast to the possibility of the same architecture having arisen twice (compare chapter 1.1) or horizontal gene transfer (compare chapter 1.1) between non *Bacteria*. However, sequencing errors or false domain prediction are also a possible source of error.



4. Later with the sequencing of more species another protein with the domain architecture A–B–C is found e.g. in the short-tailed opossum. This would lead to a new last common ancestor, in fact *Eutheria* which has arisen earlier in evolution.



2.3 Results and discussion

I used 753166 proteins from a total of 3998331 proteins stored in Smart in a version from august 2006. The proteins not used contained no domains or could not be mapped to the taxomic tree. This dataset contains 6465 domains and 32868 domain architectures (compare table 2.5) and offers a higher resolution than a previous work from Pinkert 2004 which was based only on Smart domains combined in 7540 different domain architectures and more important without the newer molecular data allowing modifications of the taxonomic structure. The final taxonomic tree of 1313 species contained 1724 nodes with new domain

architectures. The top thirty nodes are shown as a tree in Figure 2.4 and in table 2.5. Interesting to note is that one quarter of all domain architectures are based on Smart domains, although only ten percent of all domains are from Smart. This underlies the fact that Smart focuses on genetically mobile domains.

2.3.1 New domains, new domain architectures

All domain architectures originated at one taxon have at least one domain originated at this taxon.
This hypothesis was analysed for the different taxa in the human line. As shown in table 2.6 the number of arisen domain architectures based on domains orifginated in the same taxon(column 3) is small in most taxa compared to all newly arisen architectures (column 2), for example in *Eutheria* only 25 out of 441 newly arisen domain architectures are based on domains originated in this taxon. These results show that domain architectures arise mainly as a result of the recombination of allready existing domains.

2.3.2 Complexity of domain architectures

I analysed the length and complexity of domain architectures at different levels in human evolution. At the beginning of cellular life (compare figure 2.5) most domain architectures are short and the maximum number of different domains in a single architecture is six. Longer domain architectures are mainly built from duplicated domains. In an intermediate step at the *Deuterostomia group* (compare figure 2.6) where the clades of human and worm split, the maximal number of different domains is still six but the stacks show a higher ratio of more complex architectures. At the level of *Eutheria* (compare figure 2.7) can we see a further shift to more complex architectures.

2.3.3 Same number of proteins, different number of architectures and transcripts

Human, mouse and worm have roughly the same number of proteins but an different level of complexity. Where is this complexity coded? I found a remarkable difference in the percentage of newly arisen domain architectures from their last common ancestor, the *Deuterostomia group*. After their clades split the percentage of newly arisen domain architectures is highest in the human clade (compare table 2.7).
The number of transcripts per gene is slightly different between the species. Human has around eight percent genes with two transcripts per gene.

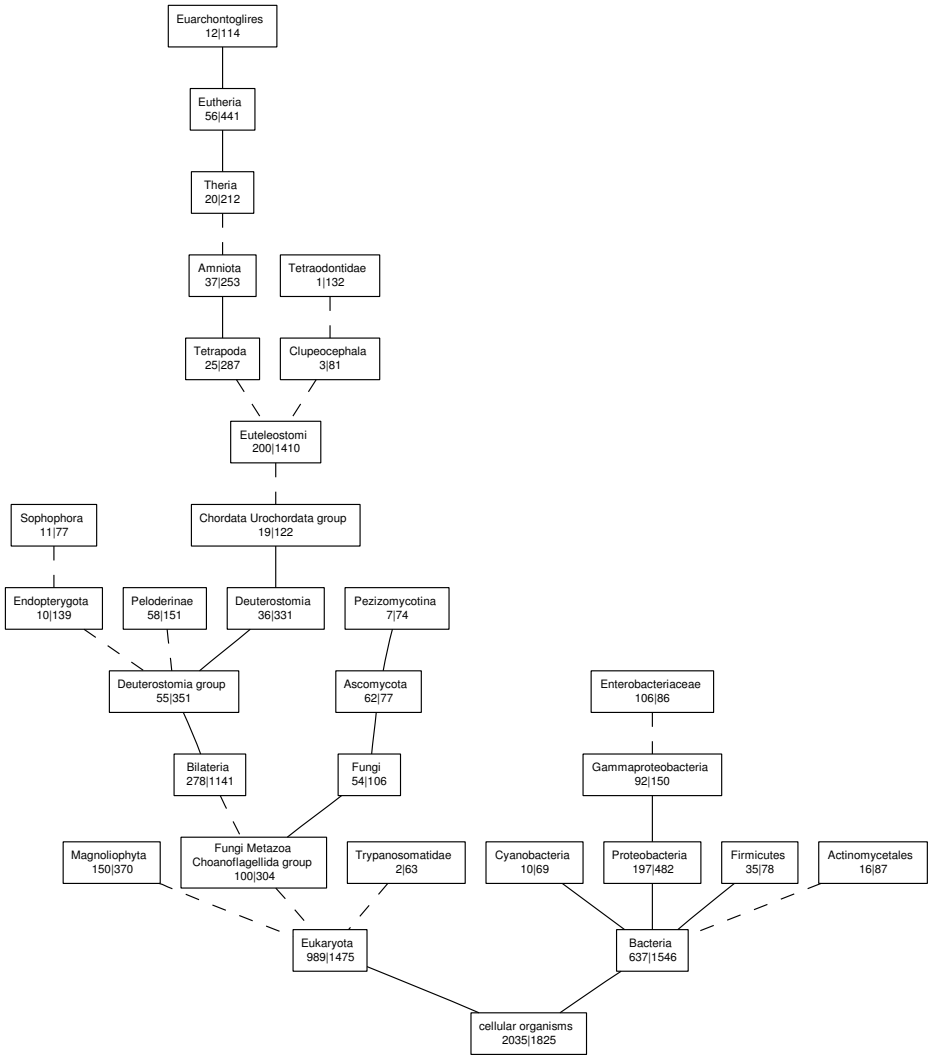


Figure 2.4 The thirty top nodes with the most new arisen domain architectures as a tree. Dashed lines indicate one or more missing taxa in between.

Table 2.5 Arisen domains and domain architectures

Taxon	Domain		Domain Architecture	
	Smart(only)	all	Smart(only)	all
whole tree	648	5817	6465	32868
cellular organisms	242	2035	357	1825
Bacteria	10	637	230	1546
Eukaryota	160	989	552	1475
Euteleostomi	23	200	628	1410
Bilateria	69	278	553	1141
Proteobacteria	3	197	65	482
Eutheria	12	56	174	441
Magnoliophyta	7	150	60	370
Deuterostomia group	10	55	161	351
Deuterostomia	4	36	157	331
Fungi Metazoa Choanoflagellida group	21	100	116	304
Tetrapoda	5	25	119	287
Amniota	7	37	97	253
Theria	1	20	106	212
Peloderinae	3	58	65	151
Gammaproteobacteria	0	92	16	150
Endopterygota	1	10	52	139
Tetraodontidae	0	1	44	132
Chordata Urochordata group	5	19	69	122
Euarchontoglires	0	12	54	114
Fungi	6	54	31	106
Actinomycetales	0	16	13	87
Enterobacteriaceae	1	106	9	86
Clupeocephala	1	3	33	81
Firmicutes	2	35	5	78
Ascomycota	3	62	22	77
Sophophora	0	11	31	77
Pezizomycotina	0	7	23	74
Cyanobacteria	0	10	9	69
Trypanosomatidae	0	2	23	63

Thirty nodes with numbers of arisen domain architectures and domains ordered by most new domain architectures descendingly. Its notable that the number of arisen domains is more rapidly decreasing than the number of originated domain architectures.

Table 2.6 Taxa in the evolution of human

Taxon	new domain-architecture	based on new domains	new domains
cellular organisms	1825	1752	2035
Eukaryota	1475	731	989
Fungi Metazoa			
Choanoflagellida group	304	25	100
Metazoa	2	0	1
Demospongiae group	34	2	8
Calcarea group	1	0	1
Placozoa Ctenenomorpha group	0	0	1
Cnidaria Bilateria group	43	5	13
Bilateria	1141	238	278
Acoelomorpha	0	0	0
Deuterostomia group	351	22	55
Deuterostomia	331	15	36
Chordata Urochordata group	122	4	19
Chordata	6	0	4
Craniata	2	0	3
Vertebrata	12	0	2
Gnathostomata	8	1	11
Teleostomi	0	0	0
Euteleostomi	1410	145	200
Sarcopterygii	0	0	0
Tetrapoda	287	20	25
Amniota	253	10	37
Mammalia	0	0	0
Theria	212	1	20
Eutheria	441	25	56
Euarchontoglires	114	1	12
Primates	1	0	0
Haplorrhini	0	0	0
Simiiformes	1	0	1
Catarrhini	11	0	3
Hominoidea	1	0	0
Hominidae	6	0	0
Homo/Pan/Gorilla group	61	0	4
Homo	0	0	0
Homo sapiens	314	8	19

For each taxon is the number of new domain architectures counted and then compared to the new domain architectures containing domains arisen at this taxon. Some domains occur only in architectures e.g. *cellular organisms*.

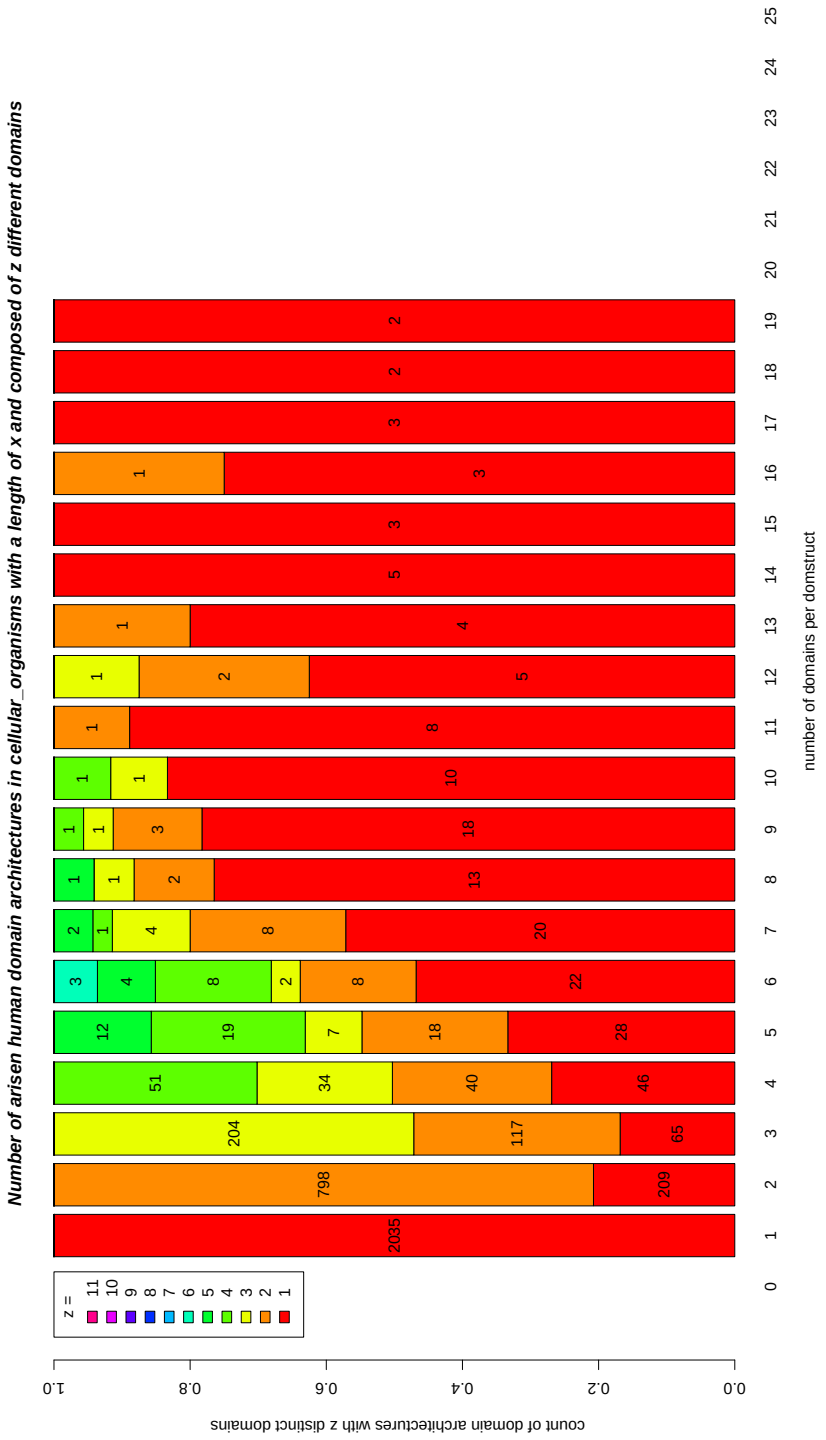


Figure 2.5 Stacked plot for domain architectures arisen at organisms with length of domain architecture on the x-axis and count of domain architectures with z different domains stacked.

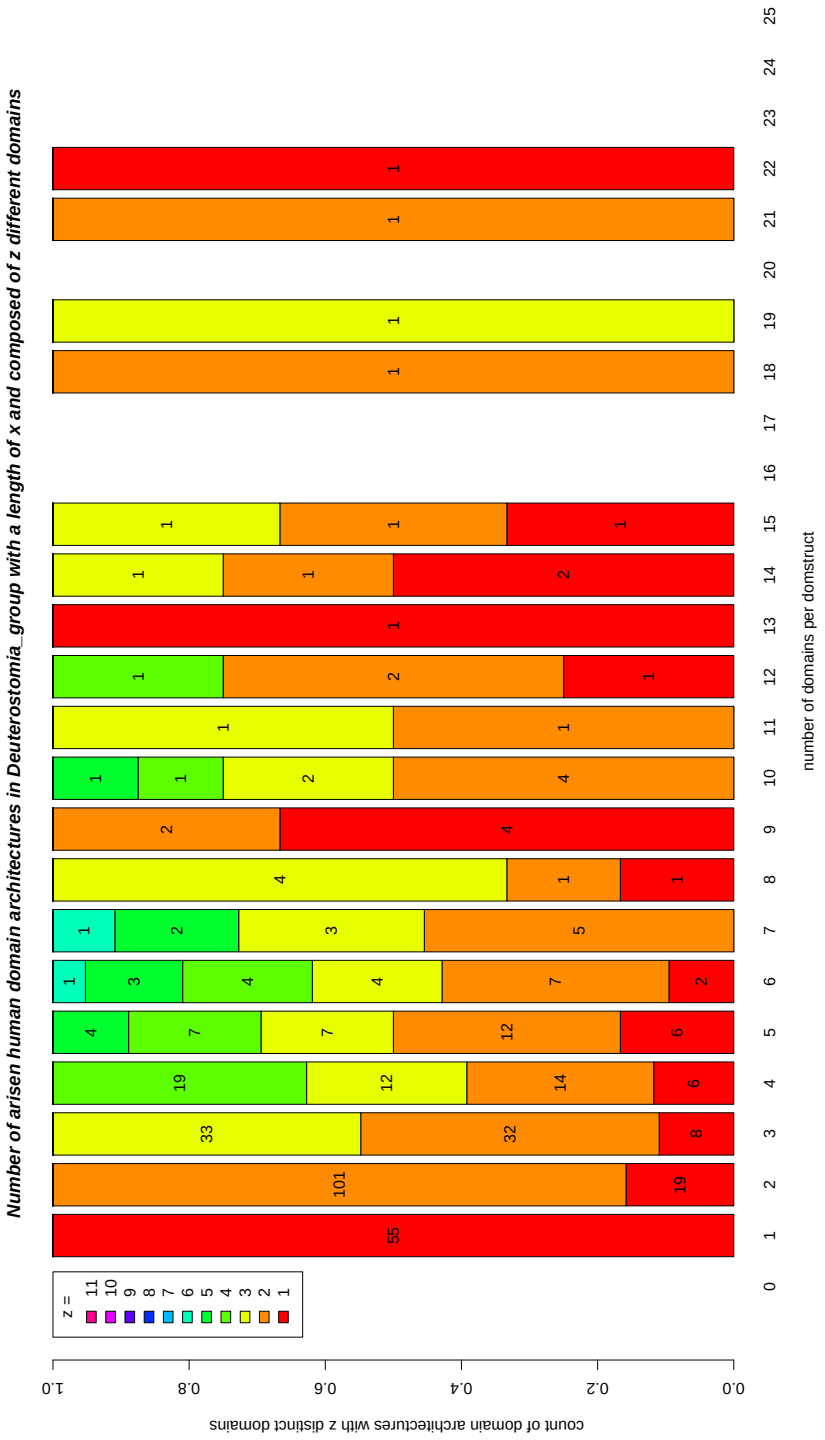


Figure 2.6 Stacked plot for domain architectures arisen at groupwith length of domain architecture on the x-axis and count of domain architectures with z different domains stacked.

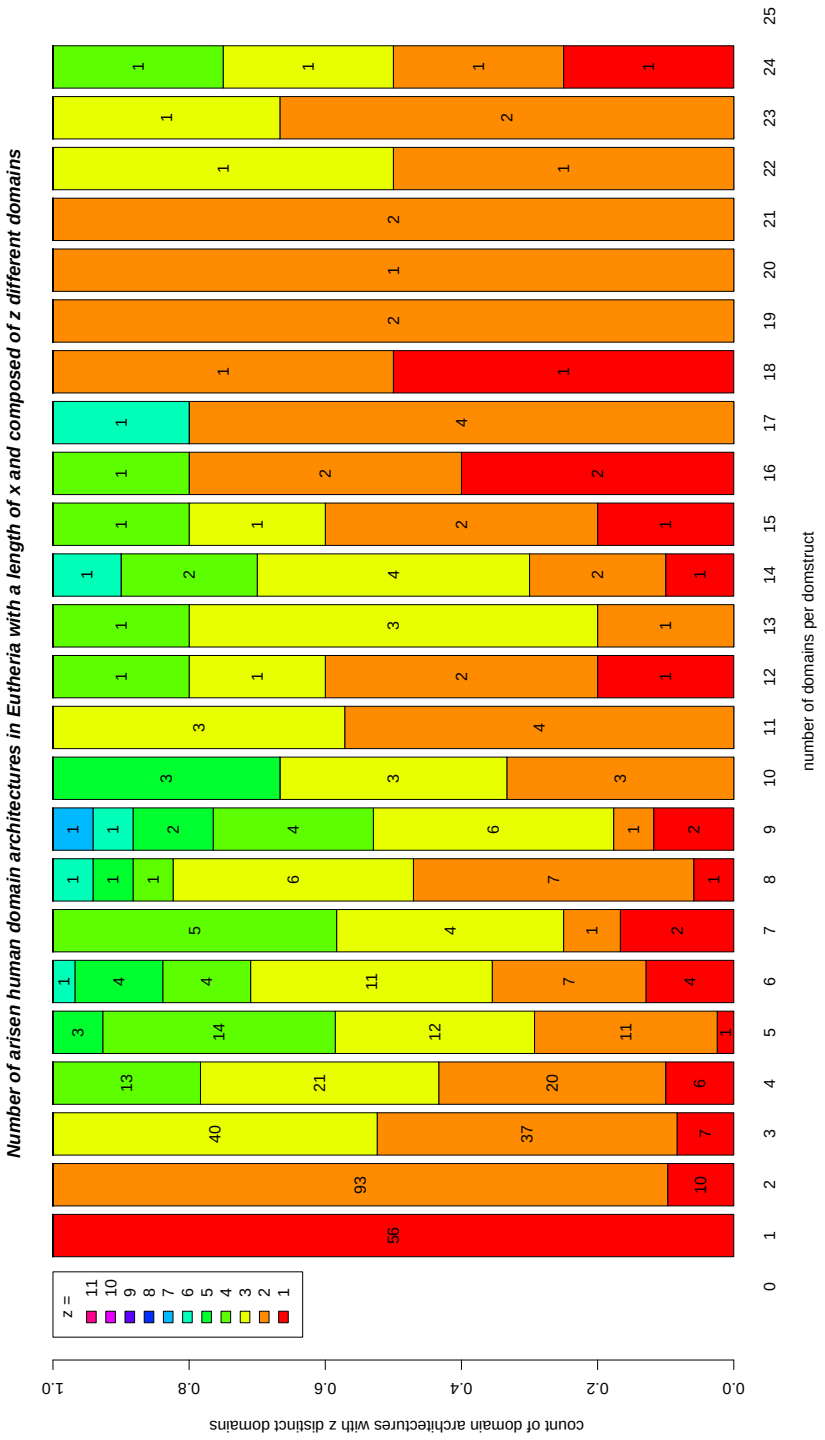


Figure 2.7 Stacked plot for domain architectures arisen at with length of domain architecture on the x-axis and count of domain architectures with z different domains stacked.

Table 2.7 Arisen domains and domain architectures in human, mouse, worm and fly

Domain architectures	Homo sapiens	Mus musculus	Drosophila melanogaster	Caenorhabditis elegans
Deuterostomia group	68.18%	71.84%	87.69%	82.78%
species	5188	4262	3200	3026

Deuterostomia group is the last common ancestor for human, mouse, fly and worm. They share all domain architectures arisen until then. At this taxon some of their clades split and new domain architectures arise. The number of new domains is very small (compare table 2.5) and therefore included.

2.3.4 Horizontal gene transfer

de la Cruz and Davies 2000 speculated that horizontal gene transfer (HGT) is a common mechanism to acquire new genes not only active in bacteria through e.g. *Bacteriophages* but in *Eukaryota* through e.g. *Bacteria* too. They believe that major leaps in evolution could be driven by massive HGT events. Stanhope et al. 2001 presented data which rejects the hypothesis of HGT from *Bacteria* to most multicellular organisms. Which impact has this on the data presented here?
If HGT events from *Bacteria* to *Eukaryota* introduced new proteins to a species, this would lead to the (new) last common ancestor *cellular organism* for the corresponding domain architectures. A possible test for critical domain architectures, would be to look for domain architectures common in *Bacteria* and appeared late in the evolution of *Eukaryota*.

2.3.5 Convergent Evolution of domain architectures

Gough 2005 has estimated that convergent evolution of domain architecture happens in less than four percent. I analysed the domain architectures for domain pairs which are part of domain architectures arisen at later evolved taxa (compare table 2.9). In most cases one direction of domain architectures is preferred. The table shows only combinations of two domains, for three or more domains, the results should be even more clear. Another point is that the reverse domain combinations can easily occur when two domain architectures / genes fuse in evolution. But overall, the chance of convergent evolution of complex domain architectures should be small. Interesting is the fact that some domain combinations have a disbalance between N-terminal and C-terminal attached domains e.g. LRR–LRRCT. This could be a hint for different duplication events

Table 2.8 Genes transcripts and domain architectures

Number of transcripts per Gene	Homo sapiens	Mus musculus	Drosophila Melanogaster	Caenorhabditis elegans
Unmapped Domain architectures	7.77	3.01	8.55	10.2
1	82.41	93.08	89	88.18
2	8.01	3.55	2.18	1.5
3	1.33	0.33	0.2	0.11
4	0.39	0.02	0.07	0.01
5	0.08	0.01	0	0.01
6	0	0	0	0
7	0.02	0	0	0
Transcripts	15837	14574	8983	11366
Protein coding Genes	22939	24498	14041	20071
Genes mapped	14607	14135	8215	10206
Unique domain architectures	5188	4262	3200	3026

The number of transcripts per gene was counted for each of the species.

followed by neofunctionalization of orthologous proteins. Or it is only an artefact of two common domains which easily combine.

2.4 Conclusions

Consistent with earlier works, this analysis shows that most domains have arisen early in evolution. The large number of genes identified in full genome sequencing projects still gives rise to some new domains. While early in evolution the number of originated domains matches the arisen domain architectures the rate of new domains has rapidly decreased in the course of evolution. Longer and more complex domain architectures have arisen. The complexity of a domain architectures is given by the number of distinct domains. New domains are only integrated in a small fraction of new domain architectures. This means that new domain architectures arise mainly from reshuffling of old architectures. The fact that new architectures are longer and more complex indicates that new architectures arise from the fusion of already established architectures with domains or other architectures. The duplication of domains of an architecture would decrease the complexity to length ratio.
Human, mouse and worm have roughly the same number of proteins but their domain architectures differ in number and complexity. They share all domain architectures arisen until their clades split. After that point more and more complex architectures arose in human compared to the other species. The number of proteins can be the same in different organisms, but there can be large differences in the number of distinct domain architectures. The complexity of domain architectures increases too. The function of a protein can be described by the sum of the function of its domains. Therefore can we conclude that even if the number of proteins stays the same a greater specialization of the proteins takes place. Horizontal gene transfer can not be excluded as a factor. This work focuses on multi cellular organism and there should be the number of events small and therefore the influence on this analysis. The probability of convergent evolution decreases with the increase of the length of domain architectures. As most domains are arisen early in evolution the influence of convergent evolution should be insignificant in this analysis.

Table 2.9 A–B domain combinations and attachment

DomainA–DomainB	N	C	NC	A–B	B–A
HisKA–HATPase_c	81	10	131	223	2
HisKA–HAT	0	11	212	223	0
LRR–LRRCT	59	8	112	180	103
KRAB–ZnF_Gen	0	73	97	171	50
HATPase_c–REC	77	5	69	152	2
LRRNT–LRR	4	56	87	148	17
DEXDc–HELICc	37	40	31	109	3
LRRCT–LRR	5	2	95	103	180
PAS–PAC	15	33	45	94	76
IG_like–FN3	20	29	42	92	53
CUB–CCP	18	7	62	88	71
LDLa–EGF_like	32	7	45	85	74
LRRCT–LRRNT	5	0	75	81	4
PAC–PAS	28	18	29	76	94
EGF_like–LDLa	10	6	57	74	85
CCP–CUB	12	6	52	71	88
PAC–HisKA	5	10	55	71	0
GAF–HisKA	4	12	47	64	0
EGF_like–LamG	40	4	13	58	27
RING–BBOX	3	41	12	57	6
PAS–HisKA	4	10	41	56	0
SET–PostSET	32	4	18	55	4
FN3–IG_like	23	3	26	53	92
LRRCT–EGF_like	42	0	8	51	0
FN3–PTPc_DSPc	28	1	20	50	1
ZnF_Gen–KRAB	9	9	32	50	171
STYKc–S_TK_X	27	13	6	47	2
ZnF_Gen–HOX	11	4	30	46	36
AT_hook–ZnF_Gen	3	8	31	43	26

Domain combinations and their occurrence in the same or later evolved taxa. N indicates the upstream and C the downstream occurrence of attached domains in the same domain architecture and NC both. The number of domain architectures with the order A–B and B–A of domains found in later arisen architectures is shown in columns A–B and B–A. Combinations of the same domain or the Smart and Pfam model detected versions of this domain are filtered out for this analysis.

Chapter 3

Evolutionary modules and evolving complexes

3.1 Introduction

With ongoing evolution the amount of proteins and the size of protein-protein interaction networks has been increasing in most cases. *How are the proteins evolutionary added to the network?* To address this question we recently developed a new feature for SMART (Letunic et al. 2006), which allows the prediction of the taxonomical origin of domain architectures and therefore proteins. Combining this with protein-protein interaction data from human would allow us to analyze the evolution of protein networks.
The protein-protein interaction networks of *S. cerevisiae* and *H. pylori* show a scale free behaviour (Jeong et al. 2001, Wagner and Fell 2001). It has been shown that scale free networks can emerge based on two mechanisms: First, the network grows by forming a connection between new proteins and a protein already in the network. Second, the probability for a new protein to form a connection with a protein already well connected is higher than for less connected proteins. This is called preferential attachment (Barabasi and Albert 1999).
Meanwhile large datasets for human complexes are available too and present the possibility to compare the composition of protein complexes with the structure of the interaction network.

3.1.1 Basic Network Features

In their work "The Architecture of Biological Networks", Wuchty et al. 2006 definend some basic features needed to describe and classify biological networks. Networks are generally *nodes* which are connected by *links*. In a protein-protein interaction network, proteins correspond to nodes and links to interactions. Links might be directed, for example in signalling networks. Protein protein interaction networks are treated as undirected. The number of links of a node is defined as a **node's degree** (connectivity).

$$Node_{ik} = \text{links } (k) \text{ of node } i$$

We could compute the average node degree for the whole network, but this would miss the potential degree variations of the network.
The **degree distribution** is a more accurate measure.

$$P(k) = \text{number of nodes with exactly } k \text{ links}$$

Most nodes are connected by many different paths. It is therefore useful to define the **shortest path**.

$$l_{ij} = \text{shortest path between nodes } i \text{ and } j$$

A good way to estimate the whole networks navigability is the **mean path length**.

$$[l] = \frac{2}{N((N-1)}$$

Networks with a low mean path length like social networks with $[l]$=6 are referred to as 'small world' networks.
Networks in real systems have a non-random distribution of links and nodes. They have a tendency to cluster. Clusters are parts of the network, in which nodes are more connected to each other. This can be measured by the **clustering coefficient of node i**.

$$C_i = \frac{2n_i}{k_i(k_i-1)}$$

Were n_i is the number of links connecting all neighbours k_i of node i to each

other. This value for the whole network can be computed as **average clustering coefficient for all nodes**.

$$[C] = \frac{1}{N} \sum_{i=n}^{N} C_i$$

Or more specific the **average clustering coefficient for all nodes with k links**.

$$C(k] = \frac{1}{N} \sum_{i=n}^{N} C_i$$

These values enable us to describe the network. If for example the $C(k]$ is independent from k, this means that the network is homogenous or consists of many small tightly linked cluster. On the other hand if $C(k) \sim k^{-1}$ than the network architecture consists of hierarchical, sparsely connected nodes and highly clustered areas. Hubs, proteins with a high k connect the highly clustered neighbourhoods. The classification of networks is done with $P(k)$ and $C(k)$ while (k) and (l) and (C) are unique for each network and therefore more specific.
Proteins in a network which have an high k, meaning many links, are called hubs. The Network can be divided into module forming proteins which are higly connected together and sparsly connected to proteins outside this module. The concept of proteins linking modules was further analysed by Yu et al. 2007. They divide nodes into four classes (compare figure 3.1). Those classes are defined by their node degree(hub↔non-hub) and their betweenness (bottleneck↔non-bottleneck). Hub-non-bottleneck nodes are party hubs and hub-bottleneck resemble the features of date hubs. Non-hub-bottleneck are essential for the network despite their low number of connections. Non-hub-non-bottleneck are the standard and make up most of the nodes.
In general, we can distinguish between two general types of networks, random and hierarchical.

3.2 Materials and methods

3.2.1 Age tagging of proteins

The taxon/taxid of arising for each protein ancestor was appointed in two steps based on their domain architectures. First, the protein's domain architecture was determined as described in chapter 2.2.1.1. And second, the resulting, if any, domain architecture was matched to the set of 39333 architectures and domains

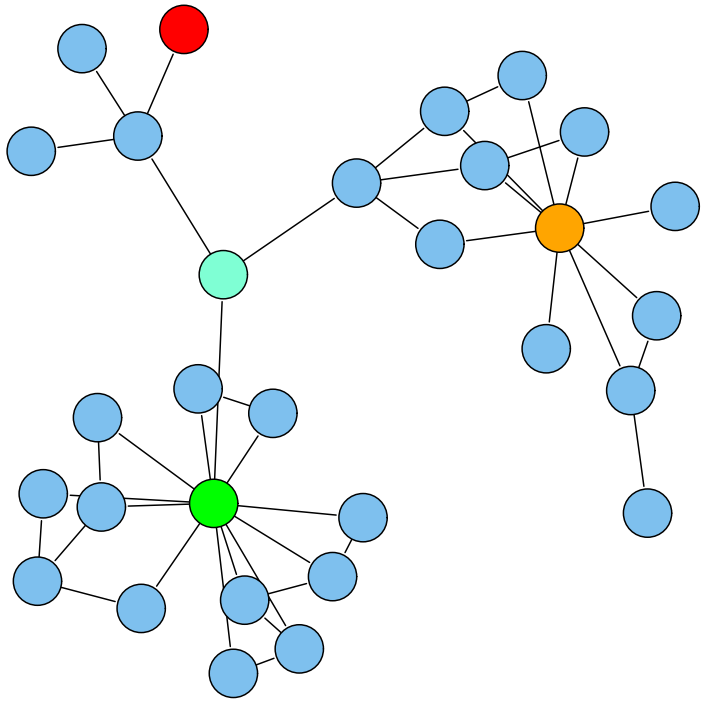


Figure 3.1 Four different node classes and their descriptors are defined by Yu et al. 2007
Colour — Type — Node degree — Betweennes
Green — Hub-bottleneck — 12 — 253.5
Blue — Non-hub-bottleneck — 3 — 256
Yellow — Hub-non-bottleneck — 9 — 133.5
Red — Non-hub-non-bottleneck — 1 — 0

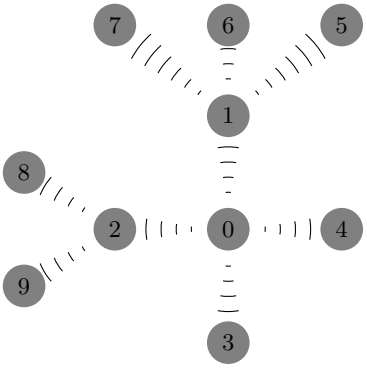
from chapter 2.3. Proteins in the network with no match were assigned the taxid zero.

3.2.2 Protein interaction network

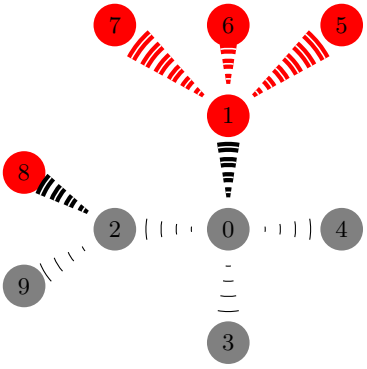
3.2.2.1 Data

The protein-protein interaction data was taken from the Human Protein Reference Database (HPRD). The work is based on the release 6 from September 2006. The interactions were found with three different methods, based on *yeast two hybrid*, *in vitro* and finally *in vivo* methods. The network contained 8503 proteins and 34013 interactions found with at least one of three methods, this means 4.00 interactions per protein on average. A data set based on *in vivo* methods only contained still 27386 interactions for 5840 proteins, resulting in 4.69 interactions per proteins on average. In the work, the *in vivo* set was used to get results of highest confidence. The basic features of the network were analysed with the statistical language R (Ihaka and Gentleman 1996) and in particular the igraph library (Csardi and Nepusz 2006).

3.2.2.2 Algorithm

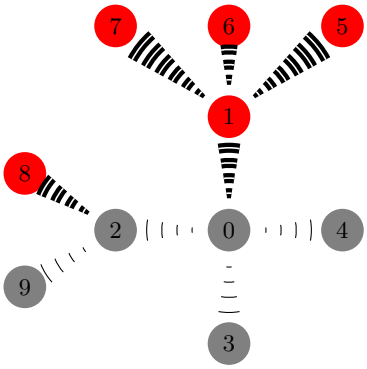


1. Tag proteins with taxon of emergence as described in chapter 3.2.1.

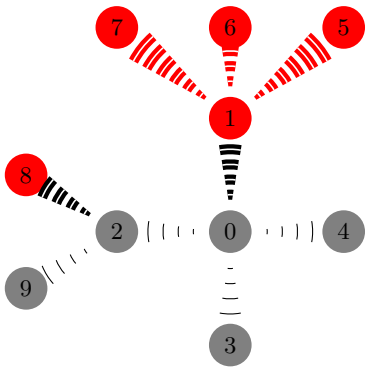


3. Count numbers of interactions:

tagged-tagged ⇔ tagged-non_tagged.
6⇔ 2.



2. Classify all interactions of tagged proteins into tagged-tagged and tagged non-tagged.



4. Calculate ratio of interactions:
tagged-tagged/all
(6 / 6+2 = 0.75).

3.2.3 Random model interaction network

The same network as before was used to calculate the random model. The random model was computed stepwise from 10 to 2000 tagged proteins of the full network with a step size of ten. Each step consisted of ten thousand repeats. In one step, x proteins were randomly tagged and then their ratio of interactions was calculated as explained in chapter 3.2.2.2.
The ten thousand ratios for each number of tagged proteins were then combined to a box plot. Each box contains 50% of all ratios and above and below respectively the lines are the 5% highest and lowest ratios. The centre of the boxes is rising with more proteins tagged. If all proteins are tagged, the ratio of connections is one. If the proteins arisen at one taxa are randomly integrated, their ratio of interaction should lie inside a box, corresponding to their number. If preferential attachment takes place the ratio should be below average for all taxa.

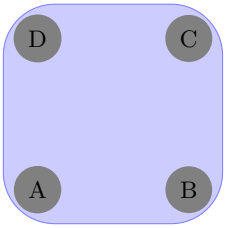
3.2.4 Complex evolution

3.2.4.1 Data

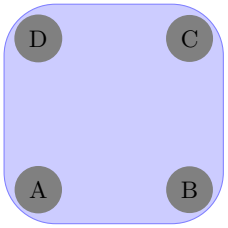
The protein-complex data were taken from HPRD too (compare chapter 3.2.2.1).

3.2.4.2 Algorithm

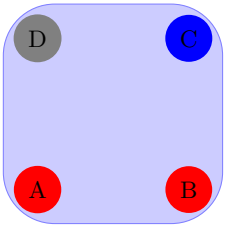
The complexes were analysed for the age of their proteins. I counted the number of complexes in which pairs of proteins tagged with different taxa of emergence occurred as shown in the following pictures. Each combination was only counted once for each complex. The complexes with at least one protein arisen at a taxon were also counted to get 100 percent of all complexes for this taxon.



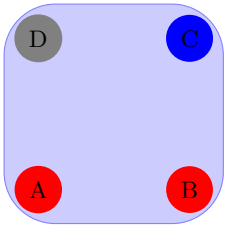
1. One protein complex consisting of 4 proteins A,B,C and D.



2. Tag proteins with their estimated point of arising.



3. The coloured proteins have the same age and for the grey one, no information is available.



4. Count the all pairs of proteins of different age.
blue:red=1; red:blue=1

3.2.4.3 Complexes sharing domain architectures

In the yeast proteome, it was estimated that 20 percent of all complexes have arisen by complex duplication (Pereira-Leal and Teichmann 2005). This is hard to analyse for all the known human complexes but I looked for complexes sharing the same domain architectures and and thus, sharing the same functions. This was done by estimating the domain architecture for all proteins in the HPRD complex data as explained in chapter 2.2.1.1. The complex data contained x

complexes with y proteins but only z different proteins which are build of a different domain architectures.

3.3 Results and discussion

3.3.1 Basic network features

I analysed the nodes of the network for betweenness (compare figure 3.2). This value reflects the number of paths going through a specific node. The distribution of the top one percent nodes is different from the overall distribution in two points. Proteins arisen at *cellular organisms* and *Eukaryota* are less prominent and bilaterian ones are more.
The node degree of the top 58 nodes shows relatively more proteins from *Bilateria* and the *Cnidaria Bilateria group* as the degree for all proteins, compare figure 3.3. The top ten percent node degree distribution looks like the distribution for all nodes. This observation argues against the theory of preferential attachment. If preferential attachment had taken place, the nodes estimated to be arisen earlier in evolution would be enriched for connections compared to the full set. On the other hand the *Bilateria* and the *Cnidaria Bilateria group* are taxa arisen relativly early. Perhaps there are other features like a special biological process, which took place at that taxa and influenced the attachment too.
Combining those sets of top node degree and top betweenness enabled me to identify the hub-bottleneck nodes. From the top ten percent nodes with the highest betweennes, 436 were also found in the top ten percent degree set. The other 148 nodes are non hub bottleneck nodes, but their number could decrease if the set of protein-protein interactions became more complete and some of those nodes gained additional links.

3.3.2 Distribution of ratio of connections for arisen proteins

The ratio of interactions is higher than in the random model for most taxa (compare figure 3.4). The variation of the connection ratio for 10 to 50 proteins is decreasing and after that constant(compare figure 3.5). Therefore, all results in this area had to be excluded from the analysis. The only other connection rate which is below the median of the ten thousand repeats is the connection rate for proteins with unknown age. Around one third of all proteins are emerged at the *cellular organisms* taxon. Therefore its not surprising that the connection

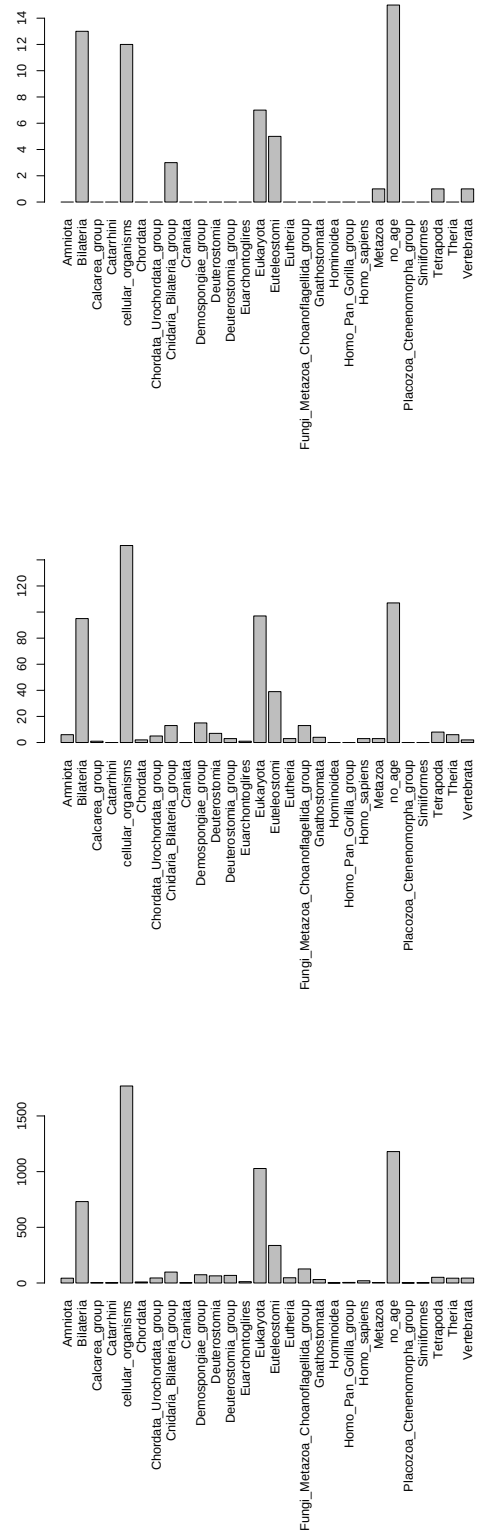


Figure 3.2 The betweenness is shown for the highest 58, 584 and all 5840 proteins of the network. It reflects the number of paths going trough the node.

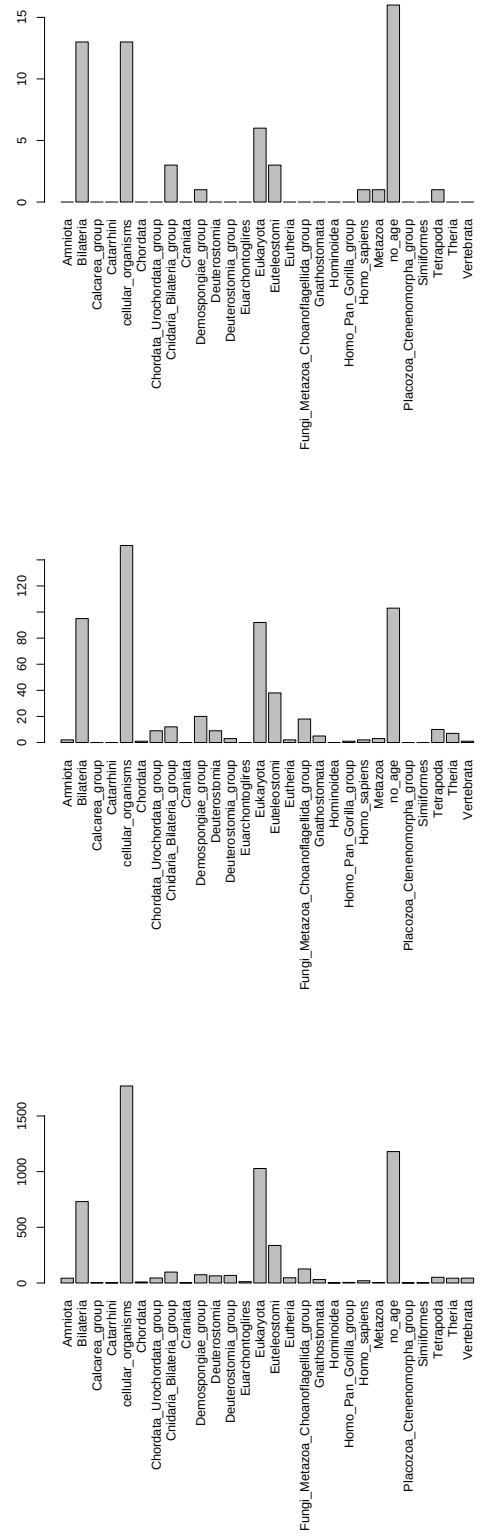


Figure 3.3 The degree is shown for the highest 58, 584 and all 5840 proteins of the network. It reflects the number of directly connected nodes.

ratio is near the median of connections. But its still an astonishing clear result that proteins arisen at one taxon form evolutionary modules. How do they look like?

It‘s depending on the number of proteins. The chance that two proteins are connected increases with the number of proteins. What we see is a higher ratio, not an ratio, of one. This means proteins arisen at one taxa are more connected than expected on random. They are not only exclusively connected to each other.

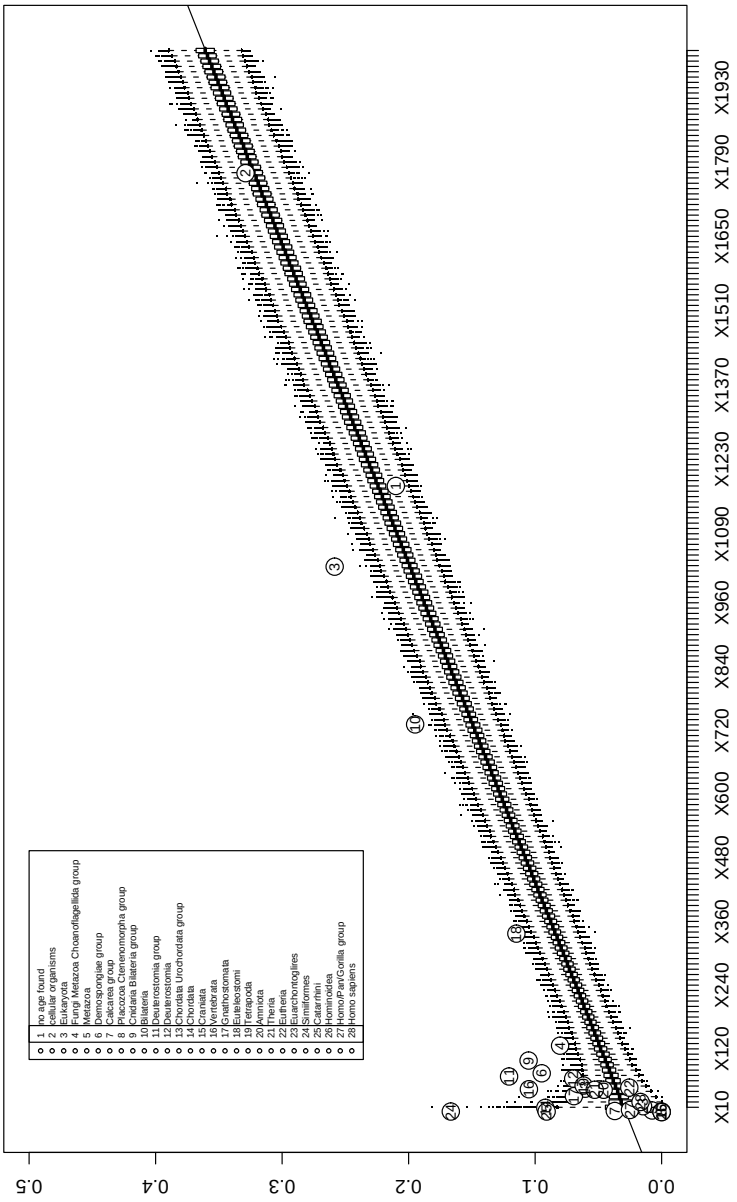


Figure 3.4 10000 rounds per step were used for the random model and the step size is 10. This was plotted as a box plot with a line of best fit and the values for the taxon specific interactions. The numbers in the circles correspond to the taxa in the legend and the position gives the interaction ratio with proteins arisen at the same taxon.

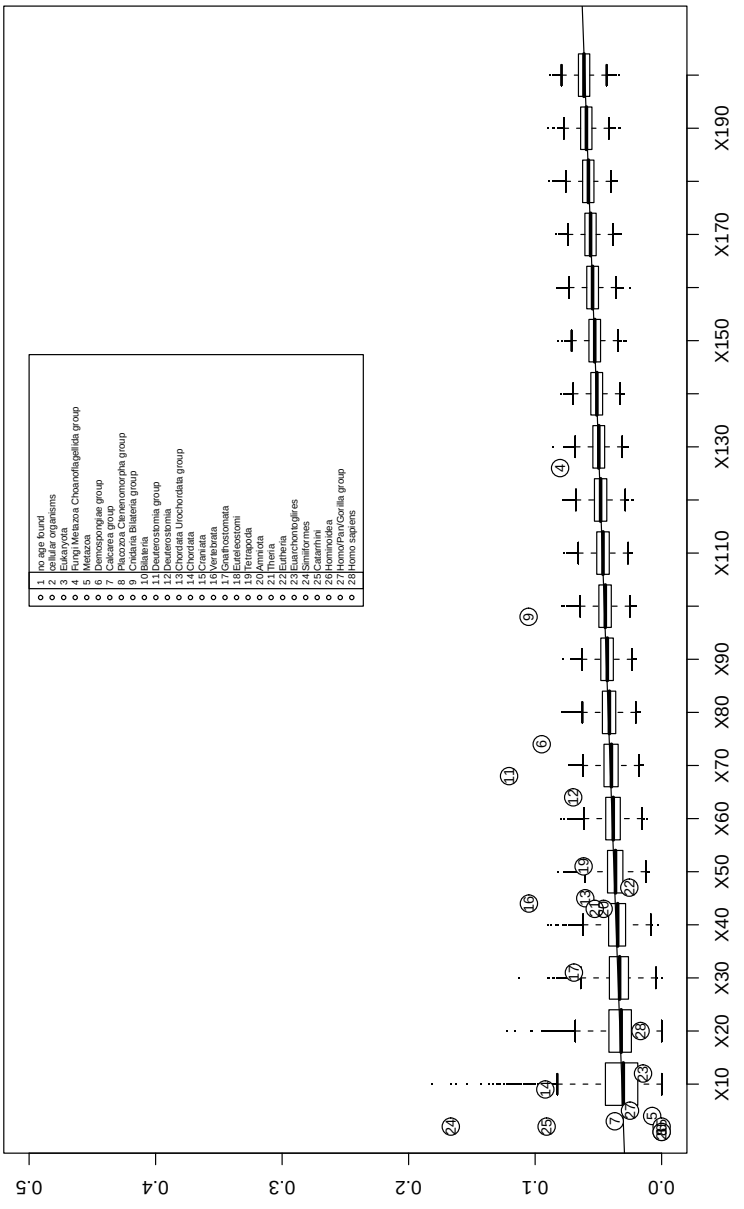


Figure 3.5 A zoomed version of plot 3.4. For description refer to figure 3.4.

3.3.3 Composition of complexes

I analysed 4978 proteins distributed over 1060 complexes. The set contained 2145 unique proteins. The biggest complex contained 29 proteins and the smallest one two. The most active date-hub protein takes place in 54 different complexes. 36 proteins are shared by more than ten complexes. They have 50 interactions per protein on average in a range from zero to 195 (compare table 3.2).
In general, the percentage of proteins based on newly originated domain architectures is the same for complexes and the full human proteome (compare table 3.1). One difference is the number of proteins with unknown age, which is higher in complexes, around one third compared to one quarter.
The proteins which are part of the most complexes are the evolutionary oldest (compare table 3.2). Sixty percent of proteins shared by more than ten complexes have arisen at the time *cellular organisms* or *Eukaryota* have arisen. Those proteins are very well connected in the human interactome, too. Roughly 90 percent of all complexes contain at least one domain architecture arisen at the time *cellular organisms* or *Eukaryota* arose (compare tables 3.3, 3.4, 3.5). Looking at the 5 complexes containing human specific domain architectures we can see that all off them contain at least one protein originating from the beginning of cellular live. The next two taxa at which proteins arose and which are part of complexes are part of 4 and respectively 3 complexes. It is interesting to note that eighty percent of the human specific complexes contain an architecture originating from the onset of euteleostomi.

3.4 Conclusions

The combination of a large protein-protein interaction network with the method to estimated a protein's origin based on its domain architectures made this analysis possible. Knowing the origin of a protein raised the question of the origin of its interaction partner. Are proteins integrated anywhere in the network or preferably origin related? In this analysis, it was shown that proteins arisen together significantly interact together more often. One possible explanation is that these proteins are the base for new sub-parts of the full network connect mainly to some proteins of the whole network. This picture is different from the theory of preferential attachment, which proposes new proteins as preferably attached to established nodes with a high degree. The enrichment of *Bilateria* and the *Bilateria Cnidaria group* in the nodes with the highest degree and the highest betweenness indicates that important hub-bottleneck proteins have

Table 3.1 Percentage of proteins based on in this taxa originated domain architectures human complexes

Taxon	Complex	Human
cellular organisms	26.46	29.22
Eukaryota	21.05	17.18
Fungi/Metazoa group	0	0
Fungi Metazoa Choanoflagellida group	2.29	2.09
Metazoa	0.02	0.07
Demospongiae group	0.56	1.1
Calcarea group	0	0.03
Placozoa Ctenenomorpha group	0	0.04
Cnidaria Bilateria group	1.15	1.43
Bilateria	8.34	10.72
Acoelomorpha	0	0
Deuterostomia group	1.15	0.97
Deuterostomia	0.84	0.98
Chordata Urochordata group	0.68	0.74
Chordata	0.02	0.15
Craniata	0	0.07
Vertebrata	0.02	0.59
Gnathostomata	0.48	0.46
Teleostomi	0	0
Euteleostomi	3.07	5.25
Sarcopterygii	0	0
Tetrapoda	0.66	0.98
Amniota	0.58	0.66
Mammalia	0	0
Theria	0.7	0.79
Eutheria	0.42	0.79
Euarchontoglires	0.04	0.17
Primates	0	0
Haplorrhini	0	0
Simiiformes	0	0.03
Catarrhini	0.08	0.03
Hominoidea	0	0.01
Hominidae	0	0
Homo/Pan/Gorilla group	0.02	0.12
Homo	0	0
Homo sapiens	0.1	0.35
no age	31.26	24.98

Evolutionary from *cellular organisms* to *Homo sapiens* ordered list of taxa with the percentage of proteins based on newly arisen domain architectures in this taxa for complexes and the full ppi dataset.

Table 3.2 Cluster protein statistics

Protein	Distributed over x cluster	Taxon of arising	degree	betweenness	length of domain architecture
NP_391987.1	54	Eukaryota	56	137279	1
NP_000203.2	26	Eukaryota	38	32302	1
NP_002201.1	25	no age	26	20355	7
NP_004955.2	24	cellular organisms	113	194062	1
NP_001789.2	21	cellular organisms	68	103093	1
NP_001017963.1	18	cellular organisms	0	0	-
NP_733779.1	17	Eukaryota	29	44327	2
NP_001518.2	16	no age	0	0	-
NP_055063.1	16	cellular organisms	14	13751	1
NP_005457.2	15	Eukaryota	6	4153	1
NP_001269.3	15	cellular organisms	51	54860	1
NP_002077.1	14	cellular organisms	185	412042	1
NP_001092.1	14	Eukaryota	87	209803	1
NP_004765.2	14	no age	26	18110	-
NP_003583.2	14	Eukaryota	27	18693	1
NP_005336.2	13	cellular organisms	54	120637	1
NP_004220.2	13	Eukaryota	5	80	1
NP_001420.2	13	Bilateria	196	408293	7
NP_002196.2	13	Bilateria	21	13089	7
NP_003191.1	12	Eukaryota	41	56097	1
NP_005601.1	12	cellular organisms	26	8669	1
NP_056292.1	12	Eukaryota	50	47359	4
NP_996896.1	12	cellular organisms	7	1728	1
NP_004259.3	12	Bilateria	6	15815	-
NP_005639.1	12	Eukaryota	8	6223	1
NP_004255.2	12	no age	8	3267	-
NP_000876.3	12	Deuterostomia group Fungi Metazoa	15	9180	6
NP_008996.1	11	Choanoflagellida group	21	8052	3
NP_000928.1	11	cellular organisms	51	77784	4
NP_000537.3	11	no age	0	0	-
NP_001895.1	11	Eukaryota	118	380591	1
NP_003630.1	11	no age	43	36775	-
NP_002065.1	11	cellular organisms	18	20649	1
NP_003063.2	11	Euteleostomi	34	59625	3
NP_003397.1	11	Eukaryota	105	177042	1
NP_068810.2	11	Bilateria	97	164569	2

The 36 top proteins which are part of more than ten different complexes, their taxa of emergence and their number of protein-protein interactions.

Table 3.3 Co Complexised proteins

		sum	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
cellular organisms	1	807	660	336	75	11	28	159	16	13	7	10	66	23	18	20	14	2	4	5
Eukaryota	2	738		587	49	10	20	151	21	16	6	2	60	15	18	16	12	0	2	4
Fungi Metazoa	3	174			94	1	0	23	1	3	0	0	13	1	0	3	2	0	0	3
Choanoflagellida group																				
Demospongiae group	4	35				23	0	7	0	2	1	0	2	1	0	0	0	0	0	0
Cnidaria Bilateria group	5	78					48	17	0	3	0	1	6	0	1	1	1	0	0	0
Bilateria	6	476						315	7	16	8	5	48	13	6	9	6	1	0	0
Deuterostomia group	7	53							39	2	0	0	3	2	1	0	0	0	0	0
Deuterostomia	8	65								36	0	0	7	2	0	1	0	0	0	0
Chordata Urochordata group	9	32									27	0	5	2	2	1	0	0	0	0
Gnathostomata	10	21										20	1	0	0	0	2	0	0	0
Euteleostomi	11	239											138	8	4	3	8	1	0	4
Tetrapoda	12	72												32	1	1	3	0	0	0
Amniota	13	53													29	2	0	0	0	0
Theria	14	59														31	1	0	1	0
Eutheria	15	50															21	0	1	0
Euarchontoglires	16	4																2	0	0
Catarrhini	17	4																	4	0
Homo sapiens	18	16																		5

Counts of complexes in which proteins arisen at taxa x and y are found together. The cross of x=y shows the sum of all complexes containing at least one protein arisen at this taxon.

E.g. Protein arisen in row (y) one *cellular organisms* and column (x) two *Eukaryota* are found together in 336 complexes.

Table 3.4 Co Complexised proteins relative counts part 1

		1	2	3	4	5	6	7	8	9
cellular organisms	1	1/1	0.51/0.57	0.11/0.8	0.02/0.48	0.04/0.58	0.24/0.5	0.02/0.41	0.02/0.36	0.01/0.26
Eukaryota	2		1/1	0.08/0.52	0.02/0.43	0.03/0.42	0.26/0.48	0.04/0.54	0.03/0.44	0.01/0.22
Fungi Metazoa Choanoflagellida group	3			1/1	0.01/0.04	-/ -	0.24/0.07	0.01/0.03	0.03/0.08	-/ -
Demospongiae group	4				1/1	-/ -	0.3/0.02	-/ -	0.09/0.06	0.04/0.04
Cnidaria Bilateria group	5					1/1	0.35/0.05	-/ -	0.06/0.08	-/ -
Bilateria	6						1/1	0.02/0.18	0.05/0.44	0.03/0.3
Deuterostomia group	7							1/1	0.05/0.06	-/ -
Deuterostomia	8								1/1	-/ -
Chordata Urochordata group	9									1/1
Gnathostomata	10									
Euteleostomi	11									
Tetrapoda	12									
Amniota	13									
Theria	14									
Eutheria	15									
Euarchontoglires	16									
Catarrhini	17									
Homo sapiens	18									

Relative counts of complexes in which proteins arisen at taxa x and y are found together. Relative count of: row/column

E.g. Protein arisen in row one *cellular organisms* and column two *Eukaryota* are found together in 0.51 (*660) complexes containing proteins arisen in *cellular organisms* and in 0.57 (*587) complexes containing proteins arisen in *Eukaryota*.

Table 3.5 Co Complexised proteins relative counts part 2

		10	11	12	13	14	15	16	17	18
cellular organisms	1	0.02/0.5	0.1/0.48	0.03/0.72	0.03/0.62	0.03/0.65	0.02/0.67	0/1	0.01/1	0.01/1
Eukaryota	2	0/0.1	0.1/0.43	0.03/0.47	0.03/0.62	0.03/0.52	0.02/0.57	-/ -	0/0.5	0.01/0.8
Fungi Metazoa Choanoflagellida group	3	-/ -	0.14/0.09	0.01/0.03	-/ -	0.03/0.1	0.02/0.1	-/ -	-/ -	0.03/0.6
Demospongiae group	4	-/ -	0.09/0.01	0.04/0.03	-/ -	-/ -	-/ -	-/ -	-/ -	-/ -
Cnidaria Bilateria group	5	0.02/0.05	0.13/0.04	-/ -	0.02/0.03	0.02/0.03	0.02/0.05	-/ -	-/ -	-/ -
Bilateria	6	0.02/0.25	0.15/0.35	0.04/0.41	0.02/0.21	0.03/0.29	0.02/0.29	0/0.5	-/ -	-/ -
Deuterostomia group	7	-/ -	0.08/0.02	0.05/0.06	0.03/0.03	-/ -	-/ -	-/ -	-/ -	-/ -
Deuterostomia	8	-/ -	0.19/0.05	0.06/0.06	-/ -	0.03/0.03	-/ -	-/ -	-/ -	-/ -
Chordata Urochordata group	9	-/ -	0.19/0.04	0.07/0.06	0.07/0.07	0.04/0.03	-/ -	-/ -	-/ -	-/ -
Gnathostomata	10	1/1	0.05/0.01	-/ -	-/ -	-/ -	0.1/0.1	-/ -	-/ -	-/ -
Euteleostomi	11		1/1	0.06/0.25	0.03/0.14	0.02/0.1	0.06/0.38	0.01/0.5	-/ -	0.03/0.8
Tetrapoda	12			1/1	0.03/0.03	0.03/0.03	0.09/0.14	-/ -	-/ -	-/ -
Amniota	13				1/1	0.07/0.06	-/ -	-/ -	-/ -	-/ -
Theria	14					1/1	0.03/0.05	-/ -	0.03/0.25	-/ -
Eutheria	15						1/1	-/ -	0.05/0.25	-/ -
Euarchontoglires	16							1/1	-/ -	-/ -
Catarrhini	17								1/1	-/ -
Homo sapiens	18									1/1

Relative counts of complexes in which proteins arisen at taxa x and y are found together. Relative count of: row=column E.g. Protein arisen in row one cellular organisms and column ten *Gnathostomata* are found together in 0.02 (*660) complexes containing proteins arisen in cellular organisms and in 0.5 (*20) complexes containing proteins arisen in *Eukaryota*.

arisen there.
Nine out of ten complexes contain one protein with a domain architecture arisen at the onset of *cellular organisms* or *Eukaryota*. Complexes seem to have arisen step by step. This conclusion has is limited by the facts that only the origin of the underlying domain architectures is calculated by this method. Another possible restrainment is the focus on a (small?) part of human complexes.

Chapter 4

Protein Interaction Networks - More than Mere Modules

4.1 Abstract

It is widely believed that the modular organization of cellular function is reflected in a modular structure of molecular networks. A common view is that a "module" in a network is a cohesively linked group of nodes, densely connected internally and sparsely interacting with the rest of the network. Many algorithms try to identify functional modules in protein-interaction networks (PIN) by searching for such cohesive groups of proteins.

Here, we present an alternative approach independent of any prior definition of what actually constitutes a "module". In a self-consistent manner, proteins are grouped into "functional roles", if they interact in similar ways with other proteins according to their functional roles. Such grouping may well result in cohesive modules again, but only if the network structure actually supports this.

We applied our method to the PIN from the Human Protein Reference Database and found that a representation of the network in terms of cohesive modules, at least on a global scale, does not optimally represent the network's structure because it focuses on finding independent groups of proteins. In contrast, a decomposition into functional roles is able to depict the structure much better as it also takes into account the interdependencies between roles and even allows groupings based on the absence of interactions between proteins in the

same functional role, as is the case for transmembrane proteins, which could never be recognized as a cohesive group of nodes in a PIN.

When mapping experimental methods onto the groups, we identified profound differences in the coverage suggesting that our method is able to capture experimental bias in the data, too. For example yeast-two-hybrid data were highly overrepresented in one particular group.

Thus, there is more structure in protein-interaction networks than cohesive modules alone and we believe this finding can significantly improve automated function prediction algorithms in the future

Abbreviations: PPI, protein-protein interaction; GO, Gene Ontology; HPRD, Human Protein Reference Database

4.2 Introduction

Biological function is believed to be organized in a modular and hierarchical fashion (Barabási and Oltvai 2004). Genes make proteins, proteins from cells, cells form organs, organs form organisms, organisms form populations and populations form ecosystems. While the higher levels of this hierarchy are well understood, and the genetic code has been deciphered, the unraveling of the inner workings of the proteome poses one of the greatest challenges in the post-genomic era (Sharan et al. 2007). The development of high-throughput experimental techniques for the delineation of protein-protein interactions as well as modern data warehousing technologies to make data available and searchable are key steps towards understanding the architecture and eventually function of the cellular network. These data now allow for searching for functional modules within these networks by computational approaches and for assigning of putative protein functions based on such data.

A recent review by Sharan (Sharan et al. 2007) surveys the current methods of network based prediction methods for protein function. Proteins must interact to function. Hence, we can expect protein function to be encoded in a protein interaction network. The basic underlying assumption of all methods of automated functional annotation is that pair wise interaction is a strong indication for common function.

Sharan differentiate two basic approaches of network based function prediction: “direct methods”, which can be seen as local methods applying a “guilt-by-association” principle (Oliver 2000) to immediate or second neighbours in

the network, and "module assisted" methods which first cluster the network into modules according to some definition and then annotate proteins inside a module based on known annotations of other proteins in the module. So instead of "guilt-by-association", one could speak of "kin-liability". The latter approach to function prediction necessarily needs a concept of what is to be considered a module in a network. Most researchers consider cohesive sets of proteins which are highly connected internally, but only sparsely with the rest of the network (Spirin and Mirny 2003; Cui et al. 2008; Hwang et al. 2006; Palla et al. 2005; **?**; Bu et al. 2003; Dunn et al. 2005; King et al. 2004; Krognan et al. 2006; Pereira-Leal et al. 2004; Przulj et al. 2004). Such methods have yielded considerable success at the level of very small scale modules and in particular protein complexes.

Does the concept of a module as a group of cohesively interacting proteins also extend to larger scales? Some researchers have argued that modularity in this sense is a universal principle such that small cohesive modules combine to form larger cohesive entities in a nested hierarchy (Ravasz et al. 2002; Clauset et al. 2008453). But is this view really adequate to describe the architecture of protein interactions? Recently, Wang and Zhang (Wang and Zhang 2007) even questioned whether cohesive clusters in protein interaction networks do carry biological information at all and suggested a simple network growth model based on gene duplication which would produce the observed structural cohesiveness as "an evolutionary by-product without biological significance". We will not go as far as questioning the content of biological information in the network structure but rather argue against the model of a cohesively linked group of nodes in a network as an adequate proxy for a functional module on all scales of the network.

Consider as first example protein complexes. Indeed, they consist of proteins working together and experimentally isolated together. Only the large scale analysis of protein complexes (Gavin et al. 2006, 2002) revealed that they are more dynamic than previously assumed. Many proteins can be found not only in a single, but in a multitude of complexes. The information of proteins connecting complexes will be lost when searching only for cohesively interacting groups of proteins. As a second example, consider transmembrane proteins, like receptors in signal transduction cascades. They tend to interact with many different cytoplasmic proteins as well as with their extracellular ligands. Still, only rarely do different transmembrane receptors interact with each other. Thus, the functional class of transmembrane receptors will not be identified when looking for cohesive modules.

Here, we asked whether these features, which are not covered by algorithms

searching for cohesive modules, are also present in the overall structure of the cellular network. If this would be the case, methods searching only for cohesive modules would not be able to identify them. We group proteins self-consistently into *functional roles* if they interact in similar ways with other proteins according to their functional roles. Such a role may well be a cohesive module, meaning that proteins in this class predominantly interact with other proteins of this class, but it does not have to. In other words, we do not impose a structure of cohesive modules on the network in our analysis but rather find the structural representation that is best supported by the data. Using the abstraction of a functional role, we generated an 'image graph' of the original network which depicts only the predominant interactions among classes of proteins and thus allowing a bird's eye view of the network.

In the case of protein interaction network studied here, we found sound evidence that cohesive modules on a global scale do not adequately represent the network's global structure. We found groups of proteins acting as intermediates and specifically connecting other groups of proteins. Furthermore, we even identified a group of proteins which was only sparsely connected within itself, but with similar patterns of interaction to other proteins. Thus, approaches searching only for cohesive modules might not be sufficient to represent all characteristics of cellular networks. Furthermore, our findings suggest that hierarchical modularity as nested, cohesively interacting groups of proteins has to be reconsidered as a universal organizing principle.

4.3 Functional Role Decomposition and Image Graphs

In which cases does a clustering of a network into cohesive modules not reflect its original architecture? Consider the toy network in Figure 4.1 a). There are four known types of proteins in this network. Type A may represents some biological process involving five proteins connected to four proteins of type B. These are linked to another biological process C which involves five further proteins which finally are linked to four proteins of type D. Not all nodes of the same type necessarily share the same set of neighbours. Some nodes of the same type do not have any neighbours in common with nodes of their type or have more neighbours in common with nodes of a different type. This shows that in this hypothetical example, direct methods of functional annotations may be limited in their accuracy.

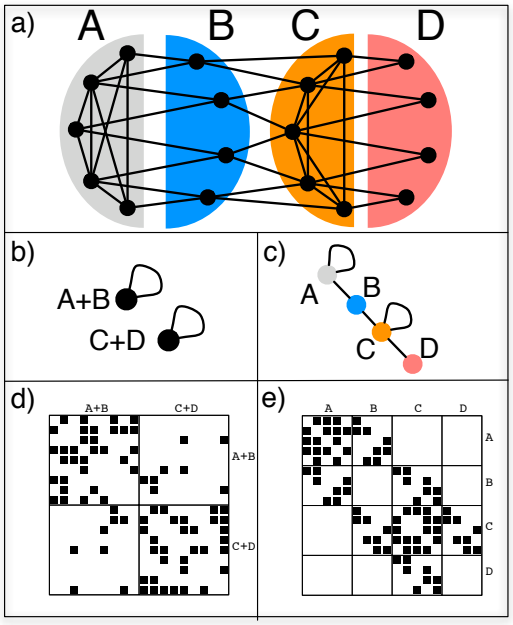


Figure 4.1 **a)** A simple example network of nodes of 4 different types identified by their structural position. Nodes of types A and C are densely connected among themselves. The nodes of type B have connections to both nodes of types A und B, but not among themselves, i.e. they mediate between types A and C. The nodes of type D only have connections to nodes of type C, but not among each other, i.e. they form a periphery to type C nodes. **(b)** and **c)** Two possible image graphs for the functional understanding of this network show the connections among groups of nodes. A typical network clustering will aggregate nodes into clusters densely connected internally but only sparsely connected to the rest, as depicted in the left image graph. This will result in grouping nodes of types A and B together and nodes of type C+D together. Because of aggregating nodes into cohesive groups any such algorithm will never recognize nodes of type C and D as different and hence miss essential part of the networkÕs structure. On the opposite, the right image graph correctly captures the network structure of the 4 different types as the 4 different nodes in the image graph. **d)** and **e)** The adjacency matrices of our example network with rows and columns ordered according to the two decompositions shown above. A black square in position (i, j) indicates the existence of a link connecting node i with node j. Rows and columns are ordered such that nodes in the same group are adjacent. The internal order of the nodes in the groups is random. Each block in the matrix corresponds to a possible edge in the image graph. The left matrix shows the adjacency matrix for the output of a typical clustering algorithm which groups nodes of type A and B, as well as C and D together. Clearly we see dense blocks along the diagonal and sparse blocks on the off-diagonal of the matrix as expected. The right matrix depicts the adjacency matrix with rows and columns according to the actual types of the nodes. All empty blocks in this matrix correspond to a missing edge in the image graph and all populated blocks are represented by an edge in the image graph. We see that for this network, the image graph perfectly captures the structure of the network.

Clustering the network into cohesive modules cannot capture the full structure of the network. The nodes of type B will never be recognized as a proper cluster, because they are not connected internally at all. An attempt to identify such groups was made by Guimera who quantified the error to such a cohesive clustering approach in a "participation coefficient" which is then used to differentiate groups of proteins by this participation coefficient. (GuimerÃ**?** and Nunes Amaral 2005).

The structure of the example network can, however, be perfectly captured by a simple image graph with 4 nodes (Fig. 4.1 c). The nodes in an image graph correspond to the types of nodes in the network. Nodes of type A are connected to other nodes of type A and to nodes of type B. Nodes of type B have connections to nodes of types A and C and so forth. The concept of defining types of nodes by their relation to other types of nodes is known as "regular equivalence" in the social sciences (White and Reitz 1983; Lorrain and White 1971). Structure recognition in networks can then be seen as finding the best fitting image graph for a network. In this context, clustering into functional modules means representing the network by an image graph consisting of isolated, self-linking nodes. Once an assignment of nodes into classes is obtained, the rows and columns of the incidence matrix can be reordered such that rows and columns corresponding to nodes in the same class are adjacent (Fig. 4.1 d and e). Since the rows and columns are not ordered within a certain class, this leads to a characteristic structure with dense blocks in the adjacency matrix corresponding to the links in the image graph and sparse or zero blocks corresponding to the links absent in the image graph. Structure recognition in networks is therefore also called "block modelling" and together with the concepts of structural and regular equivalence has a long history in the social sciences (Doreian et al. 2005; Wasserman and Faust 1994). In our further discussion, we will denote image graphs that consist only of isolated, self-linked nodes as in Figure 4.1 b), "diagonal image graphs" due to the block structure along the diagonal in the adjacency matrix that they induce. Accordingly, we will call all other image graphs "non-diagonal image graphs".

4.3.1 Calculation

But how do we find the best fitting image graph? The problem amounts essentially to aligning a small graph with q nodes to a large network with N nodes. This involves finding an image graph *and* a mapping τ of the N nodes of the network to the q types of nodes such that the mismatch between network and image graph is minimal. Suppose we were given the $q \times q$ adjacency matrix

B_{rs} of our image graph together with the $N \times N$ adjacency matrix A_{ij} of our network . Let τ be the mapping of the N nodes to the q different types, such that $\tau_i \in \{1, .., q\}$ for all $i \in \{1, .., N\}$. To optimize the mapping τ we minimize the following error function:

$$E(\tau, B) = \frac{1}{M} \sum_{i \neq j}^{N} (A_{ij} - B_{\tau_i \tau_j})(w_{ij} - p_{ij}) \tag{4.1}$$

$$= \underbrace{\frac{1}{M} \sum_{i \neq j}^{N} (w_{ij} - p_{ij}) A_{ij}}_{\mathcal{Q}_{\max} < 1} - \underbrace{\frac{1}{M} \sum_{i \neq j}^{N} (w_{ij} - p_{ij}) B_{\tau_i \tau_j}}_{\mathcal{Q}(\tau, B) \leq \mathcal{Q}_{\max}} . \tag{4.2}$$

in which A_{ij} is the $\{0, 1\}$ adjacency matrix of the network under study. w_{ij} denotes the weight given to an edge between nodes i and j. If an edge is absent in the network, w_{ij} is naturally zero. As before $B_{\tau_i \tau_j}$ is the image graph and p_{ij} is a penalty term discussed below. The normalization constant $M = \sum_{i \neq j} w_{ij}$ is used to bound the error by one. This error function gives a weight proportional to $(w_{ij} - p_{ij})$ to errors made on fitting the edges in the network and a weight of p_{ij} to errors made on fitting the absent edges in the network. The penalty term p_{ij} is chosen such that the total error weight on all edges in the network is equal to the total error weight on all absent edges in the network:

$$\sum_{i \neq j}^{N} A_{ij}(w_{ij} - p_{ij}) = \sum_{i \neq j}^{N} (1 - A_{ij}) p_{ij}. \tag{4.3}$$

This can be easily achieved by setting $p_{ij} = (\sum_{k \neq i} w_{ik} \sum_{l \neq j} w_{lj}) / \sum_{k \neq l} w_{kl}$. The first term of equation (4.2) neither depends on the mapping of nodes to types τ nor on the image graph B_{rs}. It can be interpreted as the maximum value of a quality function $\mathcal{Q}$ measuring the fit of the image graph to the network which would be obtained for a perfect fit, $B_{\tau_i \tau_j} = A_{ij}$ for all (i, j). The second term then corresponds to the quality of the actual fit for the given image graph and mapping. The error is simply the difference between the best and any sub-optimal fit and minimizing E and maximizing $\mathcal{Q}$ are equivalent.

If we assume a diagonal image graph $B_{rs} = \delta_{rs}$ we recover in $\mathcal{Q}$ of equation (4.2) a popular quality function for graph clustering known as Newman modularity (Newman and Girvan 2004; GuimerÃ**?** and Nunes Amaral 2005; Wang and Zhang 2007). We can hence directly compare the fit of different given image graphs to one network by the maximum score $\mathcal{Q}$ than can be obtained by optimizing the mapping τ of nodes in the network to the classes represented as nodes in that image graph. The overall optimal image graph with a given number of nodes q and the optimal assignment τ into the q classes can be found directly

by searching for the assignment τ which maximizes (Reichardt and White 2007; Reichardt 2008)

$$\mathcal{Q}^*(\tau) = \frac{1}{2M} \sum_{r,s}^{q} || \sum_{i \neq j}^{N} (w_{ij} - p_{ij}) \delta_{\tau_i r} \delta_{\tau_j s} ||. \tag{4.4}$$

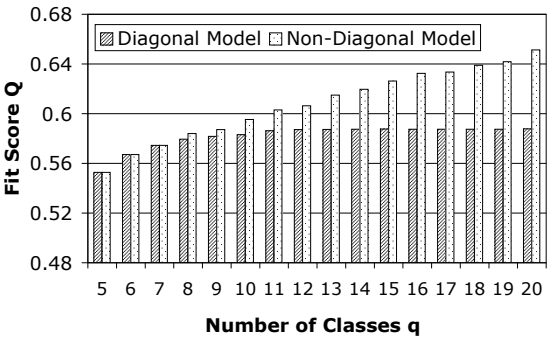
The image graph which allows the highest value of $\mathcal{Q}$ among all possible image graphs with this number of classes can be read off from the assignment τ that maximizes (4.4). It must be such that $B_{rs} = 1$, if the argument in the absolute value in (4.4) is strictly positive, and zero otherwise. One can view B_{rs} as a lossy compression of the original network, in contrast to recently introduced lossless network compression methods for biological analysis (Royer et al. 2008). Since most of the currently available data on protein interaction is noisy and incomplete, we find a lossy compression most adequate for the analysis of the large scale structure of the network.

4.4 Results

4.4.1 Network analysis

Using the quality function introduced above, we analysed the HPRD protein interaction network containing 8,500 nodes. We considered the entire network and optimised $\mathcal{Q}^*$ from (4.4) - thus finding optimal image graphs and assignments of nodes into classes. As expected, with increasing number of classes q, the fit between the actual network and the image graphs becomes better (Fig. 4.2, left panel). Restricting the image graphs to a diagonal form $B_{rs} = \delta_{rs}$ also limited the fit score. The maximum fit score was equal to $\mathcal{Q}_{\max} = 0.98$. Therefore, even with a very small number of classes, already 2/3 of the link structure in the network was captured. The maximum of $\mathcal{Q}$ for a diagonal image graph was reached at $q = 11$ and further addition of classes did not increase this value any more. For $q < 8$ the fit scores for diagonal and non-diagonal image graphs were equal because for less than 8 classes the best image graphs were in fact diagonal. Only beyond this point, the additional degrees of freedom of the non-diagonal image graphs allowed better fit scores.

The question now is, whether these additional degrees of freedom in the image graph actually convey information or only led to over fitting. We therefore divided the $32,331$ links of the network into a test- and a training-set of $1,000$ and $31,331$ links, respectively. Using the optimal image graphs obtained on the full data set and diagonal image graphs for comparison, we optimized $\mathcal{Q}$



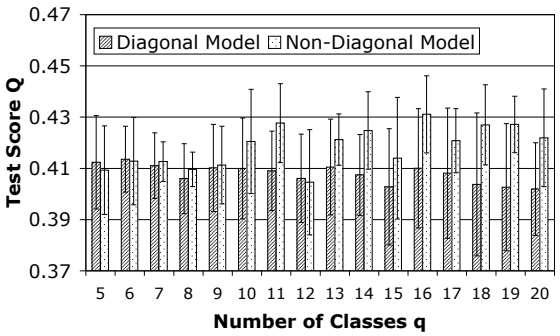


Figure 4.2 Top: Comparison of highest fit scores Q and $\mathcal{Q}^*$ for the full dataset with 32,331 interactions. Clustering methods aggregating nodes into cohesive groups (diagonal image graphs) cannot improve the score beyond a certain limit, while non-diagonal image graphs are able to capture more and more structure as the image graph gets larger and larger. **Bottom:** After removing 1000 links from the data as test-set, we optimized the assignment of nodes into classes according to (4.2) using only the remaining links and keeping the image graphs fixed to those found in the runs that lead to the figure on the left. With the assignment of nodes into classes for this training set of links, we computed the score on the test set of links. The figure shows average and standard deviation over 100 repetitions of this experiment.

from (4.2) on the training-set of links and with the resulting mapping of nodes into classes calculated the fit score $\mathcal{Q}$ on the test-set. The fit score on the training-set of links (data not shown) was close to the full data set. We fixed the non-diagonal image graphs because the comparison is made to diagonal

image graphs which were unaltered, too.

Both, diagonal and non-diagonal image graphs, showed over fitting to some extent. The score on the test set is lower than on the training set (Fig.4.2, right panel). However, for more than 8 classes, the non-diagonal image graphs not only allowed a better fit as discussed, but also scored better on the test-set, the increased fit value also generalized! The non-diagonal image graphs do contain more information about the network than the diagonal image graphs.

It has to be considered that using a test-set containing 3.2% of all links was a drastic disturbance of the system. If we assigned nodes into $q = 8$ equal sized classes, we expect approximately $2/(q(q+1)) \approx 3\%$ of all links in one block. So above this point, the test set we removed was more than the typical number of links in a block. Also, consider the average degree of $\langle k \rangle \approx 8$ interactions per protein in the network. Removing a single link means removing on average $1/8$ of the neighbourhood of the nodes connected by this edge. For the 1,000 edges in the test-set, this could have happened to 2,000 nodes and thus to almost one quarter of all nodes. This explains the large fluctuations and may also explain that for $q = 12$ the non-diagonal image graph cannot outperform the diagonal one.

Figure 4.3 shows two representations of the adjacency matrix of the PIN. On the left hand side, rows and columns are ordered according to the assignment of nodes in classes when fitting a diagonal image graph, when searching for cohesive modules. On the right hand side, the rows and columns are ordered according to the assignment of nodes into classes with the highest scoring non-diagonal image graphs. In both cases we allowed for 11 classes. We have chosen this number of classes because the diagonal models did not achieve larger scores when allowing more classes. The non-diagonal image graphs led to a different assignment of nodes into classes with higher score but further increase of the number of classes did not lead to significant improvement in the generalization error (Fig. 4.2, right panel). Note the similarities and differences in the matrix when ordered after fitting a diagonal image graph and after fitting a non-diagonal image graph.

The non-diagonal models also allowed capturing groups of proteins that mediate between cohesive clusters such as group 2 or that form a cohesive overlap between cohesive clusters, such as groups 4 and 5 or 9 and 10.

4.4.2 Biological interpretation

When comparing the cohesive module to the functional role model (Fig. 4.3) the most distinguishing feature was the existence of connections between sets of

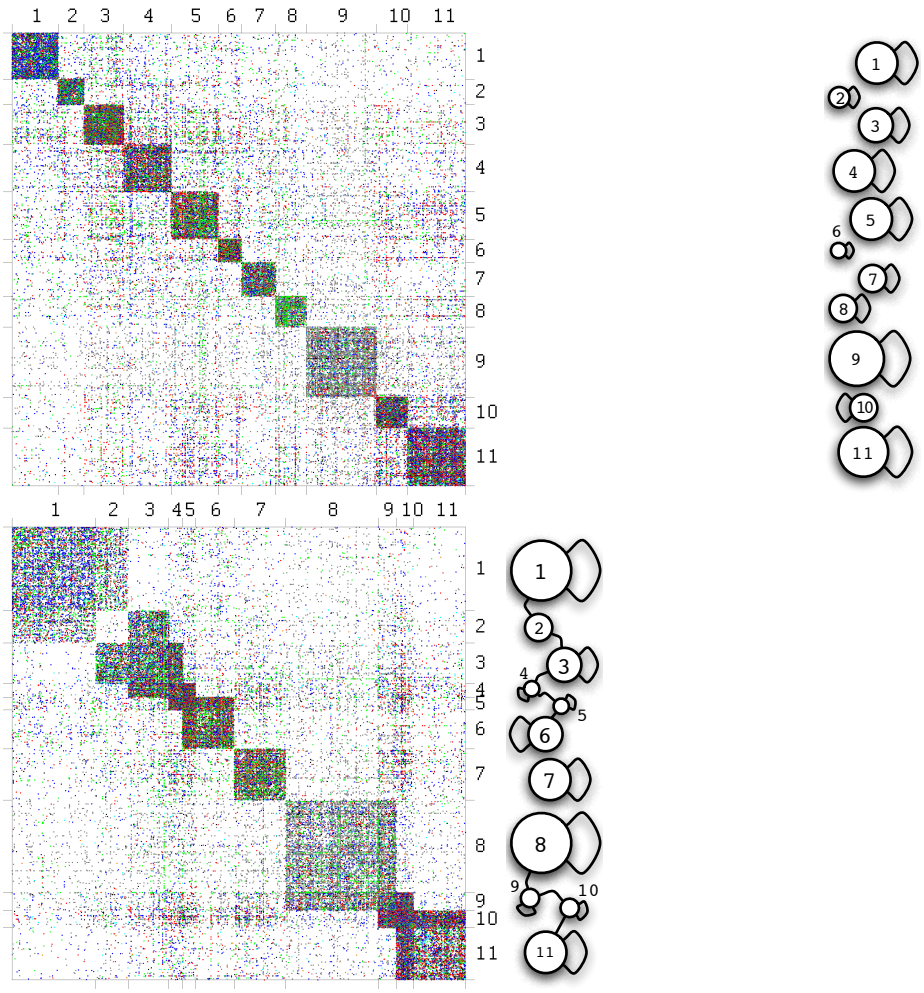


Figure 4.3 For 11 classes, we show the adjacency matrix of the HPRD protein interaction network with rows and column ordered to show diagonal and non-diagonal block structure plus the corresponding image graphs for diagonal block models and non-diagonal block models. Note how the non-diagonal models allow to capture overlap between cohesive blocks but also to detect groups of nodes which are non-cohesive but have similar connection patterns to other classes of proteins. The color of the links codes the experiment type: Y2H: grey, in-vitro: blue, in-vitro+Y2H: turquoise, in-vivo: green, in-vivo+Y2H:orange, in-vivo+in-vitro: red, in-vivo+in-vitro+y2h:black.

proteins in the latter. Groups of proteins existed, which all performed the same "functional role" of connecting two other groups of proteins. Thus, a separation of the cellular network into cohesive modules omits distinct characteristics of the network. In the functional role model, groups were connected to other groups

by a distinct set of additional proteins. These 'connector groups' may well be themselves cohesive, but do not have to. This was illustrated by class 2, where most of the proteins are not interacting with other proteins in the class, but with those of groups 1 and 3.

To evaluate the biological significance of this result, we performed a Gene Ontology enrichment analysis for all clusters. Class 2 was significantly ($E < 10^{-27}$) enriched in proteins annotated as belonging to the membrane and plasma membrane compartment. Indeed, this class contained many transmembrane proteins like for example Cadherin. These proteins typically do not interact with many other transmembrane proteins, but with their extracellular binding partners and, in the case of transmembrane receptors, with cytoplasmic signal transmitters. Indeed we found that group 1, highly interacting with proteins of class 2, mainly consisted of proteins localised in the extracellular region ($E = 2.54E^{-168}$). Furthermore, group 3 also strongly interacting with proteins of class 2, was enriched in proteins associated with the plasma membrane ($E = 2.84E^{-28}$) and involved in signal transduction ($E = 2.72E^{-20}$). Thus, the transmembrane proteins of class 2 are the perfect biological implementation of proteins not interacting with each other, but with proteins of defined other classes (nodes of type B in figure 4.1 a). A complete GO annotation of all clusters of classifications into $q = 5$ to $q = 11$ classes is given in our supporting material at `http://domains.bioapps.biozentrum.uni-wuerzburg.de/ppi`.

In the previous analyses, we considered all data from HPRD, as they are manually curated and therefore of a high quality. To unravel a possible bias between different experimental methods, we plotted the data for three different experimental approaches separately. The ordering of rows and columns, the assignment of proteins into functional roles, was kept from figure 4.3. Instead of plotting all types of interactions on top of each other, the adjacency matrices for interactions which are backed by in-vivo, in-vitro and yeast-two-hybrid (Fields and Song 1989) (Y2H) experiments were shown separatly (Fig. 4.4). The in-vitro and in-vivo data nicely resembled the overall picture while the Y2H data did not follow this pattern. To test how well the overall model described the three experimental methods, we calculated the fit function $\mathcal{Q}$ for each. Here, the assignment of nodes into functional roles was taken from figure 4.3. The fit score for the interactions backed only by Y2H experiments was much lower than the scores of any of the other experimental methods. Thus the Y2H interactions cannot depict the full range of possible protein-protein interactions. Rather, the data based on yeast two hybrid showed a prevalence for class number 8 in figure 4.4. In this cluster nuclear proteins were significantly over-represented ($8.42E^{-10}$). In the Y2H (Ito et al. 2001) assay, the

tested proteins are fused to parts of a transcription factor. Their interaction is measured by the transcription of a reporter gene. Therefore, the proteins have to be within the nucleus. Thus, a bias towards interactions of proteins which naturally reside in the nucleus can be expected in Y2H data.

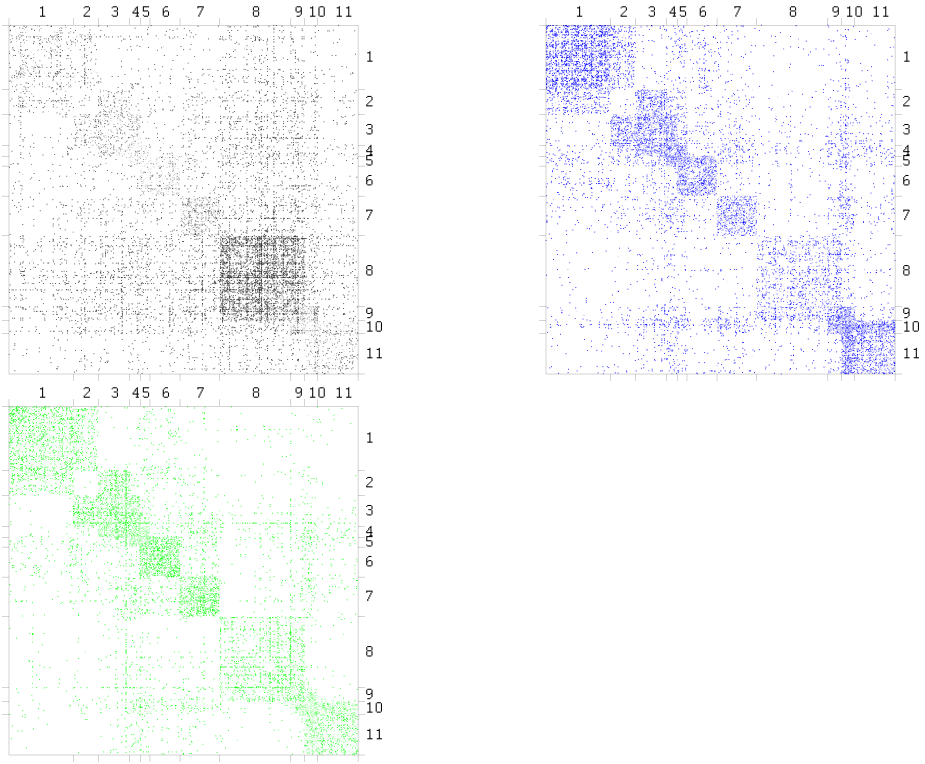


Figure 4.4 The same assignment of nodes into classes as used in Figure 4.3 for but 3 different types of interactions separately. **Left:** Interactions reported only for yeast-2-hybrid experiments (gray). **Right:** Interactions reported only in in-vitro experiments (blue). **Bottom:** Interactions reported only in in-vivo experiments (green). While in-vitro and in-vivo data is highly correlated, the interactions found in Y2H experiments are enriched in class 8.

4.5 Discussion

Using a suited algorithm, any network can be separated into cohesive groups of nodes with more internal than external connections. Accordingly, also protein-protein interaction networks can be divided into comparably independent units as putative functional modules (Spirin and Mirny 2003). Do these modules really reflect a typical characteristic of the cellular network? Here, we used

Table 4.1 Fitscore for different types of interactions.

Experiment type	Image Graph	
	Diagonal	Non-Diagonal
yeast 2-hybrid	0.28	0.30
in vitro	0.53	0.56
in vitro + yeast 2-hybrid	0.51	0.55
in vivo	0.60	0.60
in vivo + yeast 2-hybrid	0.59	0.62
in vivo + in vitro	0.59	0.61
in vivo + in vitro + yeast 2-hybrid	0.64	0.64

Given the assignment of nodes into $q = 11$ classes and the image graphs from figure 4.3, we calculated the fit score Q for each type of interaction separately with equation (4.2). Compare to figure 4.4 which singles out those links which are only supported by Y2H, or only in-vivo or only in-vitro experiments.

an alternative approach for the clustering of protein interactions. We grouped proteins of a similar functional role together. The functional role was defined by the interactions with proteins of other groups. In contrast to cohesive modules, which are more or less independent, groups which specifically linked other groups of proteins could be identified. Thus, an interconnectivity of biological units as in the case of shared components in protein complexes can also be observed at the cellular level. Using a Gene Ontology based classification of all proteins within the modules, we found that these roles are mainly determined by cellular localisation but also function. Although possibly not too surprising to the biologist, this result underlines that the classes we identified by automatic clustering do represent a biological signal.

Using HPRD as data source, a large-scale set of interactions with, on average, eight connections per protein could be analysed. As HPRD contains manually curated data, their quality should be high enough to extend the results to higher coverage. The analysis of interactions derived by different experimental methods revealed a bias in the coverage especially for yeast-two-hybrid data. The great difference of the protein interactions verified only by Y2H to the other methods reminds us to pay attention to the careful weighting of quality and quantity. As large scale binary interaction analysis were mainly based on Y2H, using high coverage data like the one from yeast or *Drosophila melanogaster* might even blur the signal. Another drawback was the small amount of interactions per protein, which is around three to four for the yeast, fly and nematode sets analysed in the study by Wang and Zhang (Wang and Zhang 2007). Still, it would be interesting to compare networks between different organisms to see whether there are changes in the clusters correlated for example with the

emergence of multicellularity. But, reliable results can only be obtained when analysing data sets of comparable quality and size (Reichardt and Leone 2008).

In summary our analysis showed that protein interaction networks are more than sparsely interacting cohesive modules. Rather, groups of proteins are connected by distinct sets of other proteins. These may be highly connected to each other, but do not have to be. Therefore, functional roles and corresponding image graphs might be better descriptors for the characteristics of a protein interaction network than cohesive modules alone. They may help to further improve protein function prediction based on protein-interaction networks.

4.6 Materials and Methods

4.6.1 PPI network.

We used the binary PPI data from the HPRD (Mishra et al. 2006) (Version 6). HPRD protein identifiers and experiment types used to support their connection were extracted. The experiment types were transformed to weights according to table 4.2. The analysis was restricted to the largest connected component containing 32,331 (out of 34,367) interactions of 8,756 proteins (out of 8,919). These interactions do not include data inferred from protein complexes which may introduce errors and bias into the network structure (Wang and Zhang 2007).

Table 4.2 Experiment type to link weight transformation.

Experiment type	Weight	interactions	distinct proteins
yeast 2-hybrid	1	6,580	3,727
in vitro	2	7,872	4,302
in vitro+yeast 2-hybrid	3	1,298	1,523
in vivo	4	6,721	3,826
in vivo+yeast 2-hybrid	5	824	1,119
in vitro+in vivo	6	6,877	3,781
in vitro+in vivo+yeast 2-hybrid	7	2,159	2,201

We valued the different experiments compiled in the HPRD database differently, giving lowest weight to interactions found in yeast-2-hybrid experiments only and highest to those interactions found in vivo, in vitro and Y2H experiments. These weights are only to represent a ranking of a practitioners belief in their validity.

4.6.2 Clustering.

We optimized (4.4) and (4.2) using Simulated Annealing (Kirkpatrick et al. 1983). Details about the implementation can be found in (Reichardt and White 2007) and (Reichardt and Bornholdt 2006), respectively. To obtain the left panel of figure 4.2, for $q = 5$ to $q = 20$ classes, we chose the best of 10 runs, each, for both the fit of a diagonal block model as well as the detection of a non-diagonal block model. The cooling factor for sets with more than ten classes was changed from 0.99 to 0.999 to decrease the false positive rate of local optima. To obtain the right panel of figure 4.2 we randomly divided the original set of links into a test-set of 1000 links and the remaining set was used as a training-set. We used the image graphs, both diagonal and non-diagonal, found in the earlier experiment to optimize the fit score on the training-set. The data shown are the fit scores of the test set, averaged over ten different partitions of the links into training- and test-set.

4.6.3 GO Term enrichment analysis.

The HPRD identifiers and their corresponding GO identifiers were taken from the same HPRD dataset as the PPI network, re-formatted and saved into a file readable by the Ontologizer (Bauer et al. 2008). For the Ontologizer the file gene_ontology.obo created by the GO project (Ashburner et al. 2000b) was be downloaded.

4.6.4 Authors

All three authors wrote and approved the manuscript. Stefan Pinkert did the clustering computations and the biological analysis. Jörg Reichard developed the clustering algorithm and supervised the computations. Jörg Schultz supervised the biological analysis.

4.7 Additional Biological Analysis

4.7.1 Connecting cluster

Cluster are described by proteins and defined by their interactions. As shown in table 4.3 proteins are in most cases found in more then one sub cellular compartment. (Gandhi et al. 2006) showed an significant enrichment of interactions between proteins in the same sub cellular compartment. This leads to higher E values for the most significant terms in the diagonal set (data not shown). The non - diagonal set in contrast has more connection enriched cluster and as pointed out before even an depleted cluster.

Table 4.3 Tags per protein.

Number tags	**Sum of proteins with x tags**		
	Molecular function	Biological process	Cellular compartment
1	2122	1204	4147
2	6298	4628	2044
3	19	141	745
4	63	2515	244
5	1	15	142
0	0	0	1181

We summed up the different tags from hprd for each protein and then counted the number of proteins with x tags.

4.7.2 PDGF pathway

How representative are the cluster? The Platelet Derived Growth Factor (PDGF) Pathway is essential for many processes in cellular proliferation and development and has been used as a model system for the regulation of biological processes by growth factors. This receptor tyrosine kinase class pathway is triggered by an dimeric ligand formed of different combinations of PDGFA, PDGF-B,PDGF-C or PDGF-D (Tallquist and Kazlauskas 2004). Accordingly to those the alpha and beta forms of PDGFR are assembled as a dimer and activated by auto phosphorylation. This leads to the activation of 3 different signalling cascades.

1. The Ras/MAP Kinase Pathway is activated by: P1 of PDGFR-alpha and P3,4,5 of PDGFR-beta.

2. The PI-3 kinase B is phosphorylated at: P2 of PDGFR-beta

3. And PLC,IP3/DAG/PKC is triggered at: P2 of PDGRF-alpha

I adapted in Figure 4.6 the PDGF pathway from *http : //genome.ib.sci.yamaguchi−u.ac.jp/ pnp/* by recoloring it according to the mapping of proteins in the q11 non-diagonal set. In the corresponding matrix plot (figure 4.6) interactions in non enriched clusters are marked with [x] and others as black squares. 75% of all interactions are in enriched clusters which is higher than the overall fit score for the clustering of 61%. As expected when looking at signaling pathways three-fourths of the proteins are part of the biological process *Signal transduction* (compare table 4.3) and they are distributed over all cluster. Another twelve percent are part of the *Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism*. So there has to be another distinguishing feature. The Molecular class is dominated by kinases which are distributed over most of the cluster except the extracellular cluster 1 and the transporter cluster 2 and cluster 8. The transcription factors are localised in three cluster. The difference lies in their primary localisation Cytoplasm for 6 and Nucleous for transcription factors in cluster 10 and 11. The cellular compartment is hard to distinguish because 85% of the proteins are localized in at least two subcellular compartments. Specially PDGF beta is a good example for this problem. This extracellular (Malden et al. 1991) growth factor is alternatively located in Cytoplasm or Nucleus (Sella et al. 1999) this denotes the fact that PDGF beta is produced in megakaryocytes and then released to the blood. So the subcellular localisation is different for different cells. In contrast to some transcription factors i.g. STAT1 which is activated in cytoplasm (Schindler et al. 1992) and then after relocalisation acts in the nucleus (Marg et al. 2004) of the same cell. As shown in table 4.3 79% of the proteins are primary or alternatively in the subcellular compartment cytoplasm in contrast to transcription factors which are exclusively in the nucleus which is the second most compartment. Finally half of the proteins are at least part time connected to the plasma membrane. As expected two third of the proteins are part of the biological process signal transduction and 42 percent of all proteins belong to the molecular class kinase. The clustering algorithm predicts correctly three quarter of all interactions of the pathway. In most of the cases at least one protein in the unpredicted interactions interacts with proteins from more than one other cluster.
Therefore those proteins can not be grouped exactly by the algorithm.

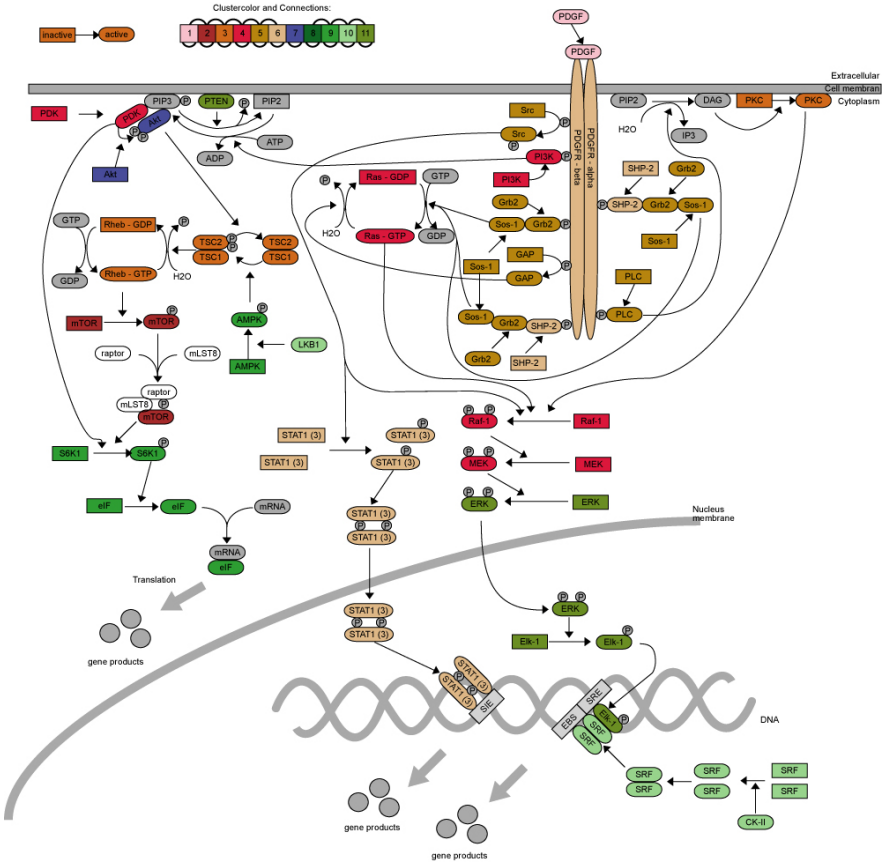


Figure 4.5 This figure is drawn after (Yamaguchi). Proteins are coloured according to the cluster of our q11 non-diagonal set and the small molecules in grey, for mLST8 and raptor was no data available. An interaction is shown as a direct protein-protein contact in the picture or as an arrow which points at an arrow between two states of one protein.

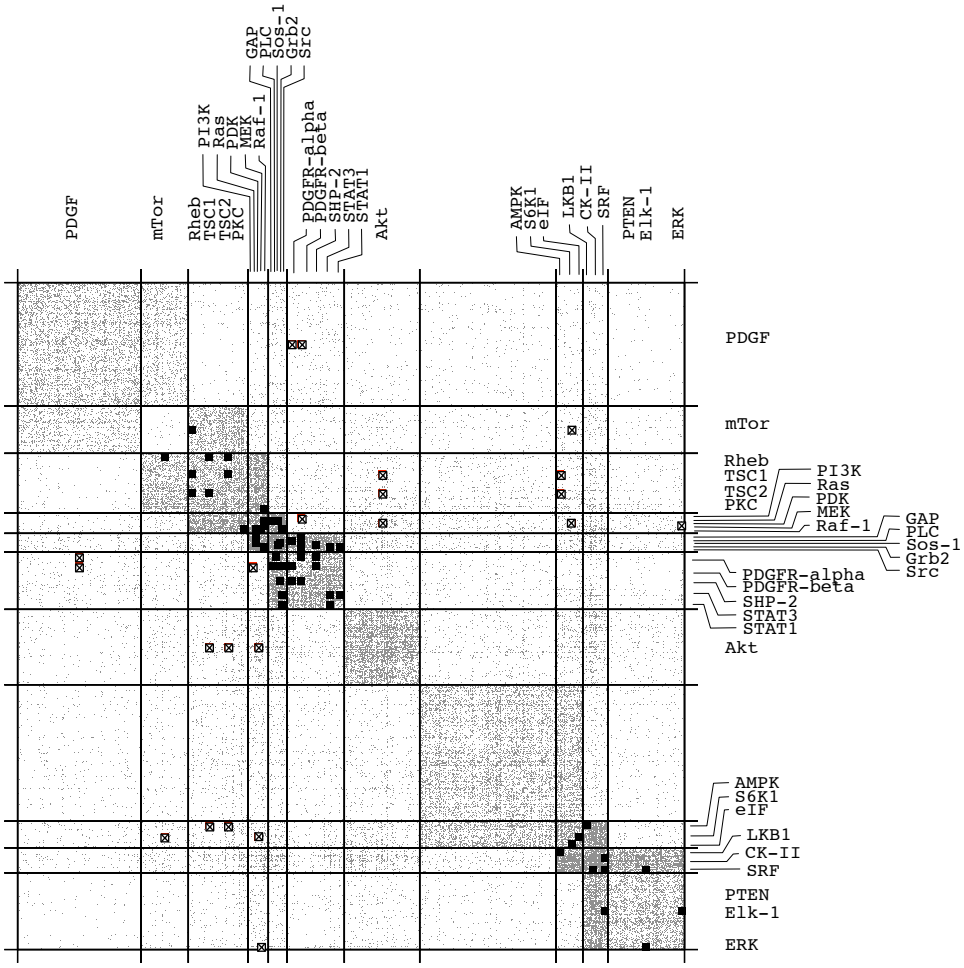


Figure 4.6 Interactions in figure 4.5 are equivalent to the black squares and $[x]$ in this matrix plot.

Chapter 5

Outlook

6465 domains and 32868 domain architectures, this is nearly all it takes to build at least 753166 proteins from 1313 species. Most proteins arisen early in the evolution consist of only one domain. With ongoing evolution proteins originate which have on the on hand more domains and on the other hand more different domains. Leading to the conclusion that new proteins arise by the recombination of domains rather than by simple duplication of already contained domains. While sequencing species it was noticed that some have roughly the same number of proteins. In this work was shown this number does not necessarily reflect the number of underlying domain architectures. Different species which have the same amount of proteins can differ greatly in their number of distinct domain architectures.
The protein-protein-interaction network of human was analysed next. All of its proteins where marked with the taxon of origin for the corresponding domain architecture, if any. Based on the theory that proteins became part of the network when they arise their interactions are analysed. It could be shown that proteins of the same taxon interact more together than expected by a random model. Leading to the assumption that proteins form evolutionary modules. Are these proteins part of one pathway? The KEGG database would be a good starting point to answer this question. The proteins could be mapped to the available pathways for human. Are proteins part of an evolutionary module cluster co expressed or alternatively expressed in different tissues or at different times if they share the same domain architecture. This question could be best answered with microarray experiments for the different tissues. The detection of evolutionary modules in protein protein networks of other species would be very interesting but it at the moment not possible due to the lack of datasets of comparable quality and size.

A different approach to functional modules was followed in the last part. The clustering of many proteins of the human protein-protein-interaction network based on their interactions and non-interactions showed promising results for an model with eleven cluster. Each cluster contained proteins with significant GO terms of the three categories. Even more important proteins of one cluster had few interactions to each other but many to the proteins of two neighbouring cluster. This cluster is significantly enriched for transport proteins, the perfect biological implementation of interacting with others but not with themselves. This is the most distinguishing feature to other clustering algorithms. As a side product it was shown that protein protein interactions detected only with yeast two hybrid methods are bias towards proteins naturally residing in the nucleus. The automated clustering will help the automated analysis of new species and their interaction network. Perhaps can this method be further developed to predict cellular compartment or molecular function for an unknown protein based on its interaction data. This could be a third predictor to sequence and structural inference of protein function.

Chapter 6

Summary

The human genome has been sequenced since 2001. Most proteins have been characterized now and with everyday more bioinformatical predictions are experimentally verified. A project is underway to sequence thousand humans. But still, little is known about the evolution of the human proteome itself. Domains and their combinations are analysed in detail but not all of the human domain architectures at once. Like no one before, we have large datasets of high quality human protein-protein-protein interactions and complexes available which allow us to characterize the human proteome with unmatched accuracy. Advanced clustering algorithms and computing power enable us to gain new information about protein interactions without touching a pipette. In this work, the human proteome is analysed at three different levels. First, the origin of the different types of proteins was analysed based on their domain architectures. The second part focuses on the protein-protein interactions. Finally, in the third part, proteins are clustered based on their interactions and non-interactions.

Most proteins are built of domains and their function is the sum of their domain functions. Proteins that share the same domain architecture, the linear order of domains are homologues and should have originated from one common ancestral protein. This ancestor was calculated for roughly 750 000 proteins from 1313 species. The relations between the species are based on the NCBI Taxonomy and additional molecular data. The resulting data set of 5817 domains and 32868 domain architectures was used to estimate the origin of these proteins based on their architectures. It could be observed, that new domain architectures are only in a small fraction composed of domains arisen at the same taxon. It was also found that domain architectures increase in length and complexity in the course of evolution and that different organisms like worm, fly and human share nearly the same amount of proteins but differ in their number

of distinct domain architectures.
The second part of this thesis focuses on protein-protein interactions. This chapter addresses the question how new evolved proteins form connections within the existing network. The network built of protein-protein interactions was shown to be scale free. Scale free networks, like the internet, consist of few hubs with many connections and many nodes with few connections. They are thought to arise by two mechanisms. First, newly emerged proteins interact with proteins of the network. Second, according to the theory of preferential attachment, new proteins have a higher chance to interact with already interaction rich proteins. The Human Protein Reference Database provides an on in-vivo interaction data based network for human. With the data obtained from chapter one, proteins were marked with their taxon of origin based on their domain architectures. The interaction ratio of proteins of the same taxa compared to all interactions was calculated and higher values than the random model showed for nearly every taxa. On the other hand, there was no enrichment of proteins originated at the taxon of *cellular organisms* for the node degree found. The node degree is the number of links for this node. According to the theorie of preferential attachment the oldest nodes should have the most interactions and newly arisen proteins should be preferably attached to them not together. Both could not be shown in this analysis, preferential attachment could therefore not be the only explanation for the forming of the human protein interaction network.
Finally in part three, proteins and all their interactions in the network are analysed. Protein networks can be divided into smaller highly interacting parts carrying out specific functions. This can be done with high statistical significance but still, it does not reflect the biological significance. Proteins were clustered based on their interactions and non-interactions with other proteins. A version with eleven clusters showed high gene ontology based ratings and clusters related to specific cell parts. One cluster consists of proteins having very few interactions together but many to proteins of two other clusters. This first cluster is significantly enriched with transport proteins and the two others are enriched with extracellular and cytoplasm/membrane located proteins. The algorithm seems therefore well suited to reflect the biological importance behind functional modules.
Although we are still far from understanding the origin of species, this work has significantly contributed to a better understanding of evolution at the protein level and has, in particular, shown the relation of protein domains and protein architectures and their preferences for binding partners within interaction networks.

Chapter 7

Zusammenfassung

Das menschliche Genom ist seit 2001 komplett sequenziert. Ein Großteil der Proteine wurde mittlerweile beschrieben und täglich werden bioinformatische Vorhersagen praktisch bestätigt. Als weiteres Großprojekt wurde kürzlich die Sequenzierung des Genoms von 1000 Menschen gestartet. Trotzdem ist immer noch wenig über die Evolution des gesamten menschlichen Proteoms bekannt. Proteindomänen und ihre Kombinationen sind teilweise sehr detailliert erforscht, aber es wurden noch nicht alle Domänenarchitekturen des Menschen in ihrer Gesamtheit miteinander verglichen. Der verwendete große hochqualitative Datensatz von Protein-Protein-Interaktionen und Komplexen stammt aus dem Jahr 2006 und ermöglicht es erstmals das menschliche Proteom mit einer vorher nicht möglichen Genauigkeit analysieren zu können. Hochentwickelte Cluster Algorithmen und die Verfügbarkeit von großer Rechenkapazität befähigen uns neue Information über Proteinnetzwerke ohne weitere Laborarbeit zu gewinnen.
Die vorliegende Arbeit analysiert das menschliche Proteom auf drei verschiedenen Ebenen. Zuerst wurde der Ursprung von Proteinen basierend auf ihrer Domänenarchitektur analysiert, danach wurden Protein-Protein-Interaktionen untersucht und schließlich erfolgte Einteilung der Proteine nach ihren vorhandenen und fehlenden Interaktionen.
Die meisten bekannten Proteine enthalten mindestens eine Domäne und die Proteinfunktion ergibt sich aus der Summe der Funktionen der einzelnen enthaltenen Domänen. Proteine, die auf der gleichen Domänenarchitektur basieren, das heißt die die gleichen Domänen in derselben Reihenfolge besitzen, sind homolog und daher aus einem gemeinsamen ursprünglichen Protein entstanden. Die Domänenarchitekturen der ursprünglichen Proteine wurden für 750 000 Proteine aus 1313 Spezies bestimmt. Die Gruppierung von Spezies und ihrer Proteine ergibt sich aus taxonomischen Daten von NCBI-Taxonomy, wel-

che mit zusätzlichen Informationen basierend auf molekularen Markern ergänzt wurden. Der resultierende Datensatz, bestehend aus 5817 Domänen und 32868 Domänenarchitekturen, war die Grundlage für die Bestimmung des Ursprungs der Proteine aufgrund ihrer Domänenarchitekturen. Es wurde festgestellt, dass nur ein kleiner Teil der neu evolvierten Domänenarchitekturen eines Taxons gleichzeitig auch im selben Taxon neu entstandene Proteindomänen enthält,. Ein weiteres Ergebnis war, dass Domänenarchitekturen im Verlauf der Evolution länger und komplexer werden, und dass so verschiedene Organismen wie der Fadenwurm, die Fruchtfliege und der Mensch die gleiche Menge an unterschiedlichen Proteinen haben, aber deutliche Unterschiede in der Anzahl ihrer Domänenarchitekturen aufweisen.
Der zweite Teil beschäftigt sich mit der Frage wie neu entstandene Proteine Bindungen mit dem schon bestehenden Proteinnetzwerk eingehen. In früheren Arbeiten wurde gezeigt, dass das Protein-Interaktions-Netzwerk ein skalenfreies Netz ist. Skalenfreie Netze, wie zum Beispiel das Internet, bestehen aus wenigen Knoten mit vielen Interaktionen, genannt Hubs, und andererseits aus vielen Knoten mit wenigen Interaktionen. Man vermutet, dass zwei Mechanismen zur Entstehung solcher Netzwerke führen. Erstens müssen neue Proteine um auch Teil des Proteinnetzwerkes zu werden mit Proteinen interagieren, die bereits Teil des Netzwerkes sind. Zweitens interagieren die neuen Proteine, gemäß der Theorie der bevorzugten Bindung, mit höherer Wahrscheinlichkeit mit solchen Proteinen im Netzwerk, die schon an zahlreichen weiteren Protein-Interaktionen beteiligt sind. Die Human Protein Reference Database stellt ein auf Informationen aus in-vivo Experimenten beruhendes Proteinnetzwerk für menschliche Proteine zur Verfügung. Basierend auf den in Kapitel I gewonnenen Informationen wurden die Proteine mit dem Ursprungstaxon ihrer Domänenarchitekturen versehen. Dadurch wurde gezeigt, dass ein Protein häufiger mit Proteinen, die im selben Taxon entstanden sind, interagiert, als mit Proteinen, die in anderen Taxa neu aufgetreten sind. Es stellte sich heraus das diese Interaktionsraten für alle Taxa deutlich höher waren, als durch das Zufallsmodel vorhergesagt wurden. Alle Taxa enthalten den gleichen Anteil an Proteinen mit vielen Interaktionen. Diese zwei Ergebnisse sprechen dagegen, dass die bevorzugte Bindung der alleinige Mechanismus ist, der zum heutigen Aufbau des menschlichen Proteininteraktion-Netzwerks beigetragen hat.
Im dritten Teil wurden Proteine basierend auf dem Vorhandensein und der Abwesenheit von Interaktionen in Gruppen eingeteilt. Proteinnetzwerke können in kleine hoch vernetzte Teile zerlegt werden, die eine spezifische Funktion ausüben. Diese Gruppen können mit hoher statistischer Signifikanz berechnet werden, haben meistens jedoch keine biologische Relevanz. Mit einem neuen Algorithmus, welcher zusätzlich zu Interaktionen auch Nicht-Interaktionen berücksichtigt,

wurde ein Datensatz bestehend aus 8,756 Proteinen und 32,331 Interaktionen neu unterteilt. Eine Einteilung in elf Gruppen zeigte hohe auf Gene Ontology basierte Werte und die Gruppen konnten signifikant einzelnen Zellteilen zugeordnet werden. Eine Gruppe besteht aus Proteinen, welche wenige Interaktionen miteinander aber viele Interaktionen zu zwei benachbarten Gruppen besitzen. Diese Gruppe enthält eine signifikant erhöhte Anzahl an Transportproteinen und die zwei benachbarten Gruppen haben eine erhöhte Anzahl an einerseits extrazellulären und andererseits im Zytoplasma und an der Membran lokalisierten Proteinen. Der Algorithmus hat damit unter Beweis gestellt das die Ergebnisse nicht bloß statistisch sondern auch biologisch relevant sind.
Wenn wir auch noch weit vom Verständnis des Ursprungs der Spezies entfernt sind, so hat diese Arbeit doch einen Beitrag zum besseren Verständnis der Evolution auf dem Level der Proteine geleistet. Im Speziellen wurden neue Erkenntnisse über die Beziehung von Proteindomänen und Domänenarchitekturen, sowie ihre Präferenzen für Interaktionspartner im Interaktionsnetzwerk gewonnen.

Bibliography

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990. 3

M. A. Andrade, C. P. Ponting, T. J. Gibson, and P. Bork. Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol*, 298(3):521–37, May 2000. 27

G. Apic, J. Gough, and S. A. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 310(2):311–25, Jul 2001. 24

I.I. Artamonova, G. Frishman, M.S. Gelfand, and D. Frishman. Mining sequence annotation databanks for association patterns. *Bioinformatics*, 21 Suppl 3: 49–57, Nov 2005. 6

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000a. 5

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000b. 82

A.-L. Barabási and Z.N. Oltvai. Network biology: Understanding the cells's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004. 68

A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, Oct 1999. 46

A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Res*, 32 Database issue:D138–41, Jan 2004. 4, 23

S. Bauer, S. Grossmann, M. Vingron, and P.N. Robinson. Ontologizer 2.0 - A Multifunctional Tool for GO Term Enrichment Analysis and Data Exploration. *Bioinformatics*, May 2008. 11, 82

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, Jan 2000. 5

E. et Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigó, T.R. Gingeras, E.H. Margulies, Z. Weng, and M... Snyder. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447:799–816, Jun 2007. 15

P. Bork. Shuffled domains in extracellular proteins. *FEBS Lett.*, 286:47–54, Jul 1991. 12

P. Bork, J. Schultz, and C. P. Ponting. Cytoplasmic signalling domains: the next generation. *Trends Biochem Sci*, 22(8):296–298, Aug 1997. 12

E. Bornberg-Bauer, F. Beaussart, S.K. Kummerfeld, S.A. Teichmann, and J. Weiner. The evolution of domain arrangements in proteins and interaction networks. *Cell. Mol. Life Sci.*, 62:435–445, Feb 2005. 12

D. Bu, Z. Zhao, L. Cai, H. Xue, H. Lu, J. Zhang, S. Sun andn L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res.*, 31:2443–2450, 2003. 69

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, Apr 1997. 15

M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, Jun 1996. 15

C. Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 357:543–544, Jun 1992. 12

A. Clauset, C. Moor, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, pages 98–101, 2008453. 69

1000 Genomes Consortium. 1000 genomes, a deep catalog of human genetic variation. URL `/www.1000genomes.org`. 15

The International HapMap Consortium. The International HapMap Project. *Nature*, 426:789–796, Dec 2003. 15

R. R. Copley, C. P. Ponting, J. Schultz, and P. Bork. Sequence analysis of multidomain proteins: past perspectives and future directions. *Adv Protein Chem*, 61:75–98, 2002. 12, 23

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 2006. URL `http://necsi.org/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf`. 51

A.L. Cuff, I. Sillitoe, T. Lewis, O.C. Redfern, R. Garratt, J. Thornton, and C.A. Orengo. The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, Nov 2008. 5

G. Cui, Y. Chen, D.S. Huang, and K. Han. An algorithm for finding functional modules and protein complexes in protein-protein interaction networks. *J. Biomed. Biotechnol.*, 2008:860270, 2008. 69

S. Das and T.F. Smith. Identifying nature's protein Lego set. *Adv. Protein Chem.*, 54:159–183, 2000. 16

R. Dawkins. The Ancestor's Tale. *Orion Books Ltd*, 2004. 2

M.O. Dayhoff, R. Schwartz, and B.C. Orcutt. A model of Evolutionary Change in Proteins. *Atlas of protein sequence and structure*, 5:345–358, 1978. 2

F. de la Cruz and J. Davies. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.*, 8:128–133, Mar 2000. 3, 42

T. Dobzansky. BIOLOGY, MOLECULAR AND ORGANISMIC. *Am. Zool.*, 4: 443–452, Nov 1964. 2

P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, New York, NY, USA, 2005. 72

R. Dunn, F. Dudbridge, and C. M. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6:39, 2005. 69

S.R. Eddy. Hidden Markov models. *Curr. Opin. Struct. Biol.*, 6:361–365, Jun 1996. 23

A.J. Enright, S. Van Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30:1575–1584, Apr 2002. 24

ENSEMBL. Embl - ebi and the sanger institute web site. URL `http://www.ensembl.org/`. 15

S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, Jul 1989. 78

R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res.*, 34:D247–251, Jan 2006. 27

W.M. Fitch. Homology a personal view on some of the problems. *Trends Genet.*, 16:227–231, May 2000. 3, 10

K.A. Frazer, D.G. Ballinger, D.R. Cox, D.A. Hinds, L.L. Stuve, R.A. Gibbs, J.W. Belmont, A. Boudreau, P. Hardenbol, S.M. Leal, S. Pasternak, D.A. Wheeler, T.D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S.B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R.C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M.M. Waye, S.K. Tsui, H. Xue, J.T. Wong, L.M. Galver, J.B. Fan, K. Gunderson, S.S. Murray, A.R. Oliphant, M.S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.F. Olivier, M.S. Phillips, S. Roumy, C. Sallée, A. Verner, T.J. Hudson, P.Y. Kwok, D. Cai, D.C. Koboldt, R.D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.C. Tsui, W. Mak, Y.Q. Song, P.K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C.P. Bird, M. Delgado, E.T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B.E. Stranger, P. Whittaker, D.R. Bentley, M.J. Daly, P.I. de Bakker, J. Barrett, Y.R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D.J. Richter, P. Sabeti, R. Saxena, S.F. Schaffner, P.C. Sham, P. Varilly, D. Altshuler, L.D. Stein, L. Krishnan, A.V. Smith, M.K. Tello-Ruiz, G.A. Thorisson, A. Chakravarti, P.E. Chen, D.J. Cutler, C.S. Kashuk, S. Lin, G.R. Abecasis, W. Guan, Y. Li, H.M. Munro, Z.S. Qin, D.J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy,

C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L.R. Cardon, G. Clarke, D.M. Evans, A.P. Morris, B.S. Weir, T. Tsunoda, J.C. Mullikin, S.T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D.R. Macer, E. Suda, C.N. Rotimi, C.A. Adebamowo, I. Ajayi, T. Aniagwu, P.A. Marshall, C. Nkwodimmah, C.D. Royal, M.F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I.F. Adewole, B.M. Knoppers, M.W. Foster, E.W. Clayton, J. Watkin, R.A. Gibbs, J.W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G.M. Weinstock, D.A. Wheeler, I. Yakub, S.B. Gabriel, R.C. Onofrio, D.J. Richter, L. Ziaugra, B.W. Birren, M.J. Daly, D. Altshuler, R.K. Wilson, L.L. Fulton, J. Rogers, J. Burton, N.P. Carter, C.M. Clee, M. Griffiths, M.C. Jones, K. McLay, R.W. Plumb, M.T. Ross, S.K. Sims, D.L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J.C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A.L. Holden, L.D. Brooks, J.E. McEwen, M.S. Guyer, V.O. Wang, J.L. Peterson, M. Shi, J. Spiegel, L.M. Sung, L.F. Zacharia, F.S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, Oct 2007. 15

T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller. Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics*, 22:2851–2857, Dec 2006. 13

T.K. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K.N. Chandrika, S.S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J.D. Boeke, G. Parmigiani, J. Schultz, J.S. Bader, and A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.*, 38:285–293, Mar 2006. 83

A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. DÃ$\frac{1}{4}$mpelfeld, A. Edelmann, M.A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J.M. Rick, B. Kuster, P. Bork, R.B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, Mar 2006. 69

A.C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse,

C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, Jan 2002. 69

J. Gough. Convergent evolution of domain architectures (is rare). *Bioinformatics*, 21:1464–1471, Apr 2005. 9, 42

R. GuimerÃ? and L.A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, Feb 2005. 72, 73

K. Halanych. The new view of animal phylogeny. *Annual Review of Ecology, Evolution, and Systematics*, 35:229–256, 2004. 1, 25, 29, 32

S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919, Nov 1992. 2

H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, 32: D452–455, Jan 2004. 17

K. Hokamp, A. McLysaght, and K.H. Wolfe. The 2R hypothesis and the human genome sequence. *J. Struct. Funct. Genomics*, 3:95–110, 2003. 6

L. Holm and C. Sander. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, 22:3600–3609, Sep 1994. 5

J.M. Howell, T.L. Winstone, J.R. Coorssen, and R.J. Turner. An evaluation of in vitro protein-protein interaction techniques: assessing contaminating background proteins. *Proteomics*, 6:2050–2069, Apr 2006. 17

N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B.A. Cuche, E. de Castro, C. Lachaize, P.S. Langendijk-Genevaux, and C.J. Sigrist. The 20 years of PROSITE. *Nucleic Acids Res.*, 36:D245–249, Jan 2008. 4

W. Hwang, Y.R. Cho, A. Zhang, and M. Ramanathan. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol Biol*, 1:24, 2006. 69

R Ihaka and R Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996. 51

T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.*, 98:4569–4574, Apr 2001. 78

M. Itoh, J.C. Nacher, K. Kuma, S. Goto, and M. Kanehisa. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol.*, 8:R121, 2007. 24

F. JACOB and J. MONOD. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, Jun 1961. 7

C.J. Jeffery. Multifunctional proteins: examples of gene sharing. *Ann. Med.*, 35:28–35, 2003. 16

R.A. Jensen. Orthologs and paralogs - we need to get it right. *Genome Biol.*, 2:INTERACTIONS1002, 2001. 3

H. Jeong, S.P. Mason, A.L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, May 2001. 46

A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3012, 2004. 69

S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi. Optimization by simmulated annealing. *Science*, 220:671–680, 1983. 82

N. J. Krognan, G. Cagney, H. Yu, G. Zhong, X. Guo, and A. Ignatchenko. Global landscape of protein complexes in yeast saccharomyces cerevisiae. *Nature*, 440: 637–643, 2006. 69

V. Kunin, I. Cases, A.J. Enright, V. de Lorenzo, and C.A. Ouzounis. Myriads of protein families, and still counting. *Genome Biol.*, 4:401, 2003. 24

N.C. Kyrpides. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15:773–774, Sep 1999. 15

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French,

D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, J. Szustakowki, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, and Y. J. Chen. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001. 15, 24

I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz,

C. P. Ponting, and P. Bork. Smart 4.0: towards genomic data integration. *Nucleic Acids Res*, 32 Database issue:D142–4, Jan 2004. 23

Ivica Letunic, Richard R Copley, Birgit Pils, Stefan Pinkert, Jorg Schultz, and Peer Bork. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, 34(Database issue):257–260, Jan 2006. 25, 46

E.D. Levy, J.B. Pereira-Leal, C. Chothia, and S.A. Teichmann. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, 2:e155, Nov 2006. 18

K. Lin, L. Zhu, and D.Y. Zhang. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, 22:2081–2086, Sep 2006. 24

F. Lorrain and H.C. White. Structural equivalence of individuals in social networks. *J. Math. Sociol.*, 1:49–80, 1971. 72

L.T. Malden, A. Chait, E.W. Raines, and R. Ross. The influence of oxidatively modified low density lipoproteins on expression of platelet-derived growth factor by human monocyte-derived macrophages. *J. Biol. Chem.*, 266:13901–13907, Jul 1991. 84

A. Marg, Y. Shan, T. Meyer, T. Meissner, M. Brandenburg, and U. Vinkemeier. Nucleocytoplasmic shuttling by nucleoporins Nup153 and Nup214 and CRM1-dependent nuclear export control the subcellular distribution of latent Stat1. *J. Cell Biol.*, 165:823–833, Jun 2004. 84

S. Mathivanan, B. Periaswamy, T.K. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y.L. Ramachandra, and A. Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7 Suppl 5:S19, 2006. 17

S. McGinnis and T.L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32:W20–25, Jul 2004. 3

MD) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore and MD) National Center for Biotechnology Information, National Library of Medicine (Bethesda. Online mendelian inheritance in man, omim (tm). URL `http://www.ncbi.nlm.nih.gov/omim/`. 17

G.R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T.M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K.S. Arun, S. Sharma, K.N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak,

G. Vishnupriya, H.G. Kumar, M. Nagini, G.S. Kumar, R. Jose, P. Deepthi, S.S. Mohan, T.K. Gandhi, H.C. Harsha, K.S. Deshpande, M. Sarker, T.S. Prasad, and A. Pandey. Human protein reference database–2006 update. *Nucleic Acids Res.*, 34:D411–414, Jan 2006. 17, 81

R. Mott, J. Schultz, P. Bork, and C. P. Ponting. Predicting protein cellular localization using a domain projection method. *Genome Res*, 12(8):1168–74, Aug 2002. 24

N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. Sigrist, R. Vaughan, and E. M. Zdobnov. The interpro database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, 31(1):315–318, Jan 2003. 23

A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, Apr 1995. 5

NCBI-Taxonomy. National center for biotechnology information taxonomy web site. URL `http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=Search&DB=taxonomy`. 25

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004. 73

H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27:29–34, Jan 1999. 6

Stephen Oliver. Guilt-by-association goes global. *Nature*, 403:601–603, 2000. 68

C.A. Orengo, D.T. Jones, and J.M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372:631–634, Dec 1994. 12

C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, Aug 1997. 5

C.A. Ouzounis, R.M. Coulson, A.J. Enright, V. Kunin, and J.B. Pereira-Leal. Classification schemes for protein structure and function. *Nat. Rev. Genet.*, 4:508–519, Jul 2003. 4

L.R. Pal and C. Guda. Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. *BMC Evol. Biol.*, 6:91, 2006. 12, 24

G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435: 814, 2005. 69

J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54:49–57, 2004. 69

J.B. Pereira-Leal and S.A. Teichmann. Novel specificities emerge by stepwise duplication of functional modules. *Genome Res.*, 15:552–559, Apr 2005. 53

S. Pinkert. Die Evolution der Domänenarchitekturen ist verbunden mit physiologischen Erfindungen,diploma thesis. *University of Wuerzburg*, Jul 2004. 34

C. P. Ponting, R. Mott, P. Bork, and R. R. Copley. Novel protein domains and repeats in drosophila melanogaster: insights into structure, function, and evolution. *Genome Res*, 11(12):1996–2008, Dec 2001. 25

C.P. Ponting and R.B. Russell. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J. Mol. Biol.*, 302:1041–1047, Oct 2000. 13, 14

N. Przulj, D. A. Wiggle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20:340–348, 2004. 69

E. Ravasz, A. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551, August 2002. 69

J. Reichardt. *Structure in Networks*, volume 766 of *Lecture Notes in Physics*. Springer-Verlag Berlin Heidelberg, 2008. 74

J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, 2006. 82

J. Reichardt and M. Leone. (Un)detectable cluster structure in sparse networks. *Phys. Rev. Lett.*, 101:078701, 2008. 81

J. Reichardt and D. R. White. Role models for complex networks. *Eur. Phys. J. B*, 60:217–224, 2007. 74, 82

J.S. Richardson. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, 34:167–339, 1981. 12

Loic Royer, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder. Unraveleing protein networks with power graph analysis. *PLoS Comput. Biol.*, 4(7):e1000108, 2008. 74

C. Schindler, K. Shuai, V.R. Prezioso, and J.E. Darnell. Interferon-dependent tyrosine phosphorylation of a latent cytoplasmic transcription factor. *Science*, 257:809–813, Aug 1992. 84

J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork. Smart: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res*, 28(1):231–234, Jan 2000. 12

J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. Smart, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, 95(11):5857–64, May 1998. 4, 12

O. Sella, G. Gerlitz, S.Y. Le, and O. Elroy-Stein. Differentiation-induced internal translation of c-sis mRNA: analysis of the cis elements and their differentiation-linked binding to the hnRNP C protein. *Mol. Cell. Biol.*, 19: 5429–5440, Aug 1999. 84

Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3:88, 2007. 68

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

B. Snel, G. Lehmann, P. Bork, and M.A. Huynen. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, 28:3442–3444, Sep 2000. 18

E.L. Sonnhammer, S.R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28:405–420, Jul 1997. 12

E.L. Sonnhammer and E.V. Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, 18:619–620, Dec 2002. 16, 21

V. Spirin and L.A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.*, 100:12123–12128, Oct 2003. 69, 79

M.J. Stanhope, A. Lupas, M.J. Italia, K.K. Koretke, C. Volker, and J.R. Brown. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 411:940–944, Jun 2001. 3, 42

L.D. Stein, Z. Bao, D. Blasiar, T. Blumenthal, M.R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D.H. Fitch, L.A. Fulton, R.E. Fulton, S. Griffiths-Jones, T.W. Harris, L.W. Hillier, R. Kamath, P.E. Kuwabara, E.R. Mardis, M.A. Marra, T.L. Miner, P. Minx, J.C. Mullikin, R.W. Plumb, J. Rogers, J.E. Schein, M. Sohrmann, J. Spieth, J.E. Stajich, C. Wei, D. Willey, R.K. Wilson, R. Durbin, and R.H. Waterston. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol.*, 1:E45, Nov 2003. 16, 24

M. Tallquist and A. Kazlauskas. PDGF signaling in cells and mice. *Cytokine Growth Factor Rev.*, 15:205–213, Aug 2004. 83

R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003. 5

J.M. Thornton, A.E. Todd, D. Milburn, N. Borkakoti, and C.A. Orengo. From structure to function: approaches and limitations. *Nat. Struct. Biol.*, 7 Suppl: 991–994, Nov 2000. 17

A. Valencia. Automatic annotation of protein function. *Curr. Opin. Struct. Biol.*, 15:267–274, Jun 2005. 16

W. Van Criekinge and R. Beyaert. Yeast Two-Hybrid: State of the Art. *Biol Proced Online*, 2:1–38, Oct 1999. 17

J. Vasilescu, X. Guo, and J. Kast. Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry. *Proteomics*, 4:3845–3854, Dec 2004. 17

J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian,

W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigó, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, Feb 2001. 15

C. Vogel, C. Berzuini, M. Bashton, J. Gough, and S.A. Teichmann. Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.*, 336:809–823, Feb 2004. 24

A. Wagner and D.A. Fell. The small world inside large metabolic networks. *Proc. Biol. Sci.*, 268:1803–1810, Sep 2001. 46

Z. Wang and J. Zhang. In search of the biological significance of modular

structures in protein networks. *PLoS Comput. Biol.*, 3:e107, Jun 2007. 69, 73, 80, 81

S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994. 72

D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 28(1):10–14, Jan 2000. 32

David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Wolfgang He lmberg, Yuri Kapustin, David L Kenton, Oleg Khovayko, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Kim D Pruitt, Gregory D Schuler, Lynn M Schriml, Stephen T Sequeira, Edwin a nd Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tugba O Suzek, Roman Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 34 (Database issue):173–180, Jan 2006. 29

D.R. White and K.P. Reitz. Graph and semigroup homomorphisms. *Soc. Networks*, 5:193–234, 1983. 72

C.R. Woese and G.E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.*, 74:5088–5090, Nov 1977. 1

S. Wuchty. Scale-free behavior in protein domain networks. *Mol Biol Evol*, 18 (9):1694–702, Sep 2001. 24

S Wuchty, E Ravasz, and AL. Barabási. The Architecture of Biological Networks. *Topics in Biomedical Engineering International Book Series*, Complex Systems Science in Biomedicine:165–181, 2006. 47

Yamaguchi. Pdgf pathway. URL `http://genome.ib.sci.yamaguchi-u.ac.jp/~pnp/frame_petri_net_pathway_pdgf.html`. 85

H. Yu, P.M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, 3:e59, Apr 2007. 48, 49

A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS Lett.*, 513:135–140, Feb 2002. 17

Chapter 8

Appendix

8.1 Curriculum vitae

8.1.1 Publications

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, **Pinkert S**, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A.
Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.
Nat Genet. 2006 Mar;38(3):285-93.
www.ncbi.nlm.nih.gov/entrez/query.fcgi

Letunic I, Copley RR, Pils B, **Pinkert S**, Schultz J, Bork P.
SMART 5: domains in the context of genomes and networks.
Nucleic Acids Res. 2006 Jan 1;34(Database issue):D257-60.
www.ncbi.nlm.nih.gov/entrez/query.fcgi

Blenk, S. and Engelmann, J.C. and **Pinkert, S.** and Weniger, M. and Schultz, J. and Rosenwald, A. and Müller-Hermelink, H.K. and Müller, T. and Dandekar, T.
Explorative data analysis of MCL reveals gene expression networks implicated in survival and prognosis supported by explorative CGH analysis.
BMC Cancer,2008, 8:106

Stefan Pinkert, Jörg Schultz Jörg Reichard
Protein Interaction Networks - More than Mere Modules
PLOS Computational Biology, submitted

8.1.2 Talks

Sep 05 Otto Warburg International Summer School and Workshop on Evolutionary Genomics
Evolutionary Modules in protein networks

8.1.3 Poster

Pinkert S, Schultz J
Evolutionary Modules in Protein Networks
Biology at work, 7th International EMBL PhD Symposium, 2005 Dec 1-3
symposium.predocs.org/symp2005/

Sep 20-22 German Conference on Bioinformatics 2006
http://www.zbit.uni-tuebingen.de/gcb2006/

8.1.4 Advanced Academical Training

Aug 29-08 Otto Warburg International Summer School and Workshop on Evolutionary Genomics
http://ows.molgen.mpg.de/

Dec 01-03 7th International EMBL PhD Student Symposium -Biology at Work-
http://symposium.predocs.org/symp2005/index.php

Jan 22-26 Second European School on Bioinformatics
http://swift.cmbi.kun.nl/euroschool/index.html

8.2 Extra tables

Table 8.1 Pfam clans and their domains

clan	domain
2H	2_5_RNA_ligase, CPDase
4H_Cytokine	CNTF, EPO_TPO, Flt3_lig, GM_CSF, Hormone_1, IFN-gamma, IL10, IL12, IL13, IL2, IL3, IL4, IL5, IL6, Interferon, Leptin, LIF_OSM, PRF, SCF
6PGD_C	3HCDH, 6PGD, IlvC, Mannitol_dh_C, NAD_Gly3P_dh_C, UDPG_MGDP_dh
6_Hairpin	DUF1680, GlcNAc_2-epim, Glyco_hydro_15, Glyco_hydro_48, Glyco_hydro_65m, Glyco_hydro_8, Glyco_hydro_88, Glyco_hydro_9
AAA	AAA, AAA_2, AAA_3, AAA_5, AAA_PrkA, ABC_tran, APS_kinase, Bac_DnaA, DNA_pol3_delta, DUF853, GSPII_E, IstB, KTI12, MCM, Mg_chelatase, NACHT, Rad17, Rep_fac_C, Sigma54_activat, SKI, SMC_N, TraG, UPF0079, Zot
ABC-2	ABC2_membrane, CcmB, DUF990
ABC_membrane	ABC_membrane, ABC_membrane_2, SbmA_BacA
AbrB	MraZ, SpoVT_AbrB
AB_hydrolase	Abhydrolase_1, Abhydrolase_2, Abhydrolase_3, Abhydrolase_4, Acyl_transf_2, AXE1, BAAT_C, Chlorophyllase, COesterase, Cutinase, DLH, DUF1023, DUF1057, DUF1100, DUF1234, DUF1400, DUF1749, DUF452, DUF676, DUF818, DUF900, DUF915, Esterase, FSH1, LACT, LIP, Lipase, Lipase_2, Lipase_3, Ndr, PAF-AH_p_II, Palm_thioest, Peptidase_S10, Peptidase_S15, Peptidase_S28, Peptidase_S37, Peptidase_S9, PGAP1, PHB_depo_C, Tannase, Thioesterase, UPF0227, VirJ
AB_Knot	DUF163, DUF171, DUF358, DUF558, SpoU_methylase, tRNA_m1G_MT
Acetyltrans-like	Acetyltransf_1, Autoind_synth, DUF738, FemAB, FR47, Gly_acyl_tr_C, NMT
ACT	ACT, Thr_dehydrat_C
Actin_ATPase	Acetate_kinase, Actin, BcrAD_BadFG, CmcH_NodU, FGGY_C, FGGY_N, FtsA, GDA1_CD39, Glucokinase, Hexokinase_1, Hexokinase_2, HSP70, Hydant_A_N, MreB_Mbl, Peptidase_M22, Ppx-GppA, ROK, StbA, UPF0075

Acyl-CoA_dh	ACOX, Acyl-CoA_dh_1, Acyl-CoA_dh_2
Acyltransferase	Acyltransferase, DAGAT, Lip_A_acyltrans
ADF	Cofilin_ADF, Gelsolin
Adhesin	Adhesin_Dr, Collagen_bind, Fimbrial, PapG_N
ADP-ribosyl	ART, Binary_toxA, Diphtheria_C, Enterotoxin_a, PARP, Pertussis_S1
AEP	D5_N, DNA_primase_S, Herpes_UL52, Replicase, VirE_N
ALDH-like	Aldedh, DUF1487, Histidinol_dh, LuxC
Alk_phosphatase	Alk_phosphatase, DUF1501, DUF229, Metalloenzyme, PglZ, Phosphodiest, Sulfatase
Amidohydrolase	Amidohydro_1, Amidohydro_2, Amidohydro_3, A_deaminase, Peptidase_M19, PHP, PTE, TatD_DNase
APC	AA_permease, Aa_trans, BenE, Branch_AA_trans, CbiQ, DUF1468, HCO3_cotransp, Na_Ala_symp, Spore_permease, SSF, Sulfate_transp, Transp_cyt_pur, Trp_Tyr_perm, Xan_ur_permease
ARM	Adaptin_N, Arm, HEAT, HEAT_PBS, IBB, IBN_N, IFRD, PC_rep, V-ATPase_H
Arrestin_N-like	Arrestin_N, Spo0M, Vps26
AT14A-like	DUF241, DUF677, DUF793
ATP-grasp	ATP-grasp, ATP-grasp_2, ATP-grasp_3, CPSase_L_D2, Dala_Dala_lig_C, DUF407, GARS_A, GSH-S_ATP, Ins134_P3_kin, RimK, STAS, Synapsin_C, TTL
ATP_synthase	ATP-synt_8, ATP-synt_B, YMF19
Beta-lactamase	Beta-lactamase, Glutaminase, Peptidase_S11, Peptidase_S13, Transpeptidase
Beta_propeller	Arylesterase, Cytochrom_D1, DPPIV_N, FG-GAP, Ldl_recept_b, Me-amine-dh_H, NHL, Nup133_N, PD40, Peptidase_S9_N, PQQ, Reg_prop, SBBP, SGL, Str_synth, WD40
Bet_V_1_like	AHSA1, Bet_v_I, COXG, IP_trans, Polyketide_cyc, Ring_hydroxyl_A, START
bZIP	bZIP_1, bZIP_2, bZIP_Maf
B_Fructosidase	DUF377, Glyco_hydro_32N, Glyco_hydro_43, Glyco_hydro_62, Glyco_hydro_68
C1	C1_1, C1_2, C1_3, C1_4
C1q_TNF	C1q, TNF
C2	C2, PI3K_C2
Cache	Cache_1, Cache_2
Calcineurin	DNA_pol_E_B, Metallophos

Calycin	His_binding, Lipocalin, Lipocalin_2, Nitrophorin, Triabin, VDE
CBD	CBM_2, CBM_3, CHB_HEX, Cohesin
CBM_14_19	CBM_14, CBM_19
CDA	AICARFT_IMPCHas, APOBEC_N, dCMP_cyt_deam_1, dCMP_cyt_deam_2
CH	CH, DUF1042
Chelatase	CbiK, CbiX, Ferrochelatase
Chemoreceptor	Sra, Srb, Sre
Chemosens_recp	7tm_6, 7tm_7, DUF267, Trehalose_recp
Chor_lyase	Chor_lyase, DUF98, UTRA
ClpP_crotonase	ACCA, Carboxyl_trans, CLP_protease, DUF114, ECH, MdcE, Peptidase_S41, Peptidase_S49
CoA-acyltrans	Carn_acyltransf, CAT, Condensation, Transferase, UPF0089
Concanavalin	Gal-bind_lectin, Glyco_hydro_11, Glyco_hydro_12, Glyco_hydro_16, Glyco_hydro_7, Laminin_G_1, Laminin_G_2, Lectin_leg-like, Lectin_legB, Pentaxin, Sialidase
CPA_AT	Asp-Al_Ex, Cons_hypoth698, DUF819, DUF897, Glt_symporter, Mem_trans, Na_H_antiport_1, Na_H_Exchanger, OAD_beta, SBF
CTPT	CTP_transf_1, DUF46
CuAO_N2_N3	Cu_amine_oxidN2, Cu_amine_oxidN3
CUB	CUB, CUB_2
Cupin	3-HAO, AraC_binding, ARD, Asp_Arg_Hydrox, Auxin_BP, CDO_I, Cupin_1, Cupin_2, Cupin_3, Cupin_4, Cupin_5, dTDP_sugar_isom, DUF1498, DUF386, Ectoine_synth, EutQ, FdtA, GPI, HgmA, JmjC, MannoseP_isomer, Pirin, Pirin_C, PMI_typeI
CU_oxidase	Copper-bind, COX2, Cu-oxidase, Cu-oxidase_2, Cu-oxidase_3, Cu_bind_like, Ephrin
Cyclin	CDK5_activator, Cyclin, Cyclin_C, Cyclin_N, RB_A, RB_B, TFIIB
Cystatin	Cathelicidins, Cystatin, Spp-24
Cystine-knot	Coagulin, Cys_knot, Hormone_6, NGF, Noggin, PDGF, TGF_beta
C_Lectin	APT, C4, Intimin_C, Lectin_C, Xlink
DALR	DALR_1, DALR_2
DBL	Duffy_binding, PFEMP
DEAD	DEAD, DEAD_2, Flavi_DEAD, ResIII, SNF2_N, UvrD-helicase

Death	CARD, Death, DED, PAAD_DAPIN
Defensin	BDS_I_II, Defensin_1, Defensin_3, Defensin_beta, Toxin_4
DHQS	DHQ_synthase, Fe-ADH
Di-copper	Hemocyanin_M, Tyrosinase
Dim_A_B_barrel	ABM, AsnC_trans_reg, Dabb, EthD, MIase
DMT	DUF1632, DUF486, DUF6, DUF803, DUF914, Multi_Drug_Res, Nuc_sug_transp, RhaT, Sugar_transport, TPT, UAA, UPF0060
DNA_clamp	DNA_pol3_beta, DNA_pol3_beta_2, DNA_pol3_beta_3, DNA_PPF, Herpes_UL42, PCNA_C, PCNA_N
DNA_ligase	DNA_ligase_aden, DNA_ligase_A_M, mRNA_cap_enzyme
DNA_pol_B-like	DNA_pol_B, DNA_pol_B_2, DUF1744
DNA_primase_lrg	Baculo_LEF-2, DNA_primase_lrg
DoxD-like	DoxD, DoxX, DUF417, SURF4
DPBB	3D, Barwin, Cerato-platanin, DPBB_1, Glyco_hydro_45
DsbD-like	DsbD, NicO
DSRM	dsrm, Ribosomal_S5
dUTPase	Cytomega_UL84, DUF570, dUTPase, Herpes_ORF11, Herpes_U55, Herpes_UL82_83
E-set	Big_1, Big_2, Big_3, Cadherin, Cadherin_2, fn3, He_PIG, HYR, phage_tail_N, PKD, PPC, Y_Y_Y
EDD	Dak1, DegV, EIIA-man
EF_hand	Caleosin, efhand, efhand_Ca_insen, S_100
EGF	EGF, EGF_2, EGF_alliinase, EGF_CA, Laminin_EGF
Endonuclease	Endonuclease_5, UvrC_HhH_N
Enolase_N	Enolase_N, MAAL_N, MR_MLE_N
Enolase_TIM	Enolase_C, MAAL_C, MR_MLE
ENTH_VHS	ANTH, ENTH, VHS
FAD_DHS	CO_dh, DS, ETF_alpha, PNTB, SIR2, TPP_enzyme_M
FAD_Lum_binding	FAD_binding_1, FAD_binding_6, FAD_binding_8, FAD_binding_9, Lum_binding
FAD_oxidored	MTHFR, Pro_dh
FAD_PCMH	FAD_binding_4, FAD_binding_5
FBA	FBA_1, FBA_3
Ferritin	FA_desaturase_2, Ferritin, Mn_catalase, Phenol_Hydrox, Ribonuc_red_sm, Rubrerythrin
Flavokinase	Citrate_ly_lig, CTP_transf_2, DUF795, FAD_syn, Pantoate_ligase
Flavoprotein	Flavodoxin_1, Flavodoxin_2, Flavodoxin_NdrI, FMN_red
FOCS	7tm_2, Dicty_CAR, Frizzled, Ocular_alb

FtsL	DivIC, FtsL
G-protein	AIG1, Arf, ATP_bind_1, DUF258, Dynamin_N, G-alpha, GTP_EFTU, IIGP, Miro, MMR_HSR1, NOG1, Ras, Septin
GAD	GAD, GAD-like
GADPH_aa-bio_dh	Gp_dh_C, Semialdhyde_dhC
GAF	CodY, DUF484, GAF, IclR
Gal_mutarotase	Aldose_epim, Bgal_small_N, Glyco_hydro_38C, Glyco_hydro_65N, Lyase_8
GBD	Allantoicase, APC10, Bac_rhamnosid_N, CBM_15, CBM_17_28, CBM_4_9, CBM_6, Endotoxin_C, Ephrin_lbd, F5_F8_type_C, FBA, Glyco_hydro_2_N, Laminin_N, Muskelin_N, PA-IL, PepX_C, P_proprotein, Sad1_UNC, XRCC1_N
GDE	Bac_rhamnosid, GDE_C, Invertase_neut, Trehalase
GFP	G2F, GFP
GH_CE	Glyco_hydro_38, Glyco_hydro_57, Polysacc_deac_1
GlnB-like	CutA1, DUF190, HisG_C, P-II
Globin	Bac_globin, Globin, Phycobilisome
Glutaminase_I	DJ-1_PfpI, GATase, GATase_3, Glyco_hydro_42M, Peptidase_C26, Peptidase_S51, SNO
Glyco_hydro_tim	Alpha-amylase, Cellulase, DUF187, Glyco_hydro_1, Glyco_hydro_10, Glyco_hydro_14, Glyco_hydro_17, Glyco_hydro_18, Glyco_hydro_20, Glyco_hydro_25, Glyco_hydro_26, Glyco_hydro_2_C, Glyco_hydro_3, Glyco_hydro_30, Glyco_hydro_31, Glyco_hydro_35, Glyco_hydro_39, Glyco_hydro_42, Glyco_hydro_53, Glyco_hydro_56, Glyco_hydro_59, Glyco_hydro_70, Glyco_hydro_72, Glyco_hydro_77, Glyco_hydro_92, Melibiase
Glyoxalase	3-dmu-9_3-mt, Glyoxalase, YecM
GME	Amidinotransf, AstB, PAD, PAD_porph
Golgi-transport	Arfaptin, BAR, DUF1208, IMD, Sec34, Vps5
GPCR_A	7tm_1, 7tm_4, Bac_rhodopsin, DUF621, Serpentine_recp, Serpentine_r_xa, TAS2R, V1R
GT-A	Chitin_synth_1, Chitin_synth_2, CTP_transf_3, DUF23, Galactosyl_T_2, Glycos_transf_2, Glyco_transf_25, Glyco_transf_34, Glyco_transf_43, Glyco_transf_6, Glyco_transf_8, GNT-I, IspD, NTP_transferase, UDPGP

GT-B	DUF1205, DUF354, Epimerase_2, Glycogen_syn, Glycos_transf_1, Glyco_transf_20, Glyco_transf_28, Glyco_transf_9, Glyco_tran_28_C, Glyphos_transf, LpxB, MGDG_synth, Phosphorylase, UDPGT
GT-C	ALG3, Alg6_Alg8, Arabinose_trans, DIE2_ALG10, DUF1420, Glucan_synthase, Glyco_transf_22, Mannosyl_trans, PMT, STT3
HAD	Acid_phosphat_B, Hydrolase, Hydrolase_3, LNS2, PMM, PNK3P, Put_Phosphatase, S6PP, Trehalose_PPase
HD_PDEase	HD, HDOD, PDEase_I
Herpes_glyco	Herpes_UL49_5, Herpes_UL73
HHH	HHH, HhH-GPD
His-Me_finger	Endonuclease_1, Endonuclease_7, Endonuclease_NS, HNH, MH1
Histone	Bromo_TP, CBFD_NFYB_HMF, Histone, TAF, TAFII28, TFIID-18kDa, TFIID-31kDa, TFIID_20kDa
His_Kinase_A	HisKA, HisKA_2, HisKA_3, HWE_HK
HMG-box	CHDNT, DUF1014, HMG_box, YABBY
HNOX-like	HNOB, V4R, XylR_N
HO	Heme_oxygenase, TENA_THI-4
HotDog	4HBT, Acyl-ACP_TE, Acyl_CoA_thio, FabA, MaoC_dehydratas
HSP20	CS, DUF1813, HSP20
HTH	Arg_repressor, Bac_DnaA_C, CENP-B_N, Crp, Dimerisation, DUF134, DUF1670, DUF24, DUF293, DUF742, DUF977, E2F_TDP, Exc, FaeA, Fe_dep_repress, FUR, GcrA, GerE, GntR, HTH_1, HTH_10, HTH_11, HTH_12, HTH_3, HTH_5, HTH_6, HTH_7, HTH_8, HTH_9, HTH_AraC, HTH_CodY, HTH_DeoR, HTH_IclR, HTH_Mga, HTH_psq, Ins_element1, LacI, LexA_DNA_bind, MarR, MerR, Mga, NUMOD1, PaaX, PadR, Pencillinase_R, Phage_AlpA, Phage_antitermQ, Phage_CII, Phage_CI_repr, Put_DNA-bind_N, RepL, Rrf2, Sigma54_DBD, Sigma70_ECF, Sigma70_r4, Sigma70_r4_2, Sulfolobus_pRN, TBPIP, Terminase_5, TetR_N, TFIIE_alpha, Transposase_8, Trans_reg_C, TrmB, Trp_repressor, UPF0122, z-alpha
Hybrid	Biotin_lipoyl, GCV_H, HlyD, PTS_EIIA_1, PYNP_C
Ig	C1-set, C2-set, C2-set_2, I-set, ig, V-set, V-set_CD47
Insulin	Insulin, Ins_beta
Ion_channel	Ion_trans, Ion_trans_2, IRK, PKD_channel, TrkH

IT	ABG_transport, ACR_tran, ArsB, CitMHS, DctM, DcuA_DcuB, DcuC, DUF1504, DUF1646, DUF401, GntP_permease, Lactate_perm, MatC_N, Na_H_antiporter, Na_sulph_symp, NhaB, SCFA_trans
Kazal	Kazal_1, Kazal_2
Kelch	Kelch_1, Kelch_2
Ketolase-like	Transket_pyr, XFP
KH	KH_1, KH_2
Kleisin	Rad21_Rec8, ScpA_ScpB
Knottin_1	Defensin_2, Gamma-thionin, Toxin_17, Toxin_2, Toxin_3, Toxin_5
KOW	EFP_N, KOW, Ribosomal_L21e, Ribosomal_L2_C
LolA_LolB	LolA, LolB
LRR	LRR_1, LRR_2, LRR_3
LysM	LysM, OapA, Phage_tail_X
Lysozyme	DUF847, Glyco_hydro_19, Glyco_hydro_46, Lys, Phage_lysozyme, SLT, Transglycosylas
M6PR	CIMR, Man-6-P_recep
MACRO	A1pp, Peptidase_M17_N
Matrix	Gag_MA, Gag_p10, Gag_p17, Gag_p19, Retro_M
MazG	MazG, PRA-PH
MBB	Ail_Lom, Autotransporter, MipA, OmpA_membrane, Omptin, OmpW, Opacity, OprB, OprD, OprF, OstA_C, Porin_1, Porin_2, Porin_O_P, Surface_Ag_2, Toluene_X, TonB_dep_Rec, YfaZ
MBD-like	AP2, Integrase_DNA, MBD
Membrane_trans	ABC-3, AmoA, BPD_transp_2, FecCD, FTSW_RODA_SPOVE
Methionine_synt	Meth_synt_1, Meth_synt_2
Methyltransfer	Bin3, CheR, CMAS, Cons_hypoth95, DNA_methylase, DOT1, DREV, DUF248, DUF574, DUF858, DUF938, Eco57I, Fibrillarin, FtsJ, GCD14, GidB, MethyltransfD12, Methyltransf_10, Methyltransf_11, Methyltransf_12, Methyltransf_2, Methyltransf_3, Methyltransf_4, Methyltransf_5, Methyltransf_8, Methyltransf_9, MetW, Met_10, Mg-por_mtran_C, MT-A70, MTS, N2227, N6_Mtase, N6_N4_Mtase, NNMT_PNMT_TEMT, NodS, Nol1_Nop2_Fmu, NSP13, PARP_regulatory, PCMT, Pox_MCEL, PrmA, RrnaAD, rRNA_methylase, Rsm22, Spermine_synth, TehB, TPMT, TRM, tRNA_U5-meth_tr, Ubie_methyltran, UPF0020

Met_repress	Arc, DUF1778, HicB, MetJ, Omega_Repress, RelB, RHH_1, RHH_2
MFS	BT1, CLN3, DUF1228, DUF1602, DUF475, DUF791, DUF894, DUF895, Folate_carrier, LacY_symp, MFS_1, MFS_Mycoplasma, Nodulin-like, Nucleoside_tran, Nuc_H_symport, OATP, PTR2, PUCC, Sugar_tr, TRI12
MIF	CHMI, MIF, Tautomerase
MORN	MORN, MORN_2
Mss4-like	Mss4, SelR, TCTP
MtN3-like	LAB_N, MtN3_slv, PQ-loop, UPF0041
MviN_MATE	MatE, MVIN, Polysacc_synt, Rft-1
NADP_Rossmann	2-Hacid_dh_C, 3Beta_HSD, 3HCDH_N, adh_short, ADH_zinc_N, AdoHcyase_NAD, AlaDh_PNT_C, Amino_oxidase, ApbA, CoA_binding, DAO, DapB_N, DXP_reductoisom, ELFV_dehydrog, Epimerase, F420_oxidored, FAD_binding_2, FAD_binding_3, FMO-like, G6PD_N, GDI, GFO_IDH_MocA, GIDA, GMC_oxred_N, Gp_dh_N, HI0933_like, IlvN, KR, Ldh_1_N, Lycopene_cycl, Malic_M, Mannitol_dh, Mqo, Mur_ligase, NAD_binding_2, NAD_binding_3, NAD_binding_4, NAD_binding_5, NAD_Gly3P_dh_N, NmrA, OCD_Mu_crystall, PDH, Polysacc_synt_2, Pyr_redox, Pyr_redox_2, RmlD_sub_bind, Saccharop_dh, SE, Semialdhyde_dh, Shikimate_DH, THF_DHG_CYH_C, Thi4, ThiF, TrkA_N, Trp_halogenase, UDPG_MGDP_dh_N
NAD_Ferredoxin	NAD_binding_1, NAD_binding_6
NagB-like	5-FTHF_cyc-lig, AcetylCoA_hydro, CoA_trans, DeoR, Glucosamine_iso, IF-2B, Rib_5-P_isom_A, Sugar-bind
NfeD-like	DUF1449, NfeD
NifU	DUF59, NifU
NTF2	CaMKII_AD, DUF1348, LEH, MecA_N, NTF2, Ring_hydroxyl_B, Scytalone_dh, SnoaL
NTN	AAT, Asparaginase_2, CBAH, GATase_2, G_glu_transpept, Penicil_amidase, Peptidase_C69, Proteasome
NTP_transf	Adenyl_transf, DUF294, GlnE, NTP_transf_2, RelA_SpoT
Nucleocapsid	Ebola_NP, Paramyxo_ncap
NUDIX	NUDIX, NUDIX-like

OB	CSD, DNA_ligase_OB, DUF388, EFP, eIF-1a, eIF-5a, mRNA_cap_C, OB_RNB, Phage_DNA_bind, Rep-A_N, Rho_RNA_bind, Ribosomal_L2, Ribosomal_S12, Ribosomal_S17, RNA_pol_Rpb8, RuvA_N, S1, SSB, Telo_bind, TOBE, TOBE_2, tRNA_anti, tRNA_bind
Omega_toxin	Albumin_I, Omega-toxin, Toxin_11, Toxin_12, Toxin_16, Toxin_21, Toxin_24, Toxin_27, Toxin_30, Toxin_7, Toxin_9
OstA	DUF1239, OstA
ox_reductase_C	GFO_IDH_MocA_C, ox_reductase_C
P53-like	NDT80_PhoG, P53, RHD, Runt, STAT_bind, T-box
PAN	PAN_1, PAN_2, PAN_3
ParBc	ParBc, ParBc_2
PAS	MEKHLA, PAS, PAS_2, PAS_3, PAS_4, PAS_5, PAS_6
PBP	Bug, Lipoprotein_9, LysR_substrate, NMT1, OpuAC, SBP_bac_1, SBP_bac_3, SBP_bac_7, Transferrin
PDDEXK	CoiA, DUF1016, DUF1052, DUF1064, DUF790, DUF91, DUF911, Herpes_alk_exo, Herpes_UL24, Hjc, Mrr_cat, RmuC, SfsA, UPF0102, VRR_NUC
Peptidase_AA	Asp, Peptidase_A3, RVP, RVP_2, Spuma_A9PTase
Peptidase_AD	DUF1119, Peptidase_A22B, Peptidase_A24, Presenilin
Peptidase_CA	Amidase_5, CHAP, NLPC_P60, Peptidase_C1, Peptidase_C10, Peptidase_C12, Peptidase_C16, Peptidase_C1_2, Peptidase_C2, Peptidase_C21, Peptidase_C23, Peptidase_C27, Peptidase_C28, Peptidase_C31, Peptidase_C32, Peptidase_C33, Peptidase_C36, Peptidase_C39, Peptidase_C42, Peptidase_C47, Peptidase_C54, Peptidase_C58, Peptidase_C6, Peptidase_C7, Peptidase_C8, Peptidase_C9, Viral_protease
Peptidase_CD	Peptidase_C11, Peptidase_C13, Peptidase_C14, Peptidase_C25, Peptidase_C50
Peptidase_CE	Peptidase_C48, Peptidase_C5, Peptidase_C55
Peptidase_MA	Astacin, DUF1695, DUF45, Peptidase_M1, Peptidase_M10, Peptidase_M11, Peptidase_M13, Peptidase_M2, Peptidase_M27, Peptidase_M3, Peptidase_M32, Peptidase_M35, Peptidase_M36, Peptidase_M4, Peptidase_M41, Peptidase_M43, Peptidase_M6, Peptidase_M61, Peptidase_M7, Peptidase_M8, Peptidase_M9, Reprolysin, WLM
Peptidase_MD	HH_signal, Peptidase_M15, Peptidase_M15_2, Peptidase_M15_3, Peptidase_M74, VanY
Peptidase_ME	LuxS, Peptidase_M16, Peptidase_M16_C, Peptidase_M44

Peptidase_MH	AstE_AspA, Nicastrin, Peptidase_M14, Peptidase_M17, Peptidase_M18, Peptidase_M20, Peptidase_M28, Peptidase_M42
Peptidase_ML	DUF1256, HycI, Peptidase_A25
Peptidase_MX	DUF955, Peptidase_M48, Peptidase_M56, Zn_peptidase
Peptidase_PA	DUF30, DUF316, Peptidase_C24, Peptidase_C3, Peptidase_C30, Peptidase_C37, Peptidase_C4, Peptidase_S29, Peptidase_S3, Peptidase_S30, Peptidase_S31, Peptidase_S32, Peptidase_S39, Peptidase_S55, Peptidase_S6, Peptidase_S7, Trypsin
Peptidase_SH	Peptidase_S21, Peptidase_U35, Peptidase_U9
Pept_Inhib_IE	CarbpepA_inh, Squash
Periplas_BP-like	ABC_sub_bind, Bmp, Peripla_BP_1
PFK	DAGK_cat, NAD_kinase, PFK
PGBD	PG_binding_1, PG_binding_2
PGM	Acid_phosphat_A, PGAM
Phage_tail_L	DUF1833, Phage_tail_L
Phosphatase	CDKN3, DSPc, Y_phosphatase, Y_phosphatase2
Phospoesterase	FBPase, FBPase_glpX, Inositol_P
PKinase	ABC1, APH, APH_6_hur, Choline_kinase, DUF1679, DUF227, Fructosamin_kin, Kdo, Pkinase, Pkinase_Tyr, Pox_ser-thr_kin, RIO1, WaaY
PK_TIM	HpcH_HpaI, Malate_synthase, PEP-utilizers_C, PK
Plasmid_toxin	Plasmid_killer, Plasmid_stabil, Plasmid_Txe
PLP_aminotran	Alliinase_C, Alum_res, Aminotran_1_2, Aminotran_3, Aminotran_5, Beta_elim_lyase, Cys_Met_Meta_PP, DegT_DnrJ_EryC1, GDC-P, OKR_DC_1, Pyridoxal_deC, SelA, SHMT, SLA_LP_auto_ag
POTRA	POTRA_1, POTRA_2, Surf_Ag_VNR, YqfD
POZ	BACK, BTB, K_tetra, Skp1_POZ
PP-loop	Arginosuc_synth, Asn_synthase, ATP_bind_3, ATP_bind_4, ExsB, NAD_synthase, PAPS_reduct, ThiI, tRNA_Me_trans
PP2C	PP2C, SpoIIE
PRD	PRD, PRD_Mga
Prefoldin	Prefoldin, Prefoldin_2
PspA	PspA_IM30, Snf7
PUA	DUF1530, DUF167, DUF437, DUF978, DUF984, PUA
RdRP	RdRP_1, RdRP_2, RdRP_3, RdRP_4
RecA-like	Rad51, RecA
Rep	DUF1424, Gemini_AL1, MobA_MobL, Phage_GPA, Relaxase, Rep_1, T_Ag_DNA_bind, Viral_Rep
Rhomboid-like	DER1, DUF1751, Rhomboid

Ribokinase	ADP_PFK_GK, Carb_kinase, HK, PfkB, Phos_pyr_kin
RIIa	Dpy-30, RIIa
RING	U-box, zf-C3HC4, zf-MIZ, zf-RING-like
RNase_H	3_5_exonuc, CAF1, DDE, DNA_pol_B_exo, Exonuc_X-T, Mu_transposase, Phage_Lacto_M3, Piwi, RnaseH, RNase_HII, RuvC, rve, Transposase_11
RNA_ribose_bind	eRF1_3, Ribosomal_L7Ae, SpoU_sub_bind
Rotavirus_VP7	Rotavirus_VP7, VP7
RRM	BRAP2, Calcipressin, MPPN, RRM_1, RRM_2, RRM_3, Smg4_UPF3
Rubredoxin	COX5B, Desulfoferrod_N, Rubredoxin
RVT	RVT_1, RVT_2
SAM	SAM_1, SAM_2, SAM_PNT
SGNH_hydrolase	DUF459, Hema_esterase, Lipase_GDSL
SH3	SH3_1, SH3_2, SH3_3, SH3_4, SH3_5
ShK-like	Crisp, ShK
SIS	PGI, SIS
SNARE	Clat_adaptor_s, Sedlin_N, Sybindin
SSRP1-like	Rtt106, SSrecog
Steroid_dh	DUF1295, ERG4_ERG24, ICMT, Steroid_dh
STIR	SEFIR, TIR
SufE_NifU	NifU_N, SufE
TetR_C	TetR_C, TetR_C_2, TetR_C_3, TetR_C_4, TetR_C_5
Thiolase	ACP_syn_III, ACP_syn_III_C, Chal_sti_synt_C, Chal_sti_synt_N, FAE1_CUT1_RppA, HMG_CoA_synt_N, ketoacyl-synt, Ketoacyl-synt_C, Thiolase_C, Thiolase_N
Thioredoxin-like	AhpC-TSA, ArsC, Calsequestrin, DSBA, DUF1687, DUF836, DUF953, ERp29_N, Glutaredoxin, GSHPx, GST_N, HyaE, OST3_OST6, Phosducin, Redoxin, SCO1-SenC, SH3BGR, T4_deiodinase, Thioredoxin
TIM_barrel	Aldolase, DeoC, DHO_dh, DUF556, DUF561, Dus, FMN_dh, G3P_antiterm, Glu_synthase, His_biosynth, IGPS, IMPDH, NanE, NAPRTase, NPD, OMPdecase, Oxidored_FMN, PcrB, PdxJ, PRAI, QRPTase_C, Ribul_P_3_epim, SOR_SNZ, ThiG, TIM, TMP-TENI, Trp_syntA
TPR	Coatomer_E, HAT, HemY_N, NSF, PPR, Sel1, TPR_1, TPR_2, TPR_3, TPR_4
Traffic	Sec20, V-SNARE
TRASH	Arc_trans_TRASH, Ribosomal_L24e, YHS, zf-MYM
TRB	BPL_C, FeoA, KorB_C

Trefoil	AbfB, Agglutinin, Botulinum_HA-17, CDtoxinA, DUF569, Fascin, FGF, FRG1, IL1, Kunitz_legume, MIR, Ricin_B_lectin, Toxin_R_bind_C
Trigger_C	SurA_N, Trigger_C
tRNA_synt_I	tRNA-synt_1, tRNA-synt_1b, tRNA-synt_1c, tRNA-synt_1d, tRNA-synt_1e, tRNA-synt_1f, tRNA-synt_1g
tRNA_synt_II	AsnA, BPL_LipA_LipB, DUF544, tRNA-synt_2, tRNA-synt_2b, tRNA-synt_2c, tRNA-synt_2d, tRNA-synt_2e
Tudor	7kD_DNA_binding, Agenet, Chromo, Chromo_shadow, MBT, PWWP, SMN, TUDOR
TypeIII_Chap	CesT, Chaperone_III, DspF
UBA	CUE, DMA, DUF1296, TAP_C, UBA, UBA_2
UBC	RWD, UEV, UQ_con
Ubiquitin	APG12, CIDE-N, DUF933, MAP1_LC3, PB1, PI3K_rbd, RA, RBD, TGS, ThiS, ubiquitin, UBX, UPF0125, UPF0185, Urm1
uPAR_Ly6_toxin	Activin_recp, BAMBI, PLA2_inh, Toxin_1, UPAR_LY6
Viral_Gag	Gag_p24, Gag_p30
Viral_NABP	Carla_C4, CTV_P23, Viral_NABP
Viral_ssRNA_CP	Bromo_coat, Calici_coat, Cucumo_coat, Peptidase_A6, Rhv, Tymo_coat, Viral_coat
vWA-like	DUF1194, DUF444, Ku_N, Sec23_trunk, Ssl1, TerD, VWA, VWA_CoxE
XI_TIM	AP_endonuc_2, DUF692, HMGL-like, iPGM_N, Orn_Arg_deC_N, UvdE, UxuA
Yip1	YIF1, Yip1
Zn_Beta_Ribbon	Auto_anti-p27, GATA, PhnA_Zn_Ribbon, Prim_Zn_Ribbon, Ribosomal_S27e, RNA_POL_M_15KD, TFIIB_Zn_Ribbon, Topo_Zn_Ribbon, Transposase_35, zf-C4_Topoisom, zf-CHC2

Chapter 9

Declaration

9.1 Erklärung

Hiermit erkläre ich ehrenwörtlich das ich die vorliegende Dissertation selbständig verfasst und keine anderen als die zitierten Quellen und Hilfsmittel verwendet habe.

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

Ausser dem Diplom in Biologie an der Universität Würzburg habe ich bisher keine anderen akademischen Grade erworben oder versucht zu erwerben.

Planegg, Dezember 2008

Stefan Pinkert