Linking Fold, Function and Phylogeny: A Comparative Genomics View on Protein (Domain) Evolution

Aartjan J.W. te Velthuis² and Christoph P. Bagowski^{1,*}

Institute of Biology ¹*Department of Integrative Zoology, Leiden University, 2333 AL Leiden, The Netherlands and*
²*Department of Molecular Virology, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA Leiden, The Netherlands*

Abstract: Domains are the building blocks of all globular proteins and present one of the most useful levels at which protein function can be understood. Through recombination and duplication of a limited set of domains, proteomes evolved and the collection of protein superfamilies in an organism formed. As such, the presence of a shared domain can be regarded as an indicator of similar function and evolutionary history, but it does not necessarily imply it since convergent evolution may give rise to similar gene functions as well as architectures.

Through the wealth of sequences and annotation data brought about by genomics, evolutionary links can be sought for *via* homology relationships and comparative genomics, structural modeling and phylogenetics. The goal hereby is not only to predict the function of newly discovered proteins, but also to spell out their pathway of evolution and, possibly, identify their most likely origin. This can ultimately help to understand protein function and functional relationships of protein families. Additionally, through comparison with transcriptional data, evolutionary data can be linked to gene (and genome) activity and thus allow for the identification of common principles behind fast evolving proteins and relatively stable ones.

In this review, we describe the basic principles of studying protein (domain) evolution and illustrate recent developments in molecular evolution and give valuable new insights in the field of comparative genomics. As an example, we include here molecular models of the multiple PDZ domain protein MUPP-1 and present a simple comparative genomic view on its structural course of evolution.

Received on: February 26, 2008 - Revised on: March 20, 2008 - Accepted on: March 25, 2008

Key Words: Domain, phylogeny, alignment, MUPP, PDZ, molecular evolution, protein folding, MPDZ, molecular modeling, multiple PDZ domain protein.

COMPARATIVE STRUCTURAL AND FUNCTIONAL GENOMICS

The genome projects of the last decade have produced a staggering amount of sequence data, but most of the identified genes lack experimental determination of biological function or even in some instances identification. The advances in bioinformatics have allowed large-scale genome comparisons, and efforts are well under way to make similar use of comparative functional and structural genomic approaches. However, the wealth of comparative genomic data generated has yet to be followed by a comparable gain of structural and functional information.

The annotation of genes, the prediction of new genes and the allocation of regulatory elements to date largely relies on evolutionary relationships for which genome comparison is fundamental [1, 2]. In essence, comparative genomics is based on the assumption that the two (or more) analyzed genomes share a common ancestor and that the bases in the sequence of each organism are the result of evolution acting on the genome of this mutual ancestor.

In general, evolution forms and molds genomes through two processes, namely mutational forces that generate random changes (*i.e.*, point mutations or insertion-deletions [indels]) and selection pressures which can be positive, negative or neutral with regard to the presence of the mutation in the next generation [3, 4]. The combined effect of mutation and selection can subsequently be calculated and presented in a rate matrix, which denotes the probability of a mutation from one amino acid (or nucleotide) into another for a given period of time [5]. In turn, the rate matrix can be used to calculate alignments of two or more functional sequences. These functional sequences are, by definition, functions that are under evolutionary selection and are often a sequence of amino acids. However, they can, for example, also be transcription factor binding sites or RNA structures (*e.g.* microRNAs or viral RNA genomes). Commonly used rate matrices are BLOSUM and PAM [5, 6], which can readily be found implemented in BLAST and other well known sequence alignment programs [7-10].

As a result, a specific gene or protein of unknown function and biological importance can be compared to the sequence of a set of proteins with characterized functions. From these, the best matching group can be selected based on the number of domains and the nature of these domains.

*Address correspondence to this author at the Institute of Biology, Leiden University, Wassenaarseweg 64, 2333 AL Leiden, The Netherlands;
Tel: ++31-(0)71-527-4802; Fax: ++31-(0)71-527-4999;
E-mail: c.p.bagowski@biology.leidenuniv.nl

This information can be used to annotate the predicted gene or protein [2, 11-13].

Indeed, comparing genomes provides new insights into the biology of organisms whose hereditary material is under scrutiny. Some recent papers of comparisons between prokaryotes (e.g., γ -proteobacteria) [14, 15], insects (e.g., *A. gambiae* to *D. melanogaster*) [16, 17], mammals (e.g., *M. musculus* to *H. sapiens*) [18, 19], but also more distant comparisons between yeast and human genomes [20] are good examples of this approach. Furthermore, these studies have shed light upon transcriptional regulation [21-25], horizontal gene transfer [14, 24, 26], conservation of proteome networks [20, 27, 28] and strain-specific adaptations [29]. The combined data in GenBank and other databases now covers sequences for over 200,000 species with at least 50 complete genomes, which makes numerous more genome comparisons feasible [30-32]. But comparative genomics, especially when combined with proteomics, protein folding and microarray data, offers far more than just that; it can be used to explicate the evolution of proteins and the structures that make up proteins: the domains. In this review we describe the approaches currently available to elucidate the evolutionary history of proteins and their domains. We also provide examples, based on the PDZ domains of the Multiple PDZ Domain Protein-1 (MUPP-1; MPDZ) [33] and the single PDZ domain protein Dishevelled (Dsh) [34]. MUPP-1 is an important scaffolding protein, which could potentially play important roles in lipid raft assembly [35], in viral entry [36] and in cancer progression [37]. Dsh, with two different additional protein binding domains, a DIX and a DEP domain, plays a central role in development of invertebrates and vertebrates [38].

SEQUENCE ALIGNMENT AND PHYLOGENY

Central biological features like metabolism, transcription and cell cycle progression are conserved from prokaryotes and single cell eukaryotes to humans [39, 40]. This conservation motivated and established the use of model organisms for studying conserved processes that are difficult or expensive to assess in higher organisms. Technological advances over the past two decades have led to the accumulation of genome-wide sequence data for many different species (see e.g., <http://www.ensembl.org>), but in order to use these sequences they have to be compared to each other in either pair-wise alignments (e.g., used in BLAST) or multiple sequence alignments, in which multiple sequences are compared simultaneously to each other (e.g., employed in ClustalX, Phylip and Muscle (see Table 1)).

Alignments can also be subdivided based on the terms *global* and *local*. When whole genomes are aligned, bases are lined up by inserting gaps in sequences to account for (hypothetical) insertions or deletions that have taken place since diversification from the common ancestor. Indeed, this can be performed from one end to the other, as global implies, but when working with small genomes of several thousand base pairs or with entire chromosomes of hundred million base pairs it will need processing power and will be time consuming. Therefore, it is mostly applied to relatively short gene or protein sequence data, although web-based alignments can also be browsed (e.g., <http://www.dcode.org>). For the longer genomic nucleic acid sequences, a focus on regions of (local) high similarity is more feasible; the low se-

quence similarity regions are then ignored, which makes the procedure altogether much faster.

Automated alignments commonly employ a scoring procedure to find the best alignment possible for the input sequences. This scoring takes into account the number of identical residues, the number of different residues, and the size and number of gaps present in the alignment. Each different residue and bigger or extra gap will result in a penalty. Additionally, different penalties are created for the differences between for example transversions and transitions; with the latter being more common and thus favored over transversions [41]. However, the optimized alignment may not be the true one, since parameters can vary from species to species [42]. It is therefore recommended to manually check alignments and improve them (see Table 1 for programs). Fig. (1A) shows an example alignment of a number of PDZ domains with different shadings representing the amount of conservation (100, 75 or 50%) at a particular position in the sequence.

Evolutionary distances can easily be estimated from small sequence alignments and can subsequently be used to create phylogenies, but also approximate divergence times, rates of evolution and ancestry sequences can be delineated from them. For phylogenetic analysis, multiple software packages are now available that often use one of these approaches: Maximum Likelihood [43, 44], Maximum Parsimony [7, 45], Neighbor Joining [7, 9] or Bayesian Estimation [46, 47] (see also Table 1). To provide an example of such a phylogenetic tree, we used MrBayes to calculate, over 100,000 generations and a mixed rate matrix set, the best tree topology for the alignment given in Fig. (1A).

Since, the MUPP-1 protein of *Tetraodon nigroviridis* has 10 domains, *Xenopus tropicalis* 12 and *Homo sapiens* 13 one hypothesis could be that the last domain of the "ten domain structure" duplicated two to three times to make up for the extra 2 or 3 domains found in the higher vertebrates. If this holds true, the last three PDZ domains should cluster closely together in the phylogenetic tree. However, this appears to be not the case: *Tetraodon nigroviridis* PDZ 8 clusters with *Xenopus tropicalis* PDZ 9 and *Homo sapiens* PDZ 10, which suggests at least one domain duplication event in the middle of the protein. The separate clustering of *Xenopus tropicalis* PDZ 8 with *Homo sapiens* PDZ 8 points to an insertion event in their common ancestor, however. Of course, we can not exclude from this small analysis that the domain was already present in the very early vertebrates and only lost in *Tetraodon*. We will try to shine more light on this with a structural model of these events in Fig. (2B).

All phylogenetic information is extremely dependent on a proper alignment and not so much on the programs used to infer phylogeny [48]. Recently, software has been developed to combine the alignment procedure and phylogenetic analysis in one single program [47]. Current versions of this software can, however, only handle a limited set of sequences.

PROTEIN DOMAIN CLASSIFICATION AND SUPER-FAMILIES

By definition, a domain is a structural, functional, but also an evolutionary component of a protein. Domain duplication and reorganization play important roles in evolution.

Table 1. List of Resources and Databases Relevant to Comparative Genomics

Internet Resource	URL
Sequence motif searches	
Pfam	http://www.sanger.co.uk/Pfam/
PRINTS	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/
Prosite	http://www.expasy.org/prosite/
SMART	http://smart.embl-heidelberg.de
Superfamily	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/hmm.html
Structural comparison	
CATH	http://www.cathdb.info/latest/index.html
DaliLite	http://www.ebi.ac.uk/DaliLite/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
SSM	http://www.ebi.ac.uk/msd-srv/ssm/
Alignment software	
BLAST	http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
ClustalW	http://www.ebi.ac.uk/Tools/clustalw2/
Dcode	http://www.dcode.org
MEGA	http://www.megasoftware.net/
Muscle	http://www.ebi.ac.uk/muscle/
T-coffee	http://www.tcoffee.org/
Alignment optimization	
GBlocks	http://molevol.ibmb.csic.es/Gblocks.html
GeneDoc	http://www.nrbsc.org/gfx/genedoc/index.html
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
Phylogenetic analysis software	
MrBayes	http://mrbayes.csit.fsu.edu/
PhyML	http://atgc.lirmm.fr/phyml/
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
Tree-Puzzle	http://www.tree-puzzle.de/
MEGA	http://www.megasoftware.net/
Visualization	
Pymol (structural)	http://pymol.sourceforge.net/
NJplot (phylogeny)	http://pbil.univ-lyon1.fr/software/njplot.html
MEGA	http://www.megasoftware.net/
Swiss-Model (structural)	http://swissmodel.expasy.org/
TreeView (phylogeny)	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
Databases	
Ensembl (Sanger project)	http://www.ensembl.org
PDB	http://www.rcsb.org/pdb/home/home.do
NCBI	http://www.ncbi.nlm.nih.gov/sites/gquery?itool=toolbar
Superfamily	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/hmm.html
UniProt	http://www.expasy.uniprot.org/

It has been estimated that at least 70% of the domains duplicated in prokaryotes. In eukaryotes this number is presumed to be even higher, ranging up to 90% [49]. Not surprisingly, many proteins comprise of more than one domain [1, 50, 51].

Domains are essential and versatile evolutionary elements that have been used to create from a relatively limited set an enormous and diverse assembly of proteins. Many protein family resources (*e.g.*, Prosite and Pfam (see Table 1)) present a hierarchical classification that is almost fully dependent on sequence similarity and motif identification. Close relatives, sharing for example >50% sequence identity and often also functional properties, are grouped into families and subfamilies (*e.g.*, PRINTS (see Table 1)). In turn, these families are grouped with other families into superfamilies [49, 52], with which they share for example ~25% sequence similarity. For a recent review on the function of these databases see reference [13].

PROTEIN DOMAIN FOLDING

After sequence analysis, the question arises whether sequence divergence is correlated with structural divergence and ultimately functional divergence. In the 1970s technologies (NMR and X-ray crystallography) for determining the 3D structure of domains and proteins became established. It was found that protein structures are primarily composed of α -helical and β -strand secondary structures (see Fig. 2 for a PDZ domain model structure) and there usually is a clear way to achieve optimal packing of the hydrophobic residues in the core of the protein (or sometimes outside, in case of a transmembrane protein).

As the number of solved structures increased it quickly became evident that protein (domain) structures are much more conserved (~50%) than the protein (amino acid) sequence (~5%) [53]. For this reason, it is possible that protein structures and their models can be used to find close as well as very distant relatives. Indeed, sometimes it is difficult to recognize divergent relatives solely through sequence comparison and often for these cases, there are no features present indicative of mutual functional properties [54]. There are two possible explanations: both domains or proteins have evolved from two different ancestral proteins; or they are two extremely distant relatives that started out from the same evolutionary ancestor [50, 54]. To distinguish between these possibilities, it is important to look at the current understanding of domain evolution. It is believed that the small set of protein domains known to date, descended from an even smaller group of ancestral domains. Unlike the raw protein sequence, the core of the protein domain is largely stable as it must be functionally conserved (*i.e.*, selection is on function) and relies on inter-residue dependence. It is likely that protein evolution took place – or rather started – at the periphery of the relatively constant core. Indeed, it was shown that in pair-wise alignments, the amount of indels correlates with the evolutionary distance of proteins [4, 55, 56]. The structures most susceptible to point mutations, insertions or deletions are typically surface loops [57]. Unless mutations in these areas are neutralized, the number of changes will accumulate and eventually generate new polypeptide folds. Subsequently, positive selection will favor some of these

newly arisen substructures when they become implemented in the biological process.

It should be clear from the above that the process of structural evolution is of a completely different order than that of sequence evolution, which is much faster. The tertiary sequence of a protein contains therefore much more phylogenetic signal and makes it far more likely to find linkages beyond the timeframe of standard sequence alignments [54]. Indeed, it may not be surprising that, like recognizing distinct sequence similarities, distinct folds and structures can be identified and classified as well. Examples are SCOP and CATH (see Table 1), which are linked to the Protein Data Bank (PDB) that stores protein structural data. Moreover, structural information can be used to verify and support phylogenetic data. As an example we modeled the differently clustering PDZ domains of MUPP-1 (the phylogenetic analyses shown in Fig. (1B) implied one insertion and one duplication event to form the extra 2/3 PDZ domains present in the *Xenopus* [PDZs 8 and 9] and *Homo sapiens* structures [PDZs 8-10]). Indeed, PDZ 8 of *Tetraodon nigroviridis* seems structurally highly related to domain 9 of *Xenopus* and domain 9 and 10 of *Homo sapiens*. In other words, either of these structures appears more structurally related to the others in this small group than to any of the other (flanking) PDZ domains, which suggests duplication. The PDZs 8 of *Xenopus* and *Homo sapiens* form, however, a separate structural group as the phylogenetic analysis predicted. We therefore propose that the *Homo sapiens* PDZ 9 originates from a duplication event of the *Xenopus* PDZ 9 and that *Homo sapiens* PDZ 8 is a result of an insertion in the common ancestor of *Xenopus* and *Homo sapiens*. Our evolutionary model shown in Fig. (2B) can thus be used to confirm the phylogenetic tree shown in Fig. (1B).

Even though domains are recognized by prediction programs, like Pfam and SMART, the actual fold may be different due to intermolecular interactions. Proteins usually contain more than one domain (*i.e.*, multidomain proteins) and have evolved through a process of duplication and recombination of the limited set of protein domains available [51]. This principle not only brought together different enzymatic functions into single protein units (*e.g.*, a catalytic domain and an ATP binding domain resulting in a helicase or kinase), but also combined domains that could co-evolve into one larger superdomain. An example of the latter can be found in the MAGUK family of proteins in which the Src homology 3 (SH3) domain and the Guanylate Kinase (GUK) domain interact intramolecularly to form a superdomain involved in protein-protein interactions [58, 59]. Not surprisingly, the GUK domain in these proteins is often only partially active or lacks activity completely and it was recently found that this loss of GUK activity corresponds with a position further away from the origin in the phylogenetic tree of the MAGUK proteins [60, 61].

GENES AND DOMAIN EVOLUTION BEYOND THE SEQUENCES

Important elements in a gene's function are its spatial and temporal expression patterns. In recent years, microarray technology has made an extraordinary number of experiments possible that were aimed to map genome-wide expres-

Fig. (1). Example of sequence alignment and phylogeny. (A) Alignment of PDZ domains of the multiple (thirteen) PDZ domain protein MUPP-1 [33] and Disheveled (Dsh) [34] of several organisms. PDZ domains are modular interaction domains that recognize and bind to 4 C-terminal residues of the target domain, although other binding principles have also been shown. Black shading indicates 100% conservation, while the lighter grays indicates 75% or 50% conservation. Abbreviations used are Hs (*Homo sapiens*), Mm (*Mus musculus*), Xt (*Xenopus tropicalis*), Tn (*Tetraodon Nigroviridis*), Dr (*Danio rerio*), Ci (*Ciona intestinalis*) and Hv (*Hydra vulgaris*). (B) Evolutionary tree inferred by Bayesian Phylogeny (MrBayes) [46], rooted to the Dsh outgroup sequences. The tree shows clustering of most PDZ domains of MUPP-1 according to their sequence number in the protein. However, after PDZ 7 the numbers mix. Tn PDZ 8 clusters with Xt PDZ 9 and Hs PDZ 10, which suggests a domain duplication event. The clustering of Xt PDZ 8 with Hs PDZ 8 (near PDZ 4) and the separate clustering of Hs PDZ 9 suggest together one insertion (of PDZ 8) and one duplication event (of PDZ 9). To explain this most plausible relationship we have also presented this more visually in our structural model in Fig. (2B). It is important to note that MUPP-1 of *Tetraodon nigroviridis* contains 10, of *Xenopus tropicalis* 12 and of *Homo sapiens* 13 PDZ domains. Numbers indicated represent Bayesian posterior support values and all sequences used were obtained and analyzed as described previously [60].

sion levels under a variety of conditions [62-65]. For example, transcriptional comparisons have been made to look at for instance aging [66], pathogenicity [67] and non-coding RNAs [68]. Equivalent data is now, in addition to the sequence data, becoming available for dozens of different species and they provide a rich resource for comparative studies.

Unfortunately, the comparison of distantly related organisms can only be done under strictly defined expression conditions since gene expressions are not static. Indeed, by thoroughly controlling research conditions, comparisons between different (sub)species were made for conditions like embryogenesis, metamorphosis, sex-dependency and mutation

rates [65, 69-72]. Other studies including diverse organisms such as yeasts, plants and primates, have revealed valuable information on promoter types and whether or not genes had previously undergone a duplication event [64, 65, 73, 74].

However, more evolutionary distant organisms may react differently to the same stimulus, which undermines comparison of gene expression data. To overcome this limitation, the association of co-expression data of genes and of expression signatures has been developed in addition to a direct comparison of individual gene expression changes [62]. Firstly, the co-expression between gene pairs is determined for each individual organism (within-species comparison) and this is

Fig. (2). Structural modeling of MUPP-1 PDZ domains and hypothetical model for internal domain duplications. (A) Molecular modeling of the thirteen human PDZ domains of MUPP-1 with Swiss-Model Workspace and Swiss-PBD Viewer 3.7 [83]. (B) In figure 1B, we compared the MUPP-1 PDZ domains of 4 different species. Of these four species, *Tetraodon nigroviridis* MUPP-1 consists of 10 PDZ domains, *Xenopus tropicalis* of 12 and *Homo sapiens* of 13 PDZ domains. Phylogenetic analyses implied that PDZ 8 of the *Tetraodon* MUPP-1 structure duplicated before at least twice to form the extra 2/3 PDZ domains present in the *Xenopus* and *Homo sapiens* structures. We therefore applied molecular modeling to these PDZ domains to visually support these findings. We modeled PDZ domains 7-9 of *Tetraodon nigroviridis*, domains 7-10 of *Xenopus tropicalis* and domains 7-11 of *Homo sapiens*. Indeed, PDZ 8 of *Tetraodon*, seems structurally related to 8 and 9 of *Xenopus* and 8-10 of *Homo sapiens*. However, within this group of six *Xenopus* PDZ 8 and *Homo sapiens* PDZ 8 appear to form a separate group. The most parsimonious explanation (and taking into account both the structural and phylogenetic data) therefore suggests one insertion event and one duplication event.

then compared to the co-expression entities of other organisms. This approach focuses on the similarity and differences of the orthologous genes within their expression networks and this can be compared when species differences do not allow direct comparison at a specific condition. This system already has been applied for several species and it has revealed that both species-specific parts of the expression networks are combinations of conserved and newly evolved modules [62, 75, 76].

Another benefit of comparing co-expression of genes is that often functional entities can be discovered and, subsequently, new leads can be gained for functional interpretation. The approach can be combined with the search for common cis-regulatory elements at the promoter regions or applied to other similarity measures between genes, such as protein-protein interactions, phosphorylation networks or ligand-binding specificities [77-79].

CONCLUDING REMARKS

Finding evolutionary relationships for genes, proteins or protein domains is mostly based on orthology and thus on best sequence matches. Identifying these and categorizing them depends largely on multiple sequence alignments and this will in most cases give good indications for function and fold. However, this approach usually discards apparent ambiguities that arise from species-specific duplications or losses and may therefore introduce extensive biases [80]. Biases may also derive from the method of alignment, the phylogenetic analysis and the sample size used [47, 48, 81]. Therefore, care should be taken to not regard orthology as a pure one-to-one relationship, but as a family of homologous relations [64] and to select for the appropriate method of analysis [48, 81].

Genome and proteome comparisons can be performed by looking at expression data and, preferably, co-expression patterns or protein-protein and phosphorylation interactions. In the end, it will be the ultimate challenge to combine all comparative data (sequence, structure, expression, interaction and function) into one biological network. Indeed, only through putting together data obtained from protein-protein interactions and co-expression networks, conserved functional cell cycle complexes shared among yeast, plants, worms and humans have been revealed [82]. Expectantly, with these approaches we will be able to clearly distinguish how different biological mechanisms integrate, mold and flow along the forces of evolution. This is certainly an exciting and stimulatory time for interdisciplinary genomic research.

REFERENCES

- [1] Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. and Rao, B.S. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **2004**, *5*: R7.
- [2] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*: 25-29.
- [3] Ureta-Vidal, A., Ettwiller, L. and Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **2003**, *4*: 251-262.
- [4] Bin Qian, R.A.G. Distribution of indel lengths. *Proteins Struct. Funct. Genet.* **2001**, *45*: 102-104.
- [5] Dayhoff, M.O., Schwarz, R.M. and Orcutt, B.C. A model of evolutionary change in proteins. In Atlas of protein sequence and structure, M.O. Dayhoff, ed. (Washington D.C.: National Biomedical Research Foundation), **1978**, 345-352.
- [6] Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*: 10915-10919.
- [7] Felsenstein, J. PHYLIP version 3.63. *Dept of Genetics, Univ of Washington, Seattle*. **2004**.
- [8] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **1997**, *25*: 3389-3402.
- [9] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **1997**, *25*: 4876-4882.
- [10] Pearson, W.R. Comparison of methods for searching protein sequence databases. *Protein Sci.* **1995**, *4*: 1145-1160.
- [11] Galperin, M.Y. and Koonin, E.V. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotech.* **2000**, *18*: 609-613.
- [12] Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. Protein function in the post-genomic era. *Nature* **2000**, *405*: 823-826.
- [13] Attwood, T.K. The role of pattern databases in sequence analysis. *Brief. Bioinformatics* **2000**, *1*: 45-49.
- [14] Lerat, E., Daubin, V. and Moran, N.A. From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the gamma-Proteobacteria. *PLoS Biol.* **2003**, *1*: e19.
- [15] Comas, I., ntild, aki, Moya, A., eacute, Gonz, aacute, and lez-Candelas, F. From Phylogenetics to Phylogenomics: The Evolutionary Relationships of Insect Endosymbiotic & gamma-Proteobacteria as a Test Case. *Syst. Biol.* **2007**, *56*: 1 - 16.
- [16] Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A. and Subramanian, G.M. The Interactive Fly: Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **2002**, *298*: 149 - 159.
- [17] Zdobnov, E.M. and Bork, P. Quantification of insect genome divergence. *Trends Genet.* **2007**, *23*: 16 - 20.
- [18] Rozen, S., Skaletsky, H., Marszalek, J.D., Minx, P.J., Cordum, H.S., Waterston, R.H., Wilson, R.K. and Page, D.C. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **2003**, *423*: 873-876.
- [19] Consortium, M.G.S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*: 520-562.
- [20] Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415*: 141-147.
- [21] Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R. and Sachidanandam, R. Comprehensive splice-site analysis using comparative genomics. *Nucl. Acids Res.* **2006**, *34*: 3955 - 3967.
- [22] Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Schmid, R., Hall, N., Barrell, B., Waterston, R.H., McCarter, J.P. and Blaxter, M.L. A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.* **2004**, *36*: 1259-1267.
- [23] Wang, Q., Prabhakar, S., Chanan, S., Cheng, J., Rubin, E. and Boffelli, D. Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol.* **2007**, *8*: R1.
- [24] Price, M., Dehal, P. and Arkin, A. Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biol.* **2008**, *9*: R4.
- [25] Price, M.N., Dehal, P.S. and Arkin, A.P. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput. Biol.* **2007**, *3*: 1739 - 1750.
- [26] Lercher, M.J. and Pal, C. Integration of Horizontally Transferred Genes into Regulatory Interaction Networks Takes Many Million Years. *Mol. Biol. Evol.* **2007**: msm283.
- [27] Wyder, S., Kriventseva, E., Schroder, R., Kadowaki, T. and Zdobnov, E. Quantification of ortholog losses in insects and vertebrates. *Genome Biol.* **2007**, *8*: R242.
- [28] Wang, X., Grus, W.E. and Zhang, J. Gene losses during human origins. *PLoS Biol.* **2006**, *4*: e52.
- [29] Chen, S.L., Hung, C.-S., Xu, J., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P., Armstrong, J.R., Fulton, R.S., Latreille, J.P., Spieth, J., Hooton, T.M., Mardis, E.R., Hultgren, S.J. and Gordon, J.I. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc. Natl. Acad. Sci. USA* **2006**, *103*: 5977-5982.
- [30] Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., Maskeri, B., Hansen, N.F., Schwartz, M.S., Weber, R.J., Kent, W.J., Karolchik, D., Bruen, T.C., Bevan, R., Cutler, D.J., Schwartz, S., Elnitski, L., Idol, J.R., Prasad, A.B., Lee-Lin, S.Q., Maduro, V.V.B., Summers, T.J., Portnoy, M.E., Dietrich, N.L., Akhter, N., Ayele, K., Benjamin, B.,

- Cariaga, K., Brinkley, C.P., Brooks, S.Y., Granite, S., Guan, X., Gupta, J., Haghighi, P., Ho, S.L., Huang, M.C., Karlins, E., Laric, P.L., Legaspi, R., Lim, M.J., Maduro, Q.L., Masiello, C.A., Matrian, S.D., McCloskey, J.C., Pearson, R., Stantripop, S., Tiongson, E.E., Tran, J.T., Tsurgeon, C., Vogt, J.L., Walker, M.A., Wetherby, K.D., Wiggins, L.S., Young, A.C., Zhang, L.H., Osoegawa, K., Zhu, B., Zhao, B., Shu, C.L., De Jong, P.J., Lawrence, C.E., Smit, A.F., Chakravarti, A., Haussler, D., Green, P., Miller, W. and Green, E.D. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **2003**, *424*: 788-793.
- [31] Premzl, M., Gready, J.E., Jermini, L.S., Simonic, T. and Marshall Graves, J.A. Evolution of Vertebrate Genes Related to Prion and Shadoo Proteins--Clues from Comparative Genomic Analysis. *Mol. Biol. Evol.* **2004**, *21*: 2210-2231.
- [32] Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W. and Richards, S. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **2005**, *15*: 1034 - 1050.
- [33] Ullmer, C., Schmuck, K., Figge, A. and Lübbert, H. Cloning and characterization of MUPP1, a novel PDZ domain protein. *FEBS Lett.* **1998**, *424*: 63-68.
- [34] Klingensmith, J., Nusse, R. and Perrimon, N. The Drosophila segment polarity gene dishevelled encodes a novel protein required for response to the wingless signal. *Genes Dev.* **1994**, *8*: 118-130.
- [35] Ackermann, F., Zittramski, N., Heydecke, D., Wilhelm, B., Gudermann, T. and Boekhoff, I. The Multi-PDZ domain protein MUPP1 as a lipid raft-associated scaffolding protein controlling the acrosome reaction in mammalian spermatozoa. *J. Cell. Physiol.* **2008**, *214*: 757-768.
- [36] Coyne, C.B., Voelker, T., Pichla, S.L. and Bergelson, J.M. The coxsackievirus and adenovirus receptor interacts with the multi-PDZ domain protein-1 (MUPP-1) within the tight junction. *J. Biol. Chem.* **2004**, *279*: 48079-48084.
- [37] Martin, T.A., Watkins, G., Mansel, R.E. and Jiang, W.G. Loss of tight junction plaque molecules in breast cancer tissues is associated with a poor prognosis in patients with breast cancer. *Eur. J. Cancer* **2004**, *40*: 2717-2725.
- [38] Wharton, Jr., K.A. Runnin' with the Dvl: proteins that associate with Dsh/Dvl and their significance to Wnt signal transduction. *Development. Biol.* **2003**, *253*: 1-17.
- [39] Kurland, C.G., Collins, L.J. and Penny, D. Genomics and the irreducible nature of eukaryote cells. *Science* **2006**, *312*: 1011-1014.
- [40] Miller, W., Makova, K.D., Nekrutenko, A. and Hardison, R.C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **2004**, *5*: 15 - 56.
- [41] Rosenberg, M.S. and Kumar, S. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Mol. Biol. Evol.* **2003**, *20*: 610-621.
- [42] Vingron, M. and Waterman, M.S. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* **1994**, *235*: 1-12.
- [43] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **1981**, *17*: 368 - 376.
- [44] Guindon, S.p. and Gascuel, O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **2003**, *52*: 696-704.
- [45] Swofford, D. PAUP* 4.0. *Sinauer Associates* **2001**.
- [46] Huelsenbeck, J.P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **2001**, *17*: 754 - 755.
- [47] Lunter, G., Miklos, I., Drummond, A., Jensen, J. and Hein, J. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **2005**, *6*: 83.
- [48] Kumar, S. and Filipinski, A. Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Res.* **2007**, *17*: 127-135.
- [49] Apic, G., Gough, J. and Teichmann, S.A. An insight into domain combinations. *Bioinformatics* **2001**, *17*: S83-89.
- [50] Han, J., Batey, S., Nickson, A.A., Teichmann, S.A. and Clarke, J. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*: 319-330.
- [51] Wolf, Y.I., Grishin, N.V. and Koonin, E.V. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **2000**, *299*: 897-904.
- [52] Wilson, D., Madera, M., Vogel, C., Chothia, C. and Gough, J. The SUPERFAMILY database in 2007: families and functions. *Nucl. Acids Res.* **2007**, *35*: D308-313.
- [53] Chothia, C. and Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*: 823-826.
- [54] Orengo, C.A. and Thornton, J.M. Protein families and their evolution: a structural perspective. *Ann. Rev. Biochem.* **2005**, *74*: 867-900.
- [55] Benner, S.A., Cohen, M.A. and Gonnet, G.H. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **1993**, *20*: 1065-1082.
- [56] Pascarella, S. and Argos, P. Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* **1992**, *224*: 461 - 471.
- [57] Panchenko, A. and Madej, T. Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evol. Biol.* **2005**, *5*: 10.
- [58] Tavares, G.A., Panepucci, E.H. and Brunger, A.T. Structural characterization of the intramolecular interaction between the SH3 and guanylate kinase domains of PSD-95. *Mol. Cell.* **2001**, *8*: 1313-1325.
- [59] McGee, A.W. and Bret, D.S. Identification of an Intramolecular Interaction between the SH3 and Guanylate Kinase Domains of PSD-95. *J. Biol. Chem.* **1999**, *274*: 17431-17436.
- [60] te Velthuis, A., Admiraal, J. and Bagowski, C. Molecular evolution of the MAGUK family in metazoan genomes. *BMC Evol. Biol.* **2007**, *7*: 129.
- [61] Olsen, O. and Bret, D.S. Functional Analysis of the Nucleotide Binding Domain of Membrane-associated Guanylate Kinases. *J. Biol. Chem.* **2003**, *278*: 6873-6878.
- [62] Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **2003**, *302*: 294-255.
- [63] Bergmann, S., Ihmels, J. and Barkai, N. Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Biol.* **2004**, *2*: e9.
- [64] Tirosh, I., Bilu, Y. and Barkai, N. Comparative biology: beyond sequence analysis. *Curr. Opin. Biotechnol.* **2007**, *18*: 371-377.
- [65] Hooper, S.D., Boue, S., Krause, R., Jensen, L.J., Mason, C.E., Ghanim, M., White, K.P., Furlong, E.E.M. and Bork, P. Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. *Mol. Syst. Biol.* **2007**, *3*.
- [66] McCarroll, S.A., Murphy, C.T., Zou, S., Pletcher, S.D., Chin, C.-S., Jan, Y.N., Kenyon, C., Bargmann, C.I. and Li, H. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* **2004**, *36*: 197-204.
- [67] Jeon, J., Park, S., Chi, M., Choi, J., Park, J., Rho, H., Kim, S., Goh, J., Yoo, S., Choi, J., Park, J., Yi, M., Yang, S., Kwon, M., Han, S., Kim, B.R., Khang, C.H., Park, B., Lim, S., Jung, K., Kong, S., Karunakaran, M., Oh, H., Kim, H., Kim, S., Park, J., Kang, S., Choi, W., Kang, S. and Lee, Y. Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nat. Genet.* **2007**, *39*: 561-565.
- [68] Torarinsson, E., Yao, Z., Wiklund, E.D., Bramsen, J.B., Hansen, C., Kjems, J., Tommerup, N., Ruzzo, W.L. and Gorodkin, J. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res.* **2008**, *18*: 242-251.
- [69] Rifkin, S.A., Houle, D., Kim, J. and White, K.P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* **2005**, *438*: 220-223.
- [70] Rifkin, S.A., Kim, J. and White, K.P. Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat. Genet.* **2003**, *33*: 138-144.
- [71] Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D. and Hartl, D.L. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* **2003**, *300*: 1742-1745.
- [72] White, K.P., Rifkin, S.A., Hurban, P. and Hogness, D.S. Microarray analysis of Drosophila development during metamorphosis. *Science* **1999**, *286*: 2179-2184.
- [73] Tirosh, I., Weinberger, A., Carmi, M. and Barkai, N. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* **2006**, *38*: 830-834.
- [74] Landry, C.R., Oh, J., Hartl, D.L. and Cavalieri, D. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* **2006**, *366*: 343-351.
- [75] Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I. and Koonin, E.V. Conservation and Coevolution in the Scale-Free Human Gene Co-expression Network. *Mol. Biol. Evol.* **2004**, *21*: 2058-2070.

- [76] Oldham, M.C., Horvath, S. and Geschwind, D.H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci. USA* **2006**, *103*: 17973-17978.
- [77] Sharan, R., Ulitsky, I. and Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **2007**, *3*.
- [78] Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A., Jørgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J.G., Samson, L.D., Woodgett, J.R., Russell, R.B., Bork, P., Yaffe, M.B. and T., P. Systematic discovery of *in vivo* phosphorylation networks. *Cell* **2007**, *129*: 1415-1426.
- [79] Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. and Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucl. Acids Res.* **2008**, *36*: D684-688.
- [80] Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I. and Hardison, R.C. Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. *Genome Res.* **2003**, *13*: 1-12.
- [81] Blouin, C., Butt, D. and Roger, A.J. Impact of Taxon Sampling on the Estimation of Rates of Evolution at Sites. *Mol. Biol. Evol.* **2005**, *22*: 784-791.
- [82] Jensen, L.J., Jensen, T.S., de Lichtenberg, U., Brunak, S. and Bork, P. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **2006**, *443*: 594-597.
- [83] Guex, N. and Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **1997**, *18*: 2714-2723.