

Proteins: Sequence to Structure and Function – Current Status

Sandhya R. Shenoy and B. Jayaram*

Department of Chemistry & Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110 016, India

Abstract: In an era that has been dominated by Structural Biology for the last 30-40 years, a dramatic change of focus towards sequence analysis has spurred the advent of the genome projects and the resultant diverging sequence/structure deficit. The central challenge of Computational Structural Biology is therefore to rationalize the mass of sequence information into biochemical and biophysical knowledge and to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences. In investigating the meaning of sequences, two distinct analytical themes have emerged: in the first approach, pattern recognition techniques are used to detect similarity between sequences and hence to infer related structures and functions; in the second *ab initio* prediction methods are used to deduce 3D structure, and ultimately to infer function, directly from the linear sequence. In this article, we attempt to provide a critical assessment of what one may and may not expect from the biological sequences and to identify major issues yet to be resolved. The presentation is organized under several subtitles like protein sequences, pattern recognition techniques, protein tertiary structure prediction, membrane protein bioinformatics, human proteome, protein-protein interactions, metabolic networks, potential drug targets based on simple sequence properties, disordered proteins, the sequence-structure relationship and chemical logic of protein sequences.

Keywords: Sequence, structure, function, proteins, membrane, bioinformatics, prediction, metabolism, drug targets, disordered, flexibility.

1. PROTEIN SEQUENCES

Remarkably, at this writing there are 512994 sequence entries in the UniProtKB/Swiss-Prot protein knowledgebase <http://ca.expasy.org/sprot/relnotes/relstat.html> [1] (Fig. 1). The availability of protein sequences has made a telling difference in countless studies of biologically important molecules. This wealth of data has transformed protein chemistry since the early pioneering efforts of Bernal and Crowfoot [2] and Perutz [3]. The vast quantity of data associated with these proteins poses enormous challenges to any attempt at sequence/structure/function annotation. In addition, structure-based programmatic initiatives now are common place, including for example a diversity of database analyses [4], taxonomic classification at the molecular level [5, 6], estimates of the number of folds [7], and pattern recognition-based approaches to prediction [8].

2. PATTERN RECOGNITION TECHNIQUES

Pattern recognition methods are built on the assumption that some underlying characteristic of a protein sequence, or of protein structure, can be used to identify similar traits in related proteins. Conserved protein sequence regions are extremely useful for identifying and studying functionally and structurally important regions [9]. Sequence conservation of homologous sequences is rarely homogeneous along

*Address correspondence to this author at the Department of Chemistry & Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110 016, India; Tel: 91-11-2659 1505; Fax: 91-11-2658 2037; E-mail: bjayaram@chemistry.iitd.ac.in

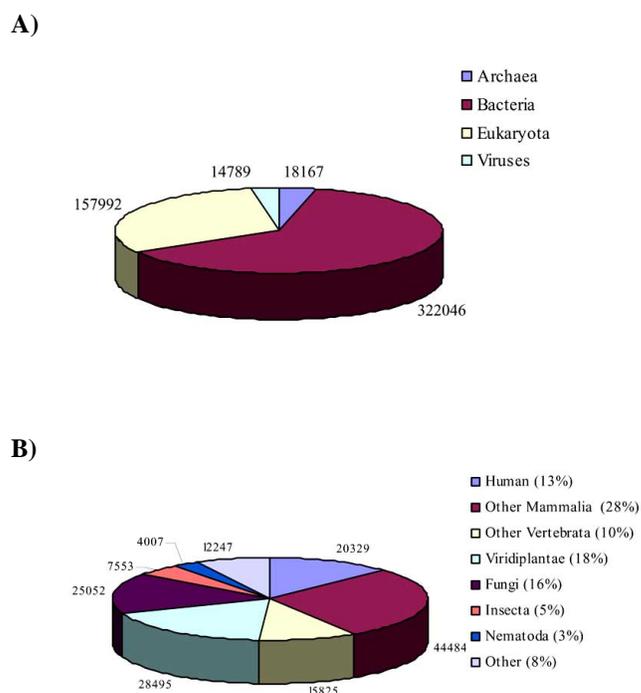


Fig. (1). A) Taxonomic Distribution of 512994 sequences (<http://au.expasy.org/sprot/>). B). Distribution of the sequences within eukaryota

their length; as sequences diverge, their conservation is localized to specific regions [9]. In order to obtain the general structural features of conserved regions of all proteins, it is necessary to decide the scale of protein clustering, conserved

regions and structural features to analyze [9]. Natural choices are generically defined protein families [10], ungapped protein sequence motifs (blocks) that separate proteins into either conserved or random signals [11] and the four basic secondary structure elements namely alpha helices, beta strands, structured turns, and loops [12].

Relations between protein sequence and structure can be analyzed by either determining the sequence features of pre-defined structures [13], or by determining structural features of conserved sequence regions. Han and Baker studied local structural features that predominate short sequence motifs, identifying correlations between specific sequence and structure motifs [14]. Secondary structure conservation was previously studied in structural alignments of protein families and secondary structure element substitution matrices were created [15]. The conservation of secondary structure element was also studied in some specific protein families [16]. Helices and strands have a regular repetitive structure [17], and this suggests that helices and strands are conserved [9]. Protein loops and their flanking regions were also found to be conserved to the same extent in an analysis of a large set of proteins [18]. Proteins with similar sequences adopt similar structure [19, 20]. However, similar structures can have less than 12% sequence similarity [21, 22-24].

Protein sequence comparison has become one of the most powerful tools for characterizing protein sequences because of the enormous amount of information that is preserved throughout the evolutionary process. One of the early attempts to measure protein sequence comparison was substitution matrices introduced by Dayhoff [25, 26]. A general approach for functional characterization of unknown proteins is to infer protein functions based on sequence similarity. One of the successful approaches is to define signatures of known families of biologically related proteins (typically at the functional or structural level). Signatures usually identify conserved regions among the family of proteins, revealing the importance for the function of their structural or physico-chemical properties. A representative example of this approach is the well-known Prosite database [27], gathering protein sequence patterns and profiles for a large number of families.

In recent years a number of different classification systems have been developed to organize proteins [9]. Among the variety of classification schemes are: (1) hierarchical families of proteins, such as the superfamilies/families [28] in the PIR-PSD, and protein groups in ProtoMap [29]; (2) families of protein domains, such as those in Pfam [30] and ProDom [31]; (3) sequence motifs or conserved regions, such as in PROSITE [32] and PRINTS [33]; (4) structural classes, such as in SCOP [34] and CATH [35]; as well as (5) integrations of various family classifications, such as iProClass [36] and InterPro [37].

The PIR superfamily/family concept [38] the original such classification based on sequence similarity, is unique in providing comprehensive and non-overlapping clustering of protein sequences into a hierarchical order to reflect their evolutionary relationships. Proteins are assigned to the same superfamily/family only if they share end-to-end sequence similarity, including common domain architecture (i.e. the same number, order, and types of domains), and do not differ

excessively in overall length (unless they are fragments or result from alternate splicing or initiators). Other major family databases are organized based on similarities of domain or motif regions alone, as in Pfam and PRINTS. There are also databases that consist of mixtures of domain families and families of whole proteins, such as SCOP and TIGRFAMs [39]. However, in all of these, the protein-to-family relationship is not necessarily one-to-one, as in PIR superfamily/family, but can also be one-to-many. The PIR superfamily classification is the only one that explicitly includes this aspect, which can serve to discriminate between multidomain proteins where functional differences are associated with presence or absence of one or more domains. Family and superfamily classification frequently allow identification or probable function assignment for uncharacterized (hypothetical) sequences. To assure correct functional assignments, protein identifications must be based on both global (whole protein, e.g. PIR superfamily) and local (domain and motif) sequence similarities [40].

3. PROTEIN TERTIARY STRUCTURE PREDICTION

Function follows form [41] and hence the need for structures. Stated alternatively, sequence to consequence [42] is the major challenge in proteomics investigations. The potential for protein tertiary structure prediction is nearly as vast as the diversity of biology itself. Folded proteins have applications in the area of sugar, chocolate, paper and pulp and textile and leather industry; *de novo* design of biocatalysts and in the area of nanobiomachines, nanofibres and quantum dots [43]. Determining the three-dimensional structure of protein molecules is a cornerstone for many aspects of modern biological research [44]. Currently close to half a million protein sequences are deposited in the UniProtKB/Swiss-Prot protein knowledgebase [1] but only ~ 61, 000 of them have experimentally solved structures [45] (Fig. 2). These numbers can be frustrating to molecular and cell biologists who need 3D models of proteins for their research. The high demand of the community for protein structures has placed computer-based protein structure prediction, the only means to rapidly alleviate the problem, at an unprecedentedly crucial position [44].

A long standing goal of Computational Biology has been to devise a computer algorithm that takes, as input, an amino acid sequence and gives, the three dimensional native structure of a protein as an output [46]. The main motivation is to understand function at a molecular level besides making drug discovery faster and more efficient by replacing slow and expensive structural biology experiments with fast and less expensive computer simulations [46]. A major milestone in computer-based native structure prediction is the creation of CASP (Critical Assessment of Techniques for Structure Prediction) by John Moult [47]. In the CASP experiments, research groups apply their prediction methods to amino acid sequences for which the native structure is not known but to be determined and to be published soon. These competitions provide a good measure to benchmark methods and progress in the field in an arguably unbiased manner [48].

Computational methods for protein tertiary structure prediction can be classified into four groups: (a) comparative modeling (Table 1), (b) fold recognition (Table 2), (c) first

principles methods with database information (Table 3), and (d) first principles methods without database information. Comparative modeling relies on the principle that sequences, which are related evolutionarily, exhibit similar three dimensional folded structures that is sequence similarity suggests structural similarity [49]. The accuracy of predictions by comparative modeling depends on the degree of sequence similarity. If the target and the template sequence have more than 50% of their sequences similar, predictions are of very good to high quality and have been shown to be as accurate as low-resolution X-ray predictions [50]. For 30-50% sequence identity, more than 80% of the C α -atoms can be expected to be within 3.5 Å of their true positions [50], while for less than 30% sequence identity, the prediction is likely to contain significant errors [50, 51]. A recent advance for automated comparative modeling is the TASSER-Lite tool, which is based on an extension of the TASSER approach [52].

Fold recognition and threading methods aim at fitting a target sequence to a known structure in a library of folds and the model built is evaluated using residue based contact potentials [49]. The method combines three different pair potentials to account for the fact that different scoring functions are capable of assigning different target sequences to the same template. By identifying structurally similar regions in multiple templates, accurate regions of structure prediction can be distinguished from less accurate ones. Skolnick and co-workers developed and successfully applied threading methods in the CASP 5-7 experiments [53, 54].

First principles methods that utilize database information can be further classified as (1) fragment-based recombination methods; (2) hybrid methods that combine multiple sequence comparison, threading, MC optimization with scoring functions, and clustering; and (3) methods that combine information from secondary structure and selected tertiary restraints with MC optimization or deterministic global optimization [49]. In fragment based recombination methods, the fundamental principle is that sequence-dependent local interactions direct the chain to sample specific sets of local conformers, which are compatible with the biased local conformers. Baker and co-workers [61] studied the distributions of local structures based on short sequence segments of up to 10 residues based on the protein database, and developed effective approaches that compare fragments of a target to fragments of known structures. Once appropriate fragments have been identified, they are assembled to a structure, often with the aid of scoring functions and optimization algorithms. In hybrid methods, Skolnick, Kolinski and co-workers [46, 62-64] developed approaches that combine multiple sequence comparison, threading, optimization with scoring functions, and clustering. The method uses a reduced representation lattice model with three or fewer atoms per residue [65]. The hierarchical approach TASSER [54] that combines template identification through threading, parallel hyperbolic MC sampling structure assembly via rearranging continuous template fragments, clustering using SPICKER [66], and post-analysis using the TM scores, was introduced and applied in CASP6 and CASP7.

First principles protein structure prediction methods aim to use purely physics-based methods, without knowledge

derived from databases (such as statistical energy functions or secondary structure predictors), to explore native structures and folding processes [46]. This class of methods can be applied to any given target sequence using only physically meaningful potentials, atomic level representations [49] and united residue representations [71]. Once 'physics-only' or 'physics-mainly' approaches succeed, the advantages would be: the ability to predict conformational changes, such as induced fit, a common and important unsolved problem in computational drug discovery; the ability to understand protein mechanisms, motions, folding processes, conformational transitions and other situations in which protein behavior requires more than just knowledge of the static native structure; the ability to design synthetic proteins for new applications or to design foldable polymers from non-biological backbones; and the ability to systematically improve protein modeling based on the laws of physics [46].

The *ab initio* methods utilize first principles to predict the three dimensional structure of proteins. These methods perform iterative conformational changes and estimate the corresponding changes in energy. Two main issues to be tackled for a successful prediction of protein structure are generation of a vast number of conformations and accurate scoring functions.

The *ab initio* methodologies have been developed along two lines. The first strategy tries to mimic the folding of proteins under physical conditions similar to those observed in nature. It involves simulating the protein-folding pathway by solving the Newton's equation of motion (molecular dynamics) [72] whereby a conformation corresponding to the global minimum of an appropriate potential energy surface is searched [73]. Early studies in this area involved the use of simplified lattice based [53, 74-76] or reduced representation [77, 78] models of proteins for carrying out simulations. These studies were carried out in vacuum. With time, increase in the computational power and efficiency allowed an all atom protein molecule to be simulated with the implicit treatment of the solvent [79, 80]. Current day compute capacities allow for microsecond and sub millisecond long simulations of the protein molecule with the solvent being treated explicitly [81-84].

The second strategy involves the generation of a number of conformations followed by evaluation of the models to determine the native-like structures. Various methods have been employed to sample the configurational space by systematic [85] or random searches in Cartesian or dihedral space [86-89]. Genetic algorithms [90], orthogonal latin squares [91, 92], distance geometry [93, 94] and hybrid search methods [95-97] have been employed as conformational search methods. Hierarchical approaches [98, 99], simulated annealing [100-103], replica exchange [104], parallel tempering [105], Monte Carlo methods [106, 107], build up procedures [108-110] and optimization of the scoring functions [111] have also been used for sampling of the conformational space of proteins.

The *ab initio* methods are rigorous in calculations but are limited by the compute power and time involved which emphasizes the need for faster structure prediction methods. Also, the accuracy of these methods is dependent upon the potential energy functions utilized during simulations [73].

Table 1. Some Popular Web Servers Available Freely Over the Internet for Homology Modeling

Sl. No.	Name of the Web Server (URL)	Description
1	CPHModels2.0 [55] (http://www.cbs.dtu.dk/services/CPHmodels/)	An automated protein structure homology-modeling server.
2	Swiss-Model [56] (http://swissmodel.expasy.org/SWISS-MODEL.html)	A fully automated protein structure homology-modeling server.
3	EsyPred3D [57] (http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/)	An automated server where the alignment is performed via a new alignment strategy using neural networks. Alignments are obtained by combining, weighting and screening the results of several multiple alignment programs. The final structure is built using the modeling package MODELLER.
4	ModWeb [58] (http://alto.compbio.ucsf.edu/modweb-cgi/main.cgi)	A web server implementation of MODELLER (comparative protein structure modeling by satisfaction of spatial restraints).
5	Geno3D [59] (http://geno3d-pbil.ibcp.fr/)	Comparative protein structure modelling by spatial restraints (distances and dihedral satisfaction).
6	3DJigSaw [60] (http://www.bmm.icnet.uk/servers/3djigsaw/)	An automated server to build three-dimensional models for proteins based on homologues of known structure.

Table 2. Some Publicly Available Servers for Fold Recognition

Sl. No.	Name of the Web Server (URL)	Description
1.	I-TASSER [54] (http://zhang.bioinformatics.ku.edu/I-TASSER/)	An internet service for protein structure and function predictions. Models are built based on multiple-threading alignments by LOMETS and iterative TASSER simulations.
2.	Threader [67] (http://bioinf.cs.ucl.ac.uk/threader)	Physically "thread" a sequence of amino acid side chains onto a backbone structure (a fold) and evaluates this proposed 3-D structure using a set of pair potentials and a separate solvation potential.
3.	LOOPP [68] (http://cbsuapps.tc.cornell.edu/loopp.aspx)	A fold recognition program based on the collection of numerous signals, merging them into a single score, and generating atomic coordinates based on an alignment into a homologue template structure. The signals we are using include straightforward sequence alignment, sequence profile, threading, secondary structure and exposed surface area prediction.
4.	GenTHREADER [69] (http://bioinf.cs.ucl.ac.uk/psipred/)	A combination of methods such as sequence alignment with structure based scoring functions and neural network based jury system to calculate final score for the alignment.
5.	LIBRA [70] (http://libra.ddbj.nig.ac.jp/top-e.html)	The target sequence and 3D profile are aligned by simple dynamic programming. According to the alignment, sequence is remounted on the structure and its fitness is evaluated by pseudo-energy potential.

In contrast, to the *ab initio* methods are the *de novo* methods, which utilize both the *ab initio* strategies as well as the database information (directly or indirectly).

The Robetta web server by Baker's group builds a multitude of protein structures from fragments of proteins [112, 113]. This is followed by clustering the final conformations and selection of representative structures of large clusters as final models. The ProtInfo web server by Samudrala *et al.* [114] predicts protein tertiary structure for sequences < 100 amino acids using *de novo* methodology, where by structures are generated using simulated annealing search phase which

minimizes a target scoring function. Scratch web server by Baldi *et al.* [115] predicts the protein tertiary structure as well as structural features starting from the sequence information alone. Astro-fold [116] an *ab initio* structure prediction framework by Klepeis and Floudas employs local interactions and hydrophobicity for the identification of helices and beta-sheets respectively followed by global optimization, stochastic optimization and torsion angle dynamics. *De novo* structure prediction by simfold energy function with the multi-canonical ensemble fragment assembly has been developed by Fujitsuka *et al.* [117]. The function has been

tested on 38 proteins along with the fragment assembly simulations and predicts structures within 6.5 Å RMSD of the native in 12 of the cases. The *Bhageerath* web server by Jayaram *et al.* [118-120] predicts 5 candidate native like structures starting with the protein sequence and secondary structure information. The *ab initio* prediction of 3D Structure of proteins where the inter residue distances are treated as random variables and the corresponding probabilities are estimated by nonparametric statistical methods and knowledge-based heuristics. has been formulated as Propainor algorithm by Joshi *et al.* [121].

The *ab initio* / *de novo* methods can be successfully employed in the design of novel protein folds [82, 122-124]. These methods are being currently employed in the large scale genome annotation projects of small genomes and are expected to have a large impact on the future of structural and molecular biology.

Based on recent CASP events (CASP 5, 6 7 and 8), it becomes evident that the first principles methods that utilize database information, more specifically, the fragment-based methods (Baker and co-workers) and the hybrid methods (Skolnick and co-workers; Zhang and co-workers) are at present leading in consistency for successful predictions primarily for medium resolution structures and for a few high resolution structures. Despite several successful medium resolution blind predictions, it is also apparent that significant advances are needed for consistent medium resolution predictions particularly in the difficult domain of free modeling, when no structurally similar templates can be successfully identified [49].

Our understanding of the folding mechanisms has also been advanced by theory and simulations. A recent view of the protein folding mechanism is the energy landscape model

[125, 126]. According to this model, all folding protein molecules are guided by an energy bias to traverse an energy landscape towards the native conformation. The concomitant decrease in conformational entropy leads to a funnel-shaped energy landscape. The road to the native state from the vast majority of individual non-native conformations is downhill and is different for each non-native starting conformation. Many different folding trajectories for individual protein molecules are envisaged and hence, multiple folding pathways are expected to be operative. Intermediates, when present, are considered as kinetic traps which slow down the folding reaction [127].

It is useful to examine the different models of protein folding in the context of how proteins begin their search for the native conformation. The nucleation model appears inapplicable to folding reactions, because it does not predict the early intermediate forms seen during the folding of many proteins [128]. Secondary structural elements do not appear to form unless some stabilizing tertiary contacts are made. Hence, a framework model is also unlikely to be a common mechanism by which proteins fold [129]. By contrast, the observation that a fast (sub-ms) collapse reaction precedes the formation of a secondary structure during the folding reactions of several proteins [130-132], suggests that many proteins indeed fold by the hydrophobic collapse mechanism. Nonetheless, the simultaneous occurrence of collapse and structure formation in the case of a few apparently two-state folding proteins [133, 134] is difficult to explain by the classical hydrophobic collapse mechanism. For such proteins, the nucleation-condensation mechanism may better describe how folding occurs [135, 136].

Protein folding no longer appears to be an insurmountable grand challenge. Current knowledge of folding codes is

Table 3. A Few *de novo* Web Servers Available in the Public Domain for Protein Tertiary Structure Prediction

Sl. No.	Name of the Web Server/Group (URL)	Description
1.	ROBETTA [112, 113] (http://robetta.bakerlab.org)	<i>De novo</i> Automated structure prediction analysis tool used to infer protein structural information from protein sequence data.
2.	PROTINFO [114] (http://protinfo.compbio.washington.edu)	<i>De novo</i> protein structure prediction web server utilizing simulated annealing for generation and different scoring functions for selection of final five conformers.
3.	SCRATCH [115] (http://www.igb.uci.edu/servers/psss.html)	Protein structure and structural features prediction server, which utilizes recursive neural networks, evolutionary information, fragment libraries and energy.
4.	ASTRO-FOLD [116]	Astro-fold: first principles tertiary structure prediction based on overall deterministic framework coupled with mixed integer optimization.
5.	ROKKY [117] (http://www.proteinsilico.org/rokky/rokky-p/)	<i>De novo</i> structure prediction by the simfold energy function with the multi-canonical ensemble fragment assembly.
6.	BHAGEERATH [118-120] (http://www.scfbio-iitd.res.in/bhageerath)	Energy based methodology for narrowing down the search space of small globular proteins.
7	PROPAINOR [121]	The <i>ab initio</i> prediction of 3D Structure of proteins where the inter residue distances are treated as random variables and the corresponding probabilities are estimated by nonparametric statistical methods and knowledge-based heuristics.

sufficient to guide the successful design of new proteins and new materials. Current computer algorithms are now predicting the native structures of small simple proteins remarkably accurately, contributing to drug discovery and proteomics [46].

4. MEMBRANE PROTEIN BIOINFORMATICS

Membrane proteins are crucial players in the cell and take centre stage in processes ranging from basic small-molecule transport to sophisticated signaling pathways [137]. Many are also prime contemporary or future drug targets, and it has been estimated that more than half of all the drugs currently on the market are directed against membrane proteins, which are responsible for the uptake, metabolism, and clearance of these pharmacologically active substances [138, 139]. Although analyses show that ~30% of the proteins coded in human genome are membrane proteins [140], it is still frustratingly hard to obtain high-resolution three-dimensional structures of membrane proteins [137], (Table 4). Even if the number of experimentally known membrane protein structures is on the rise [141, 142], methods for predicting the three dimensional structures of membrane proteins will need many more years. Therefore there exists enormous incentive for computational and theoretical studies of membrane proteins.

4.1. What the Sequences Tell

For the helix-bundle membrane proteins, the typical transmembrane segment is formed by a stretch of predominantly hydrophobic residues long enough to span the lipid bilayer as an α -helix [149-153]. The early topology prediction methods were consequently little more than plots of the segmental hydrophobicity (averaged over 10-20 residues)

along the sequence [154-156]. With more sequences came the realizations that aromatic Trp and Tyr residues tend to cluster near the ends of the transmembrane segments [157,158] and that the loops connecting the helices differ in amino acid composition, depending on whether they face the inside or outside of the cell [159-161]. More recent analyses have focused on the higher-than-random appearance of sequence motifs, such as the GxxxG-motif in transmembrane segments [162, 163] as well as other periodic patterns within the membrane helices [164], with the aim of providing information that may help in predicting helix-helix packing and 3D structure.

4.2. Topology Prediction Methods

There are a number of different topology prediction methods available today. All methods rely on hydrophobicity analyses to predict the number of TMHs and most methods use the positive-inside rule to deduce the orientation of the protein relative to the membrane. The aromatic residues Tryptophan and Tyrosine are often incorporated into the algorithm in order to better define the boundaries of the TMHs. The topology prediction methods then generate a topology model that includes: (i) how many TMHs the protein has, (ii) on which side of the membrane the loops and tails are located, (iii) the boundaries of the membrane and non-membrane domains [165]. TMHMM has been ranked as one of the top-performing topology predictors in several evaluations, with a success rate of 55-75% [166-170].

4.3. 3D Predictions

For membrane proteins, prediction methods allow us to predict (i) if a protein belongs to the class of α -helical membrane protein, (ii) topology models, (iii) re-entrant loops, (iv)

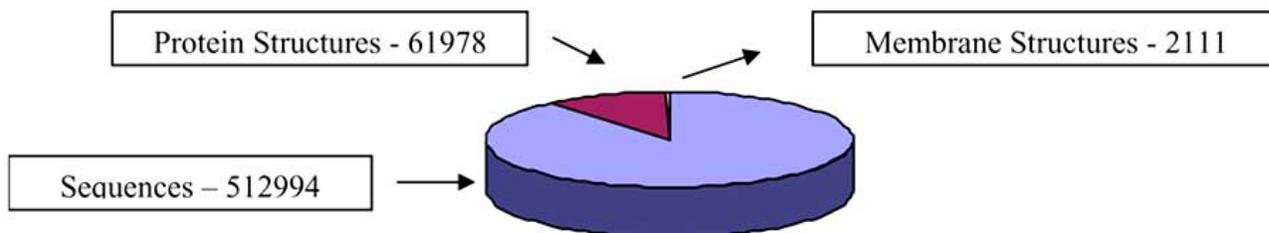


Fig. (2). Statistics from the Protein Data Bank on Proteins.

Table 4. Some Useful Membrane Protein Structure Resources

SI No	Membrane Protein Structure Resources	URL
1	Progress of membrane protein structure determination [143].	http://blanco.biomol.uci.edu/MP_Structure_Progress.html
2	Martin Caffrey's Membrane Protein Data Bank [144].	http://www.mpdb.ul.ie/
3	Bilayer Insertion of Membrane Proteins [145].	http://sbc.bioch.ox.ac.uk/cgdb/
4	Protein Data Bank of Transmembrane Proteins [146, 147].	http://pdbtm.enzim.hu/
5	TMFunction: database for functional residues in membrane proteins [148]	http://tmbeta-genome.cbrc.jp/TMFunction

presence of a signal sequence and (v) organelle localization [137]. To date there is no reliable *ab initio* 3D prediction method that is publicly available and no straight forward schemes to apply to membrane proteins because the different environment introduced by the membrane has to be modeled in some way, and because most membrane proteins are significantly larger than the globular proteins predicted so far [137]. Therefore, currently there is no general algorithm that works from protein sequence information only.

An interesting attempt to model all G-protein-coupled receptors (GPCRs) of the human proteome was made by Skolnick and coworkers [140] using the TASSER algorithm. Although the accuracy of the predicted rhodopsin structure was quite good, the correctness of the GPCR structures cannot be verified until more structures are available [124]. ROSETTA membrane folding algorithm was used to model the closed and open states of a voltage-dependent potassium channel [171].

While *ab initio* structure modeling can best predict the overall fold of a protein, homology modeling of membrane proteins is still in its infancy because very few structures are known. When a template is available, homology models of membrane proteins are comparable in quality to those that can be made for globular proteins, i.e., when the sequence identity between the template and the target is >30%, one can expect the root mean-square deviation between the modeled and the correct structure to be < 2 Å in the transmembrane regions [172].

5. HUMAN PROTEOME

Completion of sequencing of the human genome [173, 174] has ushered in an era of characterizing genes and their gene products or proteins in great detail. The Human Protein Reference Database (HPRD) is a novel comprehensive protein information resource that depicts various features of proteins such as domain architecture, post-translational modifications, tissue expression, molecular function, subcellular localization, enzyme-substrate relationships and protein-protein interactions [175, 176]. With the inclusion of most of the human protein sequences, HPRD, a community driven database has grown into an integrated knowledgebase for genomic and proteomic investigators. [176]. This database will assist in biomedical discoveries by serving as a resource of genomic and proteomic information and provid-

ing an integrated view of sequence, structure, function and protein networks in health and disease [175] (Fig. 3).

6. PROTEIN-PROTEIN INTERACTIONS

Protein-protein interactions (PPI) are essential for almost all cellular functions. Proteins seldom carry out their function in isolation; rather, they operate through a number of interactions with other biomolecules. Experimental elucidation and computational analysis of the complex networks formed by individual protein-protein interactions (PPIs) is one of the major challenges in the post-genomic era. Protein-protein interaction databases have become a major resource for investigating biological networks and pathways in cells [177]. A substantial fraction of eukaryotic proteins contains multiple domains, some of which show a tendency to occur in diverse domain architectures and are considered mobile or promiscuous. These promiscuous domains are typically involved in protein-protein interactions and play crucial roles in interaction networks, particularly those contributing to signal transduction [178]. Studies of protein-protein interaction networks across species have strong potentiality in the field of molecular evolution since protein-protein interactions are central for function and control and as well as reflect cohesive efforts in the organization of the complicated interactomes [179, 180]. The protein-protein interaction networks reveal that most of the proteins (the network nodes) are connected to relatively fewer, highly connected proteins termed as the hub proteins [176]. Further studies revealed that in a protein-protein interaction network, hub proteins that physically interact with most or all of their partners simultaneously are designated as party hubs and those that bind their different partners at different times or locations are the date hubs [181]. Protein-protein interfaces are highly attractive targets for drug discovery because they are involved in a large number of disease pathways where therapeutic intervention would bring widespread benefit [182].

A very recent review article by Tuncbag *et al.*, [183] summarizes the available tools and web servers for analysis of protein-protein interactions and interfaces. This review provides a comprehensive and organized list of the available databases and web servers of protein folding sites and their characteristics outlining how the tool was constructed, its advantages and drawbacks. These resources can be used to analyze the physico-chemical properties of interfaces and

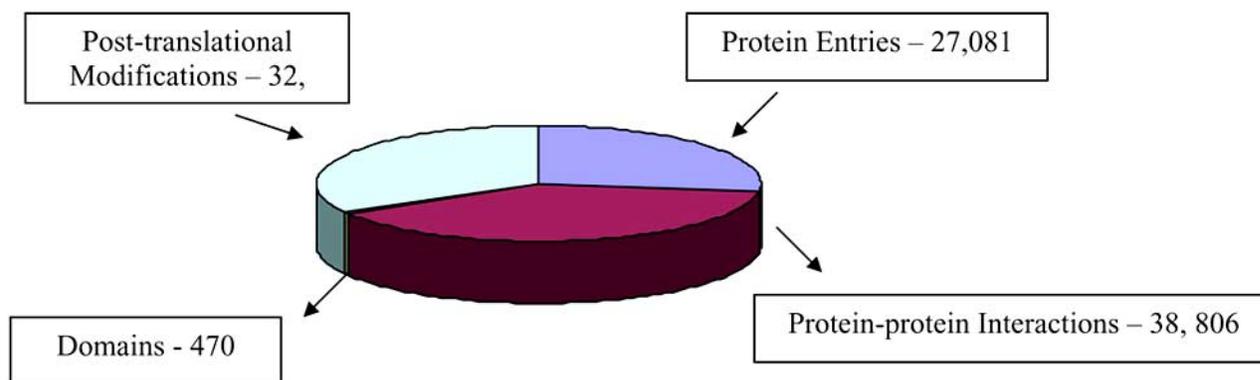


Fig. (3). Statistics in the Human Protein Reference Database (<http://www.hprd.org/>).

differentiate between biological complexes and crystal contacts and predict binding sites in protein structures and of the docked structures of two individual proteins. A combination of such resources is expected to help biologists explore protein interactions, relate these to cellular processes and design drugs to target the druggable sites.

MPIDB (Microbial protein interaction database) is a new web resource, which is a repository of all known physical microbial interactions [184] that provides unified access to available microbial interaction data. The microbial interactions have been manually curated from the literature or imported from other databases. Interactions in MPIDB are further supported by 8150 evidences based on interaction conservation, co-purification and 3D domain contacts.

6.1. Flexibility in Protein-Protein Interactions

Flexibility is of overwhelming importance for protein function, and the changes in protein structure during interactions with binding partners can be dramatic [185]. The flexibility of a protein may result in either subtle changes as when a few amino acid side chains of an enzyme move to bind a small substrate, or in more dramatic changes as when the folding of certain proteins is facilitated by the presence of the appropriate ligand. An inverse relationship between protein stability and the biological function of both enzymes and protein hormones has been described, underscoring the fact that function necessitates flexibility [186, 187]. In terms of medicinal chemistry and drug discovery, even the angiotensin-converting enzyme inhibitor captopril, credited as the first drug discovered using a protein-binding site, binds to a protein (carboxypeptidase A) that was known to be highly flexible [188].

7. METABOLIC NETWORKS

There is current interest in the processes underlying the biology of network because these offer insight into the organization and evolution of life [189]. Elucidation of cellular metabolism, one of the greatest achievements of science, is clearly the best-studied biological network. It represents a complex collection of enzymatic reactions and transport processes that convert metabolites into molecules capable of supporting cellular life [190]. A very recent study has uncovered the origins and evolutionary patterns of modern metabolism. Using phylogenomic information linked to the structure of metabolic enzymes, Mittenthal and his coworkers have sorted out recruitment processes and discovered that most enzymatic activities were associated with the nine most ancient and widely distributed protein fold architectures. Their analysis of newly discovered functions showed enzymatic diversification occurred early, during the onset of the modern protein world. Their observation of phylogenetic reconstruction exercises, strongly suggested that metabolism originated in enzymes with the P-loop hydrolase fold in nucleotide metabolism, probably in pathways linked to the purine metabolic network [191].

The most challenging issue in life sciences today is the study of metabolic pathways for the identification of suitable drug targets against infectious agents. Association of targets with less number of metabolic pathways tends to reduce the chance of unwanted interference with other processes and these targets are more likely to be successfully discovered

and explored for generating a higher number of clinical drugs [192].

8. POTENTIAL DRUG TARGETS BASED ON SIMPLE SEQUENCE PROPERTIES

Although great efforts have been exerted on drug research and development during the past decades, ~324 drug targets have been identified for clinically useful drugs to date [193], which indicates that current pharmaceutical industry actually relies on only a small pool of drug targets, compared to the large number of proteins available in human genome [194] and those of the pathogens. A significant number of drugs that fail in the pipeline of modern drug discovery can be attributed to the wrong drug target definition at the early preclinical stages [195]. In a recent study, a drug target prediction method based on support vector machine has been developed. Independent of homology annotation and protein 3D structures, the current method employs simple physico-chemical properties based on primary protein sequence to construct the SVM model. The method can successfully distinguish known drug targets from putative non drug targets at an accuracy of 84% in 10-fold cross-validation test. Limitations of this method are, only human proteins are covered and only protein drug targets are taken into account [182]. Bhakeet and Doig [195] used wider range of sequence properties and reported eight key properties of human drug targets that differed significantly from non-drug targets and applied these to identify new potential drug targets. The properties have been summarized as druggability rules in the Table 5. They then used support vector machines to make a classifier to distinguish targets from non-targets using the eight key features calculated from protein sequences. Their method identified 23% of the human proteins with target-like properties, some of which have been annotated by primary EC number, giving 17 oxidoreductases, 12 transferases, 44 hydrolases, 6 lyases, 5 isomerases and no ligases.

Structure-based drug design is playing a growing role in modern drug discovery, with numerous approved drugs tracing their origins, at least in part, to the use of structural information from X-ray-crystallography or nuclear magnetic resonance analysis of protein targets and their ligand-bound complexes [196]. Protein structure information is the bread and butter of structure-based drug discovery. An explosion in technological and computational advances in structural genomics projects have substantially increased the number of protein structures of hundreds or thousands of medically relevant targets from infectious disease organisms. This new information provides both academic and for-profit scientists with an unprecedented opportunity to accelerate the development of new and improved chemotherapeutic agents against these pathogens [196].

One of the major challenges in drug development is the accurate assessment of human drug toxicity [197]. Given the very high attrition rates in drug discovery besides the cost and time factors [198], there is also an ethical issue of causing harm to the patient population [197]. For that reason, in recent years pharmaceutical companies have brought toxicity testing, as well as ADME (absorption, distribution, metabolism, excretion) studies, early on in the drug development process [199]. The ultimate here would be to use *in silico*

Table 5. Sequence Properties as Druggability Rules

Property	p(Target)	p(Non-Target)
Hydrophobicity > -142.4	0.66	0.60
Length > 550 amino acids	0.39	0.29
SignalP motif present	0.45	0.27
No PEST motif	0.21	0.35
More than 2 N-glycosylated amino acids	0.52	0.38
Not more than one O-glycosylated Ser	0.16	0.24
Mean pI < 7.2	0.37	0.51
Membrane location	0.49	0.24

methods to predict toxicity even before a drug candidate is being synthesized [199].

The existing commercially available *in silico* tools for predicting potential toxicity issues can be roughly classified into two groups. The first group uses expert systems that derive models on the basis of abstracting and codifying knowledge from human experts and scientific literature. The second group relies primarily on the generation of descriptors of chemical structures and statistical analyses of the relationships between the descriptors and toxicological endpoints [200].

9. DISORDERED PROTEINS

Interest in disordered proteins has swelled as a result of the realization that such proteins are unexpectedly and perhaps astonishingly common in human and other genomes [201-203]. Disordered proteins are associated with a variety of biological functions, many of them intimately related to human disease [204-206]. Database analysis indicates that proteins that are involved in eukaryotic signal transduction or that are associated with cancer have an increased propensity for intrinsic disorder [207]. A signature of probable intrinsic disorder is the presence of low sequence complexity and amino-acid compositional bias, with a low content of bulky hydrophobic amino acids (Val, Leu, Ile, Met, Phe, Trp and Tyr), and a high proportion of particular polar and charged amino acids (Gln, Ser, Pro, Glu, Lys, Gly and Ala) [208, 209]. The presence of such regions in transcriptional regulatory proteins was recognized more than 25 years ago [210]. Many of these regions function in transcriptional activation and they are often classified according to their amino acid composition – for example, there are glutamine-rich, proline-rich and acidic activation domains [211]. A number of computer programs are now available for the prediction of unstructured regions from amino acid sequences and Table 6 summarizes some popular disorder predictors, their URL addresses and the principles they are based on.

Many disordered proteins do adopt more highly ordered conformations upon interactions with other cellular components [222]. A role for induced protein folding in sequence-specific DNA binding was proposed more than a decade ago by Spolar and Record [223], on the basis of the large heat-

capacity changes that result from DNA-protein complex formation. Some of these processes involve the large-scale folding of entire domains – for example, the basic region of the basic leucine-zipper (bZip) DNA-binding domain [224] – whereas others involve the folding of local disordered loops or linkers between folded domains [225]. Many RNA-binding proteins also contain unstructured regions [226]. For example, the ribosomal protein L5 seems to associate with 5S ribosomal RNA by mutual induced-fit mechanism: both RNA and protein are significantly more structured in the complex than in the free state [227].

10. THE SEQUENCE-STRUCTURE RELATIONSHIP AND PROTEIN FUNCTION PREDICTION

The knowledge of the relationship between structure and function, combined with a rise in the number of structures solved with no biochemical annotations, has motivated the development of computational tools for the prediction of molecular function using sequence and structural information [228, 229]. Despite methodological improvements in this area, determining function directly from tertiary structure has proven to be a difficult problem to crack [230]. Much of the problem in assigning function from structure comes from functional convergence, where although a stable structure is required to perform many functions, it is not always necessary to adopt a particular structure to carry out a particular function – any one of the several folds might be equally as good as any other [231-235]. Despite the complexities unique to proteins, the principal means by which protein function is defined remains rooted in a set of labels derived from gene ontology [221]. The most significant limitation of gene ontology annotations when applied to proteins is that GO terms are non-positional and there is no defined relationship between, metal-binding sites in crystal structure and concomitant GO terms [230]. The Protein Feature Ontology (PFO [236]) is an important new development that is likely to be very useful in bridging this gap. The PFO has been developed to provide a structured controlled vocabulary for features on a protein sequence or structure and comprises ~100 positional terms, now integrated into the Sequence Ontology (SO) and 40 non-positional terms which describes features relating to the whole protein sequence [236]. The

Table 6. Some Popular Disorder Predictors, Their URL Addresses and the Principles They are Based on

Sl. No.	Predictor	URL	Principle/Method
1	DisEMBL [212]	http://dis.embl.de	Neural Network
2	DISOPRED2 [213]	http://bioinf.cs.ucl.ac.uk/disopred	Support vector machines, neural networks
3	FoldUnfold [214]	http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi	Amino acid propensity
4	FoldIndex [215]	http://bip.weizmann.ac.il/fldbin/findex	Amino acid propensity
5	GlobPlot [216]	http://globplot.embl.de	Amino acid propensity, preference for ordered secondary structure
6	IUPred [217]	http://iupred.enzim.hu	Estimated pairwise interaction energy
7	NORSp [218]	http://cubic.bioc.columbia.edu/services/NORSp	Secondary structure propensity
8	PONDR VSL2 [219]	http://www.ist.temple.edu/disprot/predictorVSL2.php	Support vector machines with non linear kernel
9	PreLink [220]	http://genomics.eu.org/spip/PreLink	Amino acid propensity, hydrophobic cluster analysis
10	Ucon [221]	http://www.predictprotein.org/submit_uccon.html	Amino acid contact potential

Distributed Annotation System (DAS [237]) specification has been modified to incorporate PFO term codes along with accompanying evidence codes and represents an important step towards combining disparate function annotation sources which should enable many future developments in reliable function prediction [238].

11. CHEMICAL LOGIC OF PROTEIN SEQUENCES

Protein nucleic acid recognition and protein folding are among some of the unsolved problems in molecular biophysics/molecular biology and at the heart of these reside interactions involving amino acid side chains. It is likely that a new view of amino acid side chains may benefit this field. A novel chemical analysis of amino acid side chains reported recently [239] explains the chemical logic of protein sequences. That only 20 amino acids occur naturally accounting for the structural and functional diversity of proteins is a consequence of the action of a symmetry group. The analysis identifies the presence of hydrogen bond donor groups, presence of sp³ hybridized γ carbons, absence of δ carbons and linearity as properties central to side chain design and quantify the chemical logic of protein sequences. Naturally occurring proteins are dominated by amino acids with sp³ hybridized γ carbons and short side chains. From a chemistry point of view, the 20 amino acid side chains constitute a near comprehensive complete chemical template (monomer) library to build polymers with the diversity of functions. From a biology perspective, the 20 amino acids contain all the information necessary for the regulation of gene expression involving protein-nucleic acid interactions. It is conceivable that the chemical model presenting a clue to the language of

amino acids could facilitate a better understanding of the structure and function of proteins and structure based drug design efforts.

CONCLUSIONS

Despite much effort in structural genomics, the amount of protein structures determined by time consuming and expensive experimental methods is significantly smaller when compared to large-scale DNA sequencing methods. As genome-sequencing projects provide biologists with ready access to rapidly increasing pool of protein sequences, there will be a growing demand for developing advanced computational methods for predicting structure and function from sequence information only without knowing the structural data. Development of computational technologies that enable the complex web of relationships that characterize protein structure/function space will lead to a far more productive exploitation of the information contained in sequence/structure/function databases than is currently possible. The implications of such an approach on the impact of structural genomic initiatives could be enormous. Development of smarter algorithms and sophisticated automated computer modeling approaches will enlarge the scope of model-able proteins for structural genomics. Based on the recent CASP events (CASP 7, 8), a competition that has often been called as the “Olympic Games of Protein Structure Prediction” Zhang server constructed based on first principles and hybrid methods at present is leading in consistency for successful predictions primarily for medium resolution structures and for a few high resolution structures. The eventual solution to the problem is however free modeling con-

sidered to be a holy grail of the computational molecular biology. The best current free modeling softwares so far in business is from the Baker's group (University of Washington). However, free modeling methods are not advanced enough for routine medical applications as the models are not experimental structures determined with known accuracy but are predictions, which heavily depend on model quality.

Membrane proteins constitute ~30% of all proteins and reliable methods to predict the structure of membrane proteins are crucial as experimental determination of high resolution membrane protein structures remain very difficult given their complexity and size. The current understanding and representation of membrane protein topology as a simple string of membrane spanning α -helices and β -sheets does not fully capture the structural diversity observed in membrane proteins. Faster computational protocols to map membrane interactome should prove useful to guide and rationalize experimental investigations and a conglomeration of dry and wet lab approaches should hold a key to the answer.

New functional proteins are built on advances in modeling and structure prediction. Current computational design methodology can provide close to atomic resolution predictions [240] and can be useful in unanticipated ways, including improvement of catalytic efficiency, creating therapeutically useful proteins. Protein and peptide products for therapeutic use include a very diverse range of products as hormones, growth factors, cytokines, vaccines and monoclonal antibodies. There have been a number of successful examples in which computer-predicted models were used to guide the design of new drug [241]. Of note, Becker *et al.* used the predicted structural models of the serotonin receptors to screen a compound library [242].

A flood of data is emerging from genome research, including sequence data on proteins. To help science keep pace with this flow of knowledge, biophysicists, biochemists and structural biologists across the world have been developing tools and resources for management of data and the integration of information from varied sources. HPRD is an online repository for experimentally derived information about the human proteome. This rich resource can be browsed and searched for protein-protein interactions, post translational modifications and tissue expression. Knowledge of protein interactions is crucial for elucidating their functional role. It is also essential in correcting biological dysfunction related to diseases. One of the current drawbacks of PDB holdings is the low coverage of crystallized protein complexes, which makes structure based prediction methods for protein-protein interactions a pressing necessity. Metabolic networks have evolved to be exceptionally robust, adopting organizational structures. The relationship between metabolic structure and function is an important question for researchers in areas such as bioengineering and disease treatment, where one goal is to manipulate metabolic network structure in order to obtain desired behaviors. Drug target discovery has received much attention in both academia and pharmaceutical industry. With novel methods discovered everyday with high prediction accuracies for target identification from the sequence information alone would fasten modern drug discovery with reduction in time and cost.

Computational analyses for structure-function prediction based on sequence information are increasingly becoming an essential and integral part of modern biology. With rapid advances in the area, there is a growing need to develop efficient versatile bioinformatics software packages which are hypotheses driven. 'Gene to Drug' developed at SCFBio, IIT Delhi is an attempt in this pursuit and an integration of heterologous applications of different technologies developed in-house and their translation into *in silico* products that cater to a majority of bioinformatics applications and how grid services and high performance computing platforms can be harnessed to bridge the gap between biomolecular sequence, structure and function [243].

The topics presented here have glossed over many details. In seeking to prompt a fresh mindset, we were motivated to frame the whole picture in sweeping strokes for which we seek the indulgence of the readership.

ACKNOWLEDGEMENTS

SRS would like to thank DST for Women Scientist Fellowship. Funding from the DBT, DIT and DST-JST are gratefully acknowledged. Authors also thank Kazuhiko Fukui, Michael Gromiha, Kana Shimizu, Wataru Nemoto, Kumkum Bhushan, Priyanka Dhingra and Avinash Mishra for scientific discussions and Shashank Shekar for maintaining the web servers. The authors would like to thank the anonymous reviewers for invaluable suggestions for improving the manuscript.

ABBREVIATIONS

CASP	=	Critical assessment of methods of protein structure prediction
DAS	=	Distributed Annotation System
GO	=	Gene Ontology
GPCR	=	G protein coupled receptor
HPRD	=	Human Protein reference database
MD	=	Molecular Dynamics
MPIDB	=	Microbial protein interaction database
PDB	=	Protein data bank
Pfam	=	Protein families
PFO	=	Protein feature ontology
PIR	=	Protein information resource
PIR-PSD	=	Protein information resource – protein sequence database
PPI	=	Protein protein interactions
PRINTS	=	Compendium of protein fingerprints
ProDom	=	Protein domain
PROSITE	=	Database of protein domains, families and functional sites
ProtoMap	=	Automatic classification of protein sequences and hierarchy of protein Families

ROSETTA	=	Server for protein tertiary structure prediction
SCOP	=	Structural Classification of proteins
SO	=	Sequence ontology
TASSER	=	Server for protein tertiary structure prediction based on threading and assembly refinement
TMHs	=	Transmembrane helices
TMHMM	=	Server for prediction of transmembrane helices in proteins

REFERENCES

- [1] UniProtKB/Swiss-Prot Available at: <http://ca.expasy.org/sprot/relnotes/relstat.html>
- [2] Bernal, J.D.; Crowfoot, D. X-ray photographs of crystalline pepsin. *Nature*, **1934**, *133*, 794-795.
- [3] Perutz, M.F.; Rossmann, M.G.; Cullis, A.F.; Muirhead, H.; Will, G.; North, A.C. T. Structure of haemoglobin: a three dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature*, **1960**, *185*, 416-422.
- [4] Schuler, G.D.; Epstein, J.A.; Ohkawa, H.; Kans, J.A. Molecular biology database and retrieval system. *Methods Enzymol.*, **1996**, *266*, 141-162.
- [5] Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **1995**, *247*, 536-540.
- [6] Orengo, C.A.; Michie, A.D.; Jones, S.; Jones, D.T.; Swindells, M. B.; Thornton, J.M. CATH: a hierarchic classification of protein domain structures. *Structure (London)*, **1997**, *5*, 1093-1108.
- [7] Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature*, **1992**, *357*, 543-544.
- [8] Rohl, C.A.; Strauss, C.E.; Misura, K.M.; Baker, D. Protein structure prediction using rosetta. *Methods Enzymol.*, **2004**, *383*, 66-93.
- [9] Sitbon, E.; Pietrokovski, S. Occurrence of protein structure elements in conserved sequence regions *BMC Struct. Biol.*, **2007**, *7*, 1-15.
- [10] Henikoff, S.; Greene, E.A.; Pietrokovski, S.; Bork, P.; Attwood, T.K.; Hood, L. Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **1997**, *278*, 609-614.
- [11] Henikoff, J.G.; Greene, E.A.; Pietrokovski, S.; Henikoff, S. Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res.*, **2000**, *28*, 228-230.
- [12] Henikoff, S.; Henikoff, J.G.; Alford, W.J.; Pietrokovski, S. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, **1995**, *163*, GC17-26.
- [13] Bystroff, C.; Simons, K.T.; Han, K.F.; Baker D. Local sequence-structure correlations in proteins. *Curr. Opin. Biotechnol.*, **1996**, *7*, 417-421.
- [14] Han, K.F.; Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. USA*, **1996**, *93*, 5814-5818.
- [15] Mizuguchi, K.; Blundell, T. Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics*, **2000**, *16*, 1111-1119.
- [16] Cygler, M.; Schrag, J.D.; Sussman, J.L.; Harel, M.; Silman, I.; Gentry, M.K.; Doctor, B.P. Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins. *Protein Sci.*, **1993**, *2*, 366-382.
- [17] Rose, G.D. Secondary structure in protein analysis. In *Encyclopedia of Biological Chemistry* New York, Elsevier Inc; **2004**, pp. 1-6.
- [18] Liu, J.; Tan, H.; Rost, B. Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **2002**, *322*, 53-64.
- [19] Chothia, C.; Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **1986**, *5*, 823-826.
- [20] Doolittle, R.F. Similar amino acid sequences: chance or common ancestry? *Science*, **1981**, *214*, 149-159.
- [21] Hubbard, T.J.; Murzin, A.G.; Brenner, S.E.; Chothia, C. SCOP: a structural classification of proteins database. *C. Nucleic Acids Res.*, **1997**, *25*, 236-239.
- [22] Holm, L.; Sander, C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.*, **1996**, *24*, 206-209.
- [23] Brenner, S.E.; Chothia, C.; Hubbard, T.J.; Murzin, A.G. Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.*, **1996**, *266*, 635-643.
- [24] Rost, B. Protein structures sustain evolutionary drift. *Fold Des.*, **1997**, *2*, S19-24.
- [25] Dayhoff, M.O.; Eck, R.V.; Park, C.M. Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC, **1972**, Vol. 5, pp. 89-99.
- [26] Dayhoff, M.O.; Schwartz, R.M.; Orcutt, B.C. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Washington, DC, **1978**, Vol. 5(3), pp. 345-352.
- [27] Hulo, N.; Sigrist, C.J.; Le Saux, V.; Langendijk-Genevaux, P.S.; Bordoli, L.; Gattiker, E.; De Castro, Bucher, P.; Bairoch, A.P. Recent improvements to the PROSITE database *Nucleic Acids Res.*, **2004**, *32* (90001), 134-137.
- [28] Barker, W.C.; Pfeiffer, F.; George D.G. Superfamily classification in PIR-International protein sequence database. *Methods Enzymol.*, **1996**, *266*, 59-71.
- [29] Yona, G.; Linial, N.; Linial M. ProtoMap: Automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **2000**, *28*, 49-55.
- [30] Bateman, A.; Birney, E.; Cerutti, L.; Durbin, R.; Etwiller, L.; Eddy, S.R.; Griffiths-Jones, S.; Howe, K.L.; Marshall, M.; Sonnhammer, E.L.L. The Pfam protein families database. *Nucleic Acids Res.*, **2002**, *30*, 276-280.
- [31] Corpet, F.; Servant, F.; Gouzy, J.; Kahn D. Tools for protein domain analysis and whole genome comparisons, *Nucleic Acids Res.*, **2000**, *28*, 267-269.
- [32] Falquet, L.; Pagni, M.; Bucher, P.; Hulo, N.; Sigrist, C.J.A.; Hofmann, K.; Bairoch, A. The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **2002**, *30*, 235-238.
- [33] Attwood, T.K.; Blythe, M.J.; Flower, D.R.; Gaulton, A.; Mabey, J.E.; Maudling, N.; McGregor, L.; Mitchell, A.L.; Moulton, G.; Paine, K.; and Scordis, P. PRINTS and PRINTS-s shed light on protein ancestry. *Nucleic Acids Res.*, **2002**, *30*, 239-241.
- [34] Lo Conte, L.; Brenner, S.E.; Hubbard, T.J.P.; Chothia, C.; and Murzin, A.G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **2002**, *30*, 264-267.
- [35] Pearl, F.M.G.; Martin, N.; Bray, J.E.; Buchan, D.W.A.; Harrison, A.P.; Lee, D.; Reeves, G.A.; Shepherd, A.J.; Sillitoe, I.; Todd, A.E.; Thornton, J.M.; Orengo, C.A. The CATH extended protein family database: providing structural annotations for genome sequences. *Nucleic Acids Res.*, **2001**, *29*, 223-227.
- [36] Huang, H.; Barker, W.C.; Chen, Y.; Wu, C.H. ProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.*, **2003**, *31*, 390-392.
- [37] Apweiler, R.; Attwood, T.K.; Bairoch, A.; Bateman, A.; Birney, E.; Biswas, M.; Bucher, P.; Cerutti, L.; Corpet, F.; Croning, M.D.R.; Durbin, R.; Falquet, L.; Fleischmann, W.; Gouzy, J.; Hermjakob, H.; Hulo, N.; Jonassen, I.; Kahn, D.; Kapapin, A.; Karavidopoulou, Y.; Lopez, R.; Marx, B.; Mulder, N.J.; Oinn, T.M.; Pagni, M.; Servant, F.; Sigrist, C.J.A.; Zdobnov, E.M. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **2001**, *29*, 37-40.
- [38] Dayhoff, M.O. The origin and evolution of protein superfamilies. *Fed Proc.*, **1976**, *35*, 2132-2138.
- [39] Haft, D. H.; Loftus, B.J.; Richardson, D.L.; Yang, F.; Eisen, J.A.; Paulsen, I.T.; White, W. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **2001**, *29*, 41-43.
- [40] Wu, C.H.; Huang, H.; Yeh, L.L.; Barker, W.C. Protein family classification and functional annotation. *Comp. Biol. Chem.*, **2003**, *27*, 37-47.
- [41] Dickerson, R.E.; Geis, I. The structure and action of proteins. Benjamin/Cummings, Menlo Park, California, **1969**.
- [42] Petsko, G.A. From sequence to consequence. *Genome Biol.*, **2000**, *1*, 406.
- [43] Baker, D. Proteins by design. Protein structure prediction: when is it useful? *The Scientist*, **2006**, 26-32.
- [44] Zhang, Y. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.*, **2009**, *19*, 145-155.

- [45] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.*, **2000**, *28*, 235-242.
- [46] Dill, K.A.; Ozkan, S.B.; Weikl, T.R.; Chodera, J.D.; Voelz, V.A. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.*, **2007**, *17*, 342-346.
- [47] Moulton, J. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **2006**, *361*, 453-458.
- [48] Murzin, A.G. Protein structure watch: Making "predictions". **2004**. <http://www.forcas.org>.
- [49] Floudas, C.A. Computational methods in protein structure prediction. *Biotechnol. Bioeng.*, **2007**, *97*, 207-213.
- [50] Kopp, J.; Schwede, T. Automated protein structure homology modeling: a progress report. *Pharmacogenomics J.*, **2004**, *5*(4), 405-416.
- [51] Vitkup, D.; Melomud, E.; Moulton, J.; Sander, C. Completeness in structural genomics. *Nat. Struct. Biol.*, **2001**, *8*(6), 559-566.
- [52] Pandit, S.B.; Zhang, Y.; Skolnick, J. TASSER-Lite: an automated tool for protein comparative modeling. *Biophys. J.*, **2006**, *91*(11), 4180-4190.
- [53] Skolnick, J.; Zhang, Y.; Arakaki, A.K.; Kolinski, A.; Boniecki, M.; Szilagyi, A.; Kihara, D. TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, **2003**, *53*, 469-479.
- [54] Zhang, Y.; Arakaki, A.K.; Skolnick, J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*, **2005**, *61*(S7), 91-98.
- [55] Lund, O.; Nielsen, M.; Lundegard, C.; Worning, P. CPHmodels 2.0: X3M a computer program to extract 3dmodels. *Abstract at the CASP5 conference*, **2002**, A102.
- [56] Guex, N.; Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **1997**, *18*, 2714-2723.
- [57] Lambert, C.; Leonard, N.; De Bolle, X.; Depiereux, E. ESyPred3D: prediction of proteins 3D structures. *Bioinformatics*, **2002**, *18*, 1250-1256.
- [58] (a) Sali, A.; Blundell, T. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.*, **1993**, *234*, 779-815. (b) Pokarowski, P.; Kozlowski, A.; Jernigan, R.L.; Kothari, N.S.; Pokarowska, M.; Kolinski, A. Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins*, **2005**, *59*, 49-57. (c) Jha, A.N.; Vishveshwara, S.; Banavar, J.R. Amino acid interaction preferences in proteins. *Protein Sci.*, **2010**, *19*, 603-616.
- [59] Combat, C.; Jambon, M.; Deleage, G.; Geourjon, C. Geno3D: automatic comparative molecular modelling of protein. *Bioinformatics*, **2002**, *18*, 213-214.
- [60] Bates, P.A.; Kelley, L.A.; MacCallum, R.M.; Sternberg, M. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **2001**, *45*, 39-46.
- [61] Simons, K.T.; Kooperberg, C.; Huang, C.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **1997**, *268*, 209-225.
- [62] Skolnick, J.; Kolinski, A.; Kihara, D.; Betancourt, M.; Rotkiewicz, P.; Boniecki, M. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins*, **2001**, *55*, 149-156.
- [63] Zhang, Y.; Skolnick, J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys. J.*, **2004**, *87*, 2647-2655.
- [64] Zhang, Y.; Skolnick, J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA*, **2004**, *102*, 1029-1034.
- [65] Zhang, Y.; Kolinski, A.; Skolnick, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.*, **2003**, *85*, 1145-1164.
- [66] Zhang, Y.; Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.*, **2004**, *25*, 865-871.
- [67] Jones, D.T.; Taylor, W.R.; Thornton, J.M. A new approach to protein fold recognition enriching the sequence substitution. *Nature*, **1992**, *358*, 86-89.
- [68] Teodorescu, O.; Galor, T.; Pillardy, J.; Elber, R. Enriching the sequence substitution. *Proteins*, **2004**, *54*, 41-48.
- [69] Jones, D.T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **1999**, *287*, 797-815.
- [70] Ota, M.; Mizunuma, T.; Nishikawa, K. Introduction to LIBRA. *DDBJ off-line news*, **1997**, no. 8.
- [71] Oldziej, S.; Czaplowski, C.; Liwo, A.; Chinchio, M.; Nanas, M.; Vila, J.A.; Khalili, M.; Arnautova, Y.A.; Jagielska, A.; Makowski, M.; Schafroth, H.D.; Kazmierkiewicz, R.; Ripoll, D.R.; Pillardy, J.; Saunders, J.A.; Kang, Y.K.; Gibson, K.D.; Scheraga, H.A. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 7547-7552.
- [72] McCammon, J.A.; Gelin, B.R.; Karplus, M. Dynamics of folded proteins. *Nature*, **1977**, *267*, 585-590.
- [73] Karplus, M.; McCammon, J.A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, **2002**, *9*, 646-652.
- [74] Kolinski, A.; Skolnick, J.; Yaris, R. Order-disorder transitions in tetrahedral lattice. *Proc. Natl. Acad. Sci. USA*, **1986**, *83*, 7267-7271.
- [75] Ortiz, A.R.; Kolinski, A.; Rotkiewicz, P.; Ilkowsk, B.; Skolnick, J. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, **1999**, *37*, 77-185.
- [76] Shea, J.E.; Brooks, C.L. From folding theories to folding proteins: folding and unfolding. *Annu. Rev. Phys. Chem.*, **2001**, *52*, 499-535.
- [77] Levitt, M.; Warshell, A. Computer simulations of protein folding. *Nature*, **1975**, *253*, 694-698.
- [78] Schaefer, M.; Karplus, M. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.*, **1996**, *100*, 1578-1599.
- [79] Cramer, C.J.; Truhlar, D.G. Implicit solvation models: equilibria, structure, spectra, and dynamics. *Chem. Rev.*, **1999**, *99*, 2161.
- [80] Simonson, T. Refinement against fiber diffraction data. *Curr. Opin. Struct. Biol.*, **2001**, *11*, 243.
- [81] Duan, Y.; Kollman, P.A. Pathways ... in a 1-ms simulation in aqueous solution. *Science*, **1998**, *282*, 740-744.
- [82] Neidigh, J.W.; Fesinmeyer, R.M.; Anderson, N.H. Designing a 20-residue protein. *Nat. Struct. Biol.*, **2002**, *9*, 425-430.
- [83] a) Simmerling, C.; Strockbine, B.; Roitberg, A.E. All-Atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, **2002**, *124*, 11258; (b) Shaw, D.E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Eastwood, M.P.; Bank, J.A.; Jumper, J.M.; Salmon, J.K.; Shan, Y.; Wriggers, W. Atomic Level Characterization of the Structural Dynamics of Proteins. *Science*, **2010**, *330*, 341-346.
- [84] Qiu, L.; Pabit, S.A.; Roitberg, A.E.; Hagan, S.J. Clients in mixed electrolytes. *J. Am. Chem. Soc.*, **2002**, *124*, 12952.
- [85] Lipton, M.; Still, W.C. The multiple minimum problem in molecular modeling. *J. Comp. Chem.*, **1998**, *9*, 343-355.
- [86] Saunders, M. Stochastic exploration of molecular mechanics energy surfaces. Hunting for the global minimum. *J. Am. Chem. Soc.*, **1987**, *109*, 3150-3152.
- [87] Ferguson, D.M.; Raber, D.J. A comparison of methods for conformational searching. *J. Am. Chem. Soc.*, **1989**, *111*, 4371-4378.
- [88] Li, Z. Q.; Scheraga, H.A. Monte-Carlo-minimization approach to the multiple. *Proc. Natl. Acad. Sci. USA*, **1987**, *84*, 6611-6615.
- [89] Chang, G.; Guida, W.C.; Still, W.C. An internal coordinate monte carlo method for searching conformational space. *J. Am. Chem. Soc.*, **1989**, *111*, 4379-4386.
- [90] Goldberg, D.E. Genetic algorithms in search, optimization and machine learning, Kluwer Academic Publishers, Boston, MA, **1989**.
- [91] Gautham, N.; Rafi, Z.A. Global search for optimal biomolecular structures using mutually orthogonal Latin squares. *Curr. Sci.*, **1992**, *63*, 560-564.
- [92] Vengadesan, K.; Gautham, N. Enhanced sampling of the molecular ... latin squares: application to peptide structures. *Biophys. J.*, **2003**, *84*, 2897-2906.
- [93] Hanggi, G.; Braun, W. Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin. *FEBS Lett.*, **1994**, *344*, 147-153.
- [94] Huang, E.S.; Samudrala, R.; Ponder, J.W. Ab initio fold prediction of small helical proteins. *Protein Sci.*, **1998**, *7*, 1998-2003.
- [95] Shin, J.K.; John, M.S. High directional monte carlo procedure. *Biopolymers*, **1991**, *31*, 177-185.
- [96] Lee, J.; Scheraga, H.A.; Rackovsky, S. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J. Comp. Chem.*, **1997**, *18*, 1222-1232.

- [97] Klepeis, J.L.; Pieja, M.J.; Floudas, C.A. Hybrid global optimization ... prediction: alternating hybrids. *Biophys. J.*, **2003**, *84*, 869-882.
- [98] Xia, Y.; Huang, E.S.; Levitt, M.; Samudrala, R. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.*, **2000**, *300*, 171-185.
- [99] Lee, J.; Liwo, A.; Ripoli, D.R.; Pillardy, J.; Scheraga, H.A. Calculation of protein conformation by global optimization of a potential energy function. *Proteins*, **1999**, *37*, 204-208.
- [100] Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science*, **1983**, *220*, 671-680.
- [101] Wilson, S.R.; Cui, W.; Moskowitz, J.W.; Schmidt, K.E. Conformational analysis ... by the simulated annealing method. *Tet. Lett.*, **1998**, *29*, 4373-4376.
- [102] Morales, L.B.; Garduno-Juarez, R.; Romero, D. Applications of simulated annealing to the multiple-minima. *J. Biomol. Struct. Dyn.*, **1991**, *8*, 721-735.
- [103] Okamoto, Y.; Kikuchi, T.; Kawai, H. Prediction of ... by Monte Carlo simulated annealing. *Chem. Lett.*, **1992**, *1*, 1275-1278.
- [104] Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **1999**, *314*, 141-151.
- [105] Hansmann, U.H.E. Stochastic dynamics simulations in a new generalized ensemble. *Chem. Phys. Lett.*, **1997**, *281*, 140-150.
- [106] Kolinski, A.; Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins*, **1994**, *18*, 353-356.
- [107] Dinner, A.R.; Sali, A.; Karplus, M. The folding mechanism of larger model proteins: role of native structure. *Proc. Natl. Acad. Sci. USA*, **1996**, *93*, 8365-8361.
- [108] Vasquez, M.; Scheraga, H.A. Use of buildup and energy-minimization structures of the backbone of enkephalin. *Biopolymers*, **1985**, *24*, 1437-1447.
- [109] Gibson, K.D.; Scheraga, H.A. Revised algorithm for the build-up procedure for predicting protein conformation by energy minimization. *J. Comp. Chem.*, **1985**, *8*, 826-834.
- [110] Vajda, S.; DeLisi, C. Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers*, **1990**, *29*, 1755-1772.
- [111] Huang, E.S.; Samudrala, R.; Ponder, J.W. Ab initio fold prediction of small helical proteins. *J. Mol. Biol.*, **1996**, *290*, 267-281.
- [112] Kim, D.E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the rosetta server. *Nucleic Acids Res.*, **2004**, *32*, W526-W531.
- [113] Bradley, P.; Misura, K.M.S.; Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science*, **2005**, *309*, 1868-1871.
- [114] Hung, L.-H.; Ngan, S.-C.; Liu, T.; Samudrala, R. PROTINFO: new algorithms for enhanced protein structure prediction. *Nucleic Acid Res.*, **2005**, *33*, W77-W80.
- [115] Cheng, J.; Randell, A.Z.; Sweredoski, M.J.; Baldi, P. SCRATCH: a protein structure and structural feature. *Nucleic Acids Res.*, **2005**, *33*, W72-W76.
- [116] Klepeis, J.L.; Floudas, C.A. Comparative study of global minimum energy .. proteins from the amino acid sequence. *Biophys. J.*, **2003**, *85*, 2119-2146.
- [117] Fujitsuka, Y.; Chikenji, G.; Takada, S. SimFold energy function for de novo protein structure prediction: consensus with rosetta. *Proteins*, **2005**, *62*, 381-398.
- [118] (a) Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. A computational pathway for bracketing native-like structures for small alpha helical globular proteins. *Phys. Chem. Chem. Phys.*, **2005**, *7*, 2364-2375. (b) Thukral, L.; Shenoy, S.R.; Bhushan, K.; Jayaram, B. ProRegIn: a regularity index for the selection of native-like tertiary structures of proteins. *J. Biosci.*, **2007**, *32*, 71-81. (c) Mittal, A.; Jayaram, B.; Shenoy, S.R.; Bawa, T.S. A Stoichiometry driven universal spatial organization of backbones of folded proteins: are there chargeaff's rules for protein folding? *J. Biomol. Struct. Dyn.*, **2010**, *28*, 133-142.
- [119] (a) Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. Protein structure evaluation using an all-atom energy based empirical scoring function. *J. Biomol. Struct. Dyn.*, **2006**, *23*, 385-406. (b) Arora, N.; Jayaram, B. Strength of hydrogen bonds in alpha helices. *J. Comput. Chem.* **1997**, *18*, 1245-1252. (c) Dixit, S.B.; Bhasin, R.; Rajasekaran, E.; Jayaram, B. Solvation thermodynamics of amino acids : assessment of the electrostatic contribution and force-field dependence. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 1105-1113.
- [120] Jayaram, B.; Bhushan, K.; Shenoy, S.R.; Narang, P.; Bose, S.; Agrawal, P.; Sahu, D.; Pandey, V. Bhageerath : an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. *Nucleic Acids Res.*, **2006**, *34*, 6195-6204.
- [121] Joshi, R.R.; Jyothi, S. Ab-initio structure of Human Seminal Plasma Protiens. *Comput. Bio. Chem.*, **2002**, *27*, 241-252.
- [122] Bryson, J.W.; Betz, S.F.; Lu, H.S.; Suich, D.J.; Zhou, H.X.; O'Neil, K.T.; DeGrado, W.F. Protein design: a hierarchic approach. *Science*, **1995**, *270*, 935-941.
- [123] Dahiyat, B.I.; Mayo, S.L. Fully automated sequence selection. *Science*, **1997**, *278*, 82-87.
- [124] Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D. Prediction of membrane protein structures with protein fold with atomic-level accuracy. *Science*, **2003**, *302*, 1364-1368.
- [125] Bryngelson, J.D.; Onuchic, J.N.; Succi, N.D.; Wolynes, P.G. Funnel, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, **1995**, *21*, 167-195.
- [126] Dill, K.A.; Chan, H.S. From Levinthal to pathways to funnels. *Nature Struct. Biol.*, **1997**, *4*, 10-19.
- [127] Sinha, K.K.; Udgaonkar, J.B. Early events in protein folding. *Curr. Sci.*, **2009**, *96*, 1053-1070.
- [128] Kim, P.S.; Baldwin, R.L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.*, **1982**, *51*, 459-489.
- [129] De Prat Gay, G.; Ruiz-Sanz, J.; Neira, J.L.; Itzhaki, L.S.; Fersht, A.R. Folding of a nascent polypeptide chain *in vitro*: cooperative formation of structure in a protein module. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 3683-3686.
- [130] Ferguson, N.; Fersht, A.R. Early events in protein folding. *Curr. Opin. Struct. Biol.*, **2003**, *13*, 75-81.
- [131] Sadqi, M.; Lapidus, L.J.; Munoz, V. How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. USA*, **2003**, *100*, 12117-12122.
- [132] Roder, H.; Maki, K.; Cheng, H. Early events in protein folding explored by rapid mixing methods. *Chem. Rev.*, **2006**, *106*, 1836-1861.
- [133] Jackson, S.E. How do small single-domain proteins fold? *Fold Des.*, **1998**, *3*, R81-R91.
- [134] Plaxco, K.W.; Millett, I.S.; Segel, D.J.; Doniach, S.; Baker, D. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.*, **1999**, *6*, 554-556.
- [135] Fersht, A.R. Optimization of rates of protein folding: the nucleation-condensation mechanism. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 10869-10873.
- [136] Dagget, V.; Fersht, A.R. Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.*, **2003**, *28*, 18-25.
- [137] Elofsson, A.; von Heijne, G. Membrane protein structure: prediction versus reality. *Annu. Rev. Biochem.*, **2007**, *76*, 125-140.
- [138] Klabunde, T.; Hessler, G. Drug design strategies for targeting G protein-coupled receptors. *Chem. Bio. Chem.*, **2002**, *3*, 928-944.
- [139] Chen, C.-C.; Chen, C.-M. A dual-scale approach toward structure prediction of retinal proteins. *J. Struct. Bio.*, **2009**, *165*, 37-46.
- [140] Zhang, Y.; De Vries, M.E.; Skolnick, J. Structure Modeling of All Identified G. *PLoS Comput. Biol.*, **2006**, *2*, 2(e13).
- [141] White, S. H. The progress of membrane protein structure determination. *Protein Sci.*, **2004**, *13*, 1948-1949.
- [142] Oberai A.; Ihm, Y.; Kim S.; Bowie, J.U. A limited universe of membrane protein families and folds. *Protein Sci.*, **2006**, *15*, 1723-1734.
- [143] White, S.H. Biophysical dissection of membrane proteins. *Nature*, **2009**, *459*, 344-346.
- [144] Raman, P.; Cherezov, V.; Caffrey, M. The membrane protein data bank. *Cell Mol. Life Sci.*, **2006**, *63*, 36-51.
- [145] Sansom, M.S.P.; Scott, K.A.; Bond, P.J. Coarse grained simulation: a high throughput computational approach to membran protein. *Biochem. Soc. Trans.* **2008**, *36*(1), 27-32.
- [146] Tusnady, G.E.; Dosztanyi, Z.; Simon, I. Transmembrane proteins in protein data bank: identification and classification. *Bioinformatics*, **2004**, *20*, 2964-2972.
- [147] Tusnady, G.E.; Dosztanyi, Z.; Simon, I. Transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **2005**, *33*, D275-278.
- [148] Gromiha, M.M.; Yabuki, Y.; Suresh, M.X.; Thangakani, A.M.; Suwa, M.; Fukui, K.K. TMFunction: database for functional residues in membrane proteins. *Nucleic Acids Res.*, **2009**, *37*, D201-204.

- [149] Henderson, R.; Unwin, P.N.T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, **1975**, *257*, 28-32.
- [150] Tomita, M.; Marchesi, V.T. Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycoporphin. *Proc. Natl. Acad. Sci. USA*, **1975**, *72*, 2964-2968.
- [151] Ovchinnikov, Y.A.; Abdulaev, N.G.; Feigina, M.Y.; Kiselev, A.V.; Lobanov, N.A.A. Recent findings in the structure-functional characteristics of bacteriorhodopsin. *FEBS Lett.*, **1977**, *84*, 1-4.
- [152] Khorana, H.G.; Gerber, G.E.; Herlihy, W.C.; Gray, C.P.; Anderegg, R.J.; Nihel, K.; Biemann, K. Amino acid sequence of rhodopsin. *Proc. Natl. Acad. Sci. USA*, **1979**, *76*, 5046-5050.
- [153] Ovchinnikov, Y.A.; Abdulaev, N.G.; Feigina, M.Y.; Kiselev, A.V.; Lobanov, N. The structural basis of the functioning of bacteriorhodopsin. *FEBS Lett.*, **1979**, *100*, 219-224.
- [154] von Heijne, G.; Bloemberg, C. Transmembrane translocation of proteins-the direct transfer model. *Eur. J. Biochem.*, **1979**, *97*, 175-181.
- [155] Engelman, D.M.; Steitz, T.A. The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. *Cell*, **1981**, *23*, 411-422.
- [156] Kyte, J.; Doolittle, R.F. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, **1982**, *157*, 105-132.
- [157] Weiss, M. S.; Kreuzsch, A.; Schiltz, E.; Nestel, U.; Welte, W.; Weckesser, J.; Schulz, G. E. The structure of porin from *Rhodobacter*. *FEBS Lett.*, **1991**, *280*, 379-382.
- [158] Wallin, E.; Tsukihara, T.; Yoshikawa, S.; von Heijne, G.; Elofsson, A. Architecture of helix bundle. *Protein Sci.*, **1997**, *6*, 808-815.
- [159] von Heijne, G. Towards a comparative anatomy of N-terminal topogenic protein sequences. *J. Mol. Biol.*, **1986**, *189*, 239-242.
- [160] von Heijne, G. The distribution of positively charged residues in topology. *EMBO J.*, **1986**, *5*, 3021-3027.
- [161] von Heijne, G. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, **1989**, *341*, 456-458.
- [162] Senes, A.; Gerstein, M.; Engelman, D.M. Statistical analysis of amino acids. *J. Mol. Biol.*, **2000**, *296*, 921-936.
- [163] Kim, S.; Jeon, T.J.; Oberai, A.; Yang, D.; Schmidt, J.J.; Bowie, J.U. Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 14278-14283.
- [164] Samatey, F.A.; Xu, C.; Popot J-L. On the distribution of amino acid alpha-helix bundles. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 4577-4581.
- [165] Ott, C.M.; Lingappa, V.R. Integral membrane protein biosynthesis: why topology is hard to predict. *J. Cell Sci.*, **2002**, *115*, 2003-2009.
- [166] Moller, S.; Croning, M.; Apweiler, R. Apweiler Conflict-resolution for the automated annotation *Bioinformatics*, **2001**, *17*, 646-653.
- [167] Ikeda, M.; Arai, M.; Lao, D.; Shimizu, T. Transmembrane topology prediction methods: A re-assessment and improvement by a consensus method using a dataset of experimentally characterized transmembrane topology. *In Silico Biol.*, **2001**, *2*, 1-15.
- [168] Melen, K.; Krogh, A.; von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **2003**, *327*, 735-744.
- [169] Cuthbertson, J.M.; Doyle, D.A.; Sansom, M.S. Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.*, **2005**, *18*, 295-308.
- [170] Melen, K.; Krogh, A.; von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **2003**, *327*, 735-744.
- [171] Yarov-Yaravov, V.; Baker, D.; Catterall, W. A. Open and closed states in ROSETTA structural models of K⁺ channels. *Proc. Natl. Acad. Sci. USA*, **2006**, *103*, 7292-7297.
- [172] Forrest, L.R.; Tang, C.L.; Honig, B. On the accuracy of homology modeling and sequence. *Biophys. J.*, **2006**, *91*, 508-517.
- [173] Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; LeHoczy, J.; Levine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J.P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J.C.; Mungall, A.; Plumb, R.; Ross, M.; Showkneen, R.; Sims, S.; Waterston, R.H.; Wilson, R.K.; Hillier, L.W.; McPherson, J.D.; Marra, M.A.; Mardis, E.R.; Fulton, L.A.; Chinwalla, A.T.; Pepin, K.H.; Gish, W.R.; Chissoe, S.L.; Wendl, M.C.; Delehaunty, K.D.; Miner, T.L.; Delehaunty, A.; Kramer, J.B.; Cook, L.L.; Fulton, R.S.; Johnson, D.L.; Minx, P.J.; Clifton, S.W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J.F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R.A.; Muzny, D.M.; Scherer, S.E.; Bouck, J.B.; Sodergren, J.K.; Worley, K.F.; Rives, C.M.; Gorrell, J.H.; Metzker, M.L.; Naylor, S.L.; Kucherlapati, R.S.; Nelson, D.L.; Weinstock, G.M.; Sakaki, Y.; Fujiiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Winkler, P.; Smith, D.R.; Doucette-Stamm, L.; Rubinfeld, M.; Weinstock, K.; Lee, H.M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R.W.; Federspiel, N.A.; Abola, A.P.; Proctor, M.J.; Myers, R.M.; Schmutz, J.; Dickinson, M.; Grimwood, J.; Cox, D.R.; Olson, M.V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G.A.; Athanasiou, M.; Schultz, R.; Roe, B.A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W.R.; de la Bastide, M.; Dedhia, N.; Böcker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J.A.; Bateman, A.; Batzoglu, S.; Birney, E.; Bork, P.; Brown, D.G.; Burge, C.B.; Cerutti, L.; Chen, H.C.; Church, D.; Clamp, M.; Copley, R.R.; Doerks, T.; Eddy, S.R.; Eichler, E.E.; Furey, T.S.; Galagan, J.; Gilbert, J.G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L.S.; Jones, T.A.; Kasif, S.; Kasprzyk, A.; Kennedy, S.; Kent, W.J.; Kitts, P.; Koonin, E.V.; Korfi, I.; Kulp, D.; Lancet, D.; Lowe, T.M.; McLysaght, A.; Mikkelsen, T.; Moran, J.V.; Mulder, N.; Pollara, V.J.; Ponting, C.P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A.F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y.I.; Wolfe, K.H.; Yang, S.P.; Yeh, R.F.; Collins, F.; Guyer, M.S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K.A.; Patrinos, A.; Morgan, M.J.; de Jong, P.; Catanese, J.J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y.J. International human genome sequencing consortium. *Nature*, **2001**, *409*, 860-921.
- [174] Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; Gocayne, J.D.; Amanatides, P.; Ballew, R.M.; Huson, D.H.; Wortman, J.R.; Zhang, Q.; Kodira, C.D.; Zheng, X.H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P.D.; Zhang, J.; Gabor Miklos, G.L.; Nelson, C.; Broder, S.; Clark, A.G.; Nadeau, J.; McKusick, V.A.; Zinder, N.; Levine, A.J.; Roberts, R.J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A.E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T.J.; Higgins, M.E.; Ji, R.R.; Ke, Z.; Ketchum, K.A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G.V.; Milshina, N.; Moore, H.M.; Naik, A.K.; Narayan, V.A.; Neelam, B.; Nusskern, D.; Rusch, D.B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M.L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferrieres, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.;

- McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y.H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N.N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J.F.; Guigó, R.; Campbell, M.J.; Sjolander, K.V.; Karlak, B.; Kejarawal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y.H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. The sequence of human genome. *Science*, **2001**, *291*, 1304-1351.
- [175] Peri, S.; Navarro, J.D.; Kristiansen, T.Z.; Amanchy, R.; Surendranath, V.; Muthusamy, B.; Gandhi, T.K.; Chandrika, K.N.; Deshpande, N.; Suresh, S.; Rashmi, B.P.; Shanker, K.; Padma, N.; Niranjana, V.; Harsha, H.C.; Talreja, N.; Vrushabendra, B.M.; Ramya, M.A.; Yatish, A.J.; Joy, M.; Shivashankar, H.N.; Kavitha, M.P.; Menezes, M.; Choudhury, D.R.; Ghosh, N.; Saravana, R.; Chandran, S.; Mohan, S.; Jonnalagadda, C.K.; Prasad, C.K.; Kumar-Sinha, C.; Deshpande, K.S.; Pandey, A. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, **2004**, *32*, D497-D501.
- [176] Keshava Prasad, T.S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; Balakrishnan, L.; Marimuthu, A.; Banerjee, S.; Somanathan, D.S.; Sebastian, A.; Rani, S.; Ray, S.; Harrys Kishore, C.J.; Kanth, S.; Ahmed, M.; Kashyap, M.K.; Mohmood, R.; Ramachandra, Y.L.; Krishna, V.; Rahiman, B.A.; Mohan, S.; Ranganathan, P.; Ramabadran, S.; Chaerkady, R.; Pandey, A. Human protein reference database-2009 update. *Nucleic Acids Res.*, **2008**, *37*, D767-D772.
- [177] Mathivanan, S.; Periaswamy, B.; Gandhi, T.K.B.; Kandasamy, K.; Suresh, S.; Mohmood, R.; Ramachandra, Y.L.; Pandey, A. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, **2006**, *7*(5), S19.
- [178] Basu, M.K.; Poliakov, E.; Rogozin, I.B. Domain mobility in proteins: functional and evolutionary implications. *Brief. Bioinform.*, **2009**, *10*, 205-216.
- [179] Gavin, A.C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L.J.; Bastuck, S.; Dümpelfeld, B.; Edelmann, A.; Heurtier, M.A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A.M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J.M.; Kuster, B.; Bork, P.; Russell, R.B. Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **2006**, *440*, 631-636.
- [180] Arifuzzaman, M.; Maeda, M.; Itoh, A.; Nishikata, K.; Takita, C.; Saito, R.; Ara, T.; Nakahigashi, K.; Huang, H.; Hirai, A.; Tsuzuki, K.; Nakamura, S.; Altaf-Ul-Amin, M.; Oshima, T.; Baba, T.; Yamamoto, N.; Kawamura, T.; Ioka-Nakamichi, T.; Kitagawa, M.; Tomita, M.; Kanaya, S.; Wada, C.; Mori, H. Identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Res.*, **2006**, *16*, 686-691.
- [181] Han, J.D.; Bertin, N.; Hao, T.; Goldberg, D.S.; Berriz, G.F.; Zhang, L.V.; Dupuy, D.; Walhout, A.J.; Cusick, M.E.; Roth, F.P.; Vidal, M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **2004**, *430*, 88-93.
- [182] Fuller, J.C.; Burgoyne, N.J.; Jackson, R.M. Predicting druggable binding sites at the protein-protein interface. *Drug Discov. Today*, **2009**, *14*, 155-161.
- [183] Tuncbag, N.; Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. A survey of available tools and web servers. *Brief Bioinformatics*, **2009**, *10*(3), 217-232.
- [184] Barabasi, A.L.; Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **2004**, *5*, 101-113.
- [185] Teilum, K.; Olsen, J.G.; Kragelund, B.B. Functional aspects of protein flexibility. *Cell. Mol. Life Sci.*, **2009**, *66*, 2231-2247.
- [186] Jomain, J.B.; Tallet, E.; Broutin, I.; Hoos, S.; van Agthoven, J.; Ducruix, A.; Kelly, P.A.; Kragelund, B.B.; England, P.; Goffin, V. Structural and thermodynamic bases for the design of pure prolactin receptor antagonists: X-ray structure of Del1-9-G129R-hPRL. *J. Biol. Chem.*, **2007**, *282*, 33118-33131.
- [187] Schoichet, B.K.; Baase, W.A.; Kuroki, R.; Mathews, B.W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 452-456.
- [188] Smith, C.G.; Vane, G.R. The discovery of captopril. *FASEB J.*, **2003**, *17*, 788-789.
- [189] Caetano-Anolles, G.; Kim, H.S.; Mitternath, J.E. From phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci. USA*, **2007**, *104*, 9358-9363.
- [190] Zheng, C.J.; Han, L.Y.; Yap, C.W.; Ji, Z.L.; Cao, Z.W.; Chen, Y.Z. Progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.*, **2006**, *58*, 259-279.
- [191] Overington, J.P.; Al-Lazikani, B.; Hopkins, A.L. How many drug targets are there? *Nat. Rev. Drug Discov.*, **2006**, *5*(12), 993-996.
- [192] Zambrowicz, B.P.; Sands, A.T. Knockouts model the 100 best-selling drugs--will they model the next 100? *Nat. Rev. Drug Discov.*, **2003**, *2*(1), 38-51.
- [193] Butcher, S.P. Target discovery and validation in the post-genomic era. *Neurochem. Res.*, **2003**, *28*(2), 367-371.
- [194] Li, Q.; Lai, L. Prediction of potential drug targets based on simple sequence. *BMC Bioinformatics*, **2007**, *8*, 353.
- [195] Bakheet, T.M.; Doig, A.J. Properties and identification of human protein drug targets. *Bioinformatics*, **2009**, *25*, 1-8.
- [196] Wesley, C.V.V.; Hol, W.G.J.; Myler, P.J.; Stewart, L.J. The role of medical structural genomics in discovering new drugs for infectious disease. *PLoS Comput. Biol.*, **2009**, *5*, 1-6.
- [197] Li, A.P. Accurate prediction of human drug toxicity: a major challenge in drug development. *Chem. Bio. Interact.*, **2004**, *150*, 3-7.
- [198] Shaikh, S.A.; Jain, T.; Sandhu, G.; Latha, N.; Jayaram, B. From drug target to leads- sketching, a physicochemical pathway for lead molecule design *in silico*. *Curr. Pharm. Des.*, **2007**, *13*, 3454-3470.
- [199] Dearden, J.C. *In silico* prediction of drug toxicity. *J. Comp. Mol. Des.*, **2003**, *17*, 119-127.
- [200] Schneider, G.; Fechner, U. Computer-based de novo design of druglike molecules. *Nat. Rev. Drug Discov.*, **2005**, *4*, 649-663.
- [201] Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **2004**, *337*, 635-645.
- [202] Oldfield, C.J.; Cheng, Y.; Cortese, M.S.; Brown, C.J.; Uversky, V.N.; Dunker, A.K. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **2005**, *44*, 1989-2000.
- [203] Tompa, P.; Dosztányi, Z.; Simon, I. Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J. Proteome Res.*, **2006**, *5*, 1996-2000.
- [204] Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **2005**, *6*, 197-208.
- [205] Fink, A.L. Natively unfolded proteins. *Curr. Opin. Struct. Biol.*, **2005**, *15*, 35-41.
- [206] Dunker, A.K.; Silman, I.; Uversky, V.N.; Sussman, J.L. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **2008**, *18*, 1-9.
- [207] Iakoucheva, L.M.; Brown, C.J.; Lawson, J.D.; Obradovic, Z.; Dunker, K. Intrinsic disorder in cell-signalling and cancer-associated proteins. *J. Mol. Biol.*, **2002**, *323*, 573-584.
- [208] Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K.; Sequence complexity of disordered protein. *Proteins*, **2001**, *42*, 38-48.
- [209] Vucetic, S.; Brown, C.J.; Dunker, A.K.; Obradovic, Z. Flavours of protein disorder. *Proteins*, **2003**, *52*, 573-584.
- [210] Mitchell, P.J.; Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **1989**, *245*, 371-378.
- [211] Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev.*, **2005**, *6*, 197-208.
- [212] Linding, R.; Jensen, L.J.; Diella, F.; Bork, P.; Gibson, T.J.; Russell, R. B. Protein disorder prediction: implications for structural proteomics. *Structure*, **2003**, *11*, 1453-1459.

- [213] Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **2004**, *337*, 635-645.
- [214] Garbuzynskiy, S.O.; Lobanov, M.Y.; Galzitskaya, O.V. To be folded or to be unfolded? *Protein Sci.*, **2004**, *13*, 2871-2877.
- [215] Prilusky, J.; Felder, C.E.; Zeev-Ben-Mordehai, T.; Rydberg, E.H.; Man, O.; Beckmann, J.S.; Silman, I.; Sussman, J.L. FoldIndex[®]: a simple tool predicts whether a given protein is intrinsically disordered. *Bioinformatics*, **2005**, *21*, 3435-3438.
- [216] Linding, R.; Russell, R.B.; Neduva, V.; Gibson, T.J. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **2003**, *31*, 3701-3708.
- [217] Dosztanyi, Z.; Csizmek, V.; Tompa, P.; Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **2005**, *21*, 3433-3434.
- [218] Liu, J.; Rost, B. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.*, **2003**, *31*, 3833-3835.
- [219] Li, X.; Romero, P.; Rani, M.; Dunker, A.K.; Obradovic, Z. Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform. Ser. Workshop Genome Inform.*, **1999**, *10*, 30-40.
- [220] Coeysaux, K.; Poupon, A. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, **2005**, *21*, 1891-1900.
- [221] Schlessinger, A.; Punta, M.; Rost, B. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **2007**, *23*, 2376-2384.
- [222] Dyson, H.J.; Wright, P.E. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, **2002**, *12*, 54-60.
- [223] Spolar, R.S.; Record, M.T. Coupling of local folding to site-specific binding of proteins to DNA. *Science*, **1994**, *263*, 777-784.
- [224] Patel, L.; Abate, C.; Curran, T. DNA binding by Fos and Jun results in altered protein conformation. *Nature*, **1990**, *347*, 572-574.
- [225] Laity, J.H.; Dyson, H.J.; Wright, P.E. DNA-induced α -helix capping in conserved linker sequences is a determinant of binding affinity in Cys₂-His₂ zinc fingers. *J. Mol. Biol.*, **2000**, *295*, 719-727.
- [226] Uversky, V.N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **2002**, *11*, 739-756.
- [227] DiNitto, J.P.; Huber, P.W. Mutual induced fit binding of *Xenopus* ribosomal protein L5 to 5S rRNA. *J. Mol. Biol.*, **2003**, *330*, 979-992.
- [228] Kim, S.H.; Shin, D.H.; Choi, I.G.; Gahmen, U.S.; Chen, S.; Kim, R. Structure-based functional inference in structural genomics. *J. Struct. Funct. Genomics*, **2003**, *4*, 129-135.
- [229] Petry, D.; Honig, B. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr. Opin. Struct. Biol.*, **2009**, *19*, 363-368.
- [230] Sadowski, M.; Jones, D.T. The sequence-structure relationship and protein function prediction. *Curr. Opin. Struct. Biol.*, **2009**, *19*, 357-362.
- [231] Bork, P.; Sander, C.; Valencia, A. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.*, **1993**, *2*, 31-40.
- [232] Russell, R.B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.*, **1998**, *279*, 121-1227.
- [233] Copley, R.R.; Russell, R.B.; Ponting, C.P. Sialidase like Asp-Boxes: Sequence similar structures within different protein folds. *Protein Sci.*, **2001**, *10*, 285-292.
- [234] Ausiello, G.; Peluso, D.; Via, A.; Helmer-Citterich, M. Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinformatics*, **2007**, *8*(1), S24.
- [235] 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.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **2000**, *25*, 25-29.
- [236] Reeves, G.A.; Eilbeck, K.; Magrane, M.; O'Donovan, C.; Montecchi-Palazzi, L.; Harris, M.A.; Orchard, S.; Jimenez, R.C.; Plic, A.; Hubbard, T.J.P.; Hermjakob, H.; Thornton, J.M. The Protein Feature Ontology: A Tool for the Unification of Protein Feature Annotations. *Bioinformatics*, **2008**, *24*, 2767-2772.
- [237] Jenkinson, A.M.; Albrecht, M.; Birney, E.; Blankenburg, H.; Down, T.; Finn, R.D.; Hermjakob, H.; Hubbard, T.J.; Jimenez, R.C.; Jones, P.; Kähäri, A.; Kulesha, E.; Macías, J.R.; Reeves, G.A.; Plic, A. Integrating biological data - the Distributed Annotation System. *BMC Bioinformatics*, **2008**, *9*(8), S3.
- [238] Reeves, G.A.; Talavera, D.; Thornton, J. M. Genome and proteome annotation: organization, interpretation and integration. *Roc. Interface*, **2009**, *6*, 129-147.
- [239] Jayaram, B. Decoding the design principles of amino acids and the chemical logic of protein sequences. *Nature Precedings*, **2008**, <http://hdl.handle.net/10101/npre.2008.2135.1>
- [240] Damborsky, J.; Brezovsky, J. Computational tools for designing and engineering biocatalysts. *Curr. Opin. Chem. Biol.*, **2009**, *13*, 26-34.
- [241] Ekins, S.; Mestres, J.; Testa, B. *In silico* pharmacology for drug discovery: applications to targets and beyond. *Br. J. Pharmacol.*, **2007**, *152*, 21-37.
- [242] Becker, O.M.; Dhanoa, D.S.; Marantz, Y.; Chen, D.; Shacham, S.; Cheruku, S.; Heifetz, A.; Mohanty, P.; Fichman, M.; Sharadendu, A.; Nudelman, R.; Kauffman, M.; Noiman, S. An integrated *in silico* 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT_{1A} agonist (PRX-00023) for the treatment of anxiety and depression. *J. Med. Chem.*, **2006**, *49*, 3116-3135.
- [243] Shenoy, S.R.; Jayaram, B.; Latha, N.; Narang, P.; Jain, T.; Bhushan, K.; Shaikh, S.A.; Bose, S.; Sharma, P.; Singhal, P.; Gandhimathi, A.; Agrawal, P.; Pandey, V.; Dutta, S.; Sandhu, G.; Gupta, A.; Shekhar, S.; Tripathi, S. From gene to drug: a proof of concept for a plausible computational pathway. *Proceedings of 6th International Conference on Intelligent Systems Design and Applications (ISDA'06)*. *IEEE Computer Society*, **2006**, pp. 1147-1152.