Indian Journal of Biotechnology
Vol 3, October 2004, pp 473-485

Influence of protein structural similarities in adding value to genome data

B Anand, S Namboori, S Sandhya and N Srinivasan*
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Received 24 January 2004; accepted 11 March 2004

One of the central problems in post-genomic era is the understanding of function of myriad of putative gene products suggested by the genome sequencing projects. Computational approaches aimed at establishing the relationships between proteins, purely on the basis of their amino acid sequences, provide a rapid and useful first step. Sequence analysis methods, which use evolutionary information on protein families perform well in terms of detecting remote homologues. Use of three-dimensional (3-D) structures provides a further edge in detecting distantly related proteins as 3-D structures are conserved better than the amino acid sequences. Also, in many cases, similarity in the fold of proteins corresponds to gross similarity in functions. Hence, knowledge of 3-D structures has profound influence in identifying the functions of newly discovered gene products. This review covers recent developments in this area of homology detection and its influence in computational genomics.

Keywords: database searching, homology detection, protein evolution, protein structures, sequence analysis

IPC Code: Int. Cl.[7] G 01 N 33/00

Introduction

A critical problem confronting the present era of genome revolution is assignment of function and structure to newly discovered proteins. The exploding rate at which genomes are sequenced is a formidable challenge for experimental scientists who attempt biological and biochemical characterization of proteins. The realm of rapid, preliminary assignment of protein function is, therefore, a challenging one and several procedures and strategies have been developed in the last decade to complement the pace of genome sequencing.

Several effective experimental techniques are aimed to understand properties of proteins at the genomic scale. Microarray and protein expression profiles quantify transcription and translation of genes in an organism. Techniques, such as mass spectrometry and 2D-Gel, also serve as important tools for genome-wide analysis to characterize the gene products. Genome-wide yeast-two-hybrid analysis serves as a powerful tool to study interactions between proteins.

While these techniques provide a variety of information about genes encoded in a genome, the biochemical or biological functional characterization of such proteins is still not available for most of the proteins in various organisms. Preliminary characterization and assignment of protein function is often performed by relating newly discovered proteins to those proteins whose structure and biochemical function are well known[1-6]. These similarities are deduced at the level of amino acid sequences, performing pair-wise string comparisons between the protein sequences by the process of protein homology detection—a central tool in genomics. Further, *in-silico* approaches to identify interacting proteins include Rossetta approach[7], comparison of phylogenetic profiles[8] and chromosomal localization of genes[9].

A major problem in homology detection is the extensive divergence of amino acid sequences of homologous proteins during evolution[10]. Protein evolution is best studied at the level of domains, as domains appear to be the most common minimal functional and evolutionary module. Although amino acid sequences diverge less extensively than the base sequences of the genes, there are several known cases of very low sequence similarity between two divergently evolved protein domains. The sequence identity can be so low (even < 10%) that the two evolutionarily related protein domains appear as entirely different proteins. Detecting such distantly related proteins is challenging and is important since

*Author for correspondence:
Tel: 91-80-2932837; Fax: 91-80-3600535
E-mail: ns@mbu.iisc.ernet.in

evolutionarily related proteins often resemble in their functional properties. While very sensitive methods have been developed in the last few years to detect such related proteins, their similarity in 3-D structures can also be exploited in the identification of remote homologues[10-12]. Indeed, detection of homologues is central in comparative modelling of 3-D structures[13], which is now performed at the genomic scale[14].

The purpose of this review is to highlight the recent developments in homology detection with and without the use of 3-D structures, a key problem in bioinformatics. The authors also review its application in the genomic scale in order to obtain clues about functions of newly discovered gene products in the genome sequencing projects.

Detection of Similarity between Closely Related Proteins

Many related proteins have similar sequences making their relationship easy to detect. However, some sequences diverge to the extent that they have very little sequence similarity between them and detection of such remote homology necessitates more sensitive procedures. Pair-wise sequence comparison methods such as BLAST[15] and FASTA[16], which are based on dynamic programming, have been traditionally used in detecting homologies between closely related proteins. The dynamic programming method was first applied for global sequence alignments by Needleman and Wunsch[17] and for local sequence alignments by Smith and Waterman[18]. Briefly, these programs align a query sequence with all sequences in the database by following a scoring scheme for matches, mismatches and gaps and then maximize matches of all possible residue combinations between the sequences. This generates a matrix of numbers representing all possible alignments between the query and every target sequence. The highest set of sequential scores in the matrix represents the most optimal alignment between the two sequences. Thus, alignment algorithms find the best possible alignment between sequences by penalizing gaps and rewarding matches. For alignment of protein sequences, amino acid substitution matrices, such as the Dayhoff Percent accepted mutation matrix 250 (PAM 250) or the Block substitution matrix 62 (BLOSUM 62)[19], are used to score matches and mismatches. FASTA and BLAST are two implementations of the dynamic programming method and report those query-target pairs that represent a statistically significant match. A significant match is represented by a statistical score called the expectation (*E*)-value[20], a probability measure of an alignment appearing by chance. Lower the *E*-value more significant is the match, because it is less likely that such alignments appear by chance. These programs are very effective in detecting relationships between sequences having identities >30%. However, the sensitivity of detection of remote relationships becomes increasingly small when identities between the sequences are very poor. Brenner *et al*[21] demonstrated that only half as many evolutionary relationships are detected in proteins with 20-30% sequence identities by pair-wise comparison methods.

To overcome these limitations, procedures based on shared characteristics of sets of related proteins have been developed. Examples include templates[22-25], profiles[26-28], Hidden Markov Models[29-35], position specific iterated version of Blast (PSI-BLAST)[36] and intermediate sequence searches (ISS)[37].

Intermediate Sequences Enhance Capability of Detection of Homology

When sequences of related proteins diverge to such an extent that simple pair-wise search procedures fail to detect their relationship, it is possible to relate them through a third sequence whose sequence characteristics are intermediate between the two being matched[37,38]. High scores for a sequence match between the first and third sequence and between the second and third sequence imply that the first and the third sequence are related even though their own match score is low. The procedure, therefore, relates two distantly related proteins by collecting from a large databank of homologues that match both with high significant scores. These intermediates with high match scores were detected using pair-wise procedures such as FASTA.

The usefulness of this procedure was demonstrated on a database containing 971 protein domains whose structures are known and residue identities with each other are < 40%. On the basis of sequence and structural information, 2143 pairs of these sequences are known to have evolutionary relationships. Whilst FASTA, in an all against all comparison of the sequences in the database, detected only 15% of these relationships, an intermediate sequence procedure using FASTA on these sequences against the non-redundant OWL database[39] increased the sensitivity of the detection procedure by 70%.

Multiple Intermediates in Detection of Remote Relationships

Salamov *et al*[40] used data from CATH database[11] to assess the BLASTP, FASTA, Smith-Waterman, gapped blast algorithms and intermediate sequence based approaches. The sensitivity and specificity of these approaches were estimated by performing an all against all comparison within a dataset of 841 protein domains with varying sequence identities that was derived from the CATH database. Of all the possible pairs, 3442 relationships are known to be true from structural and known evolutionary information. The intermediate sequence based procedure was assessed for the same dataset by querying the non-redundant OWL database in a manner similar to the Park *et al*[37] approach. It clearly demonstrates that among the direct pair-wise comparison procedures blastp performs the worst and detects only the most obvious similarities, while the performance of gapped blast is marginally better. The variation of sensitivity and specificity when tested by choosing BLOSUM 62 and BLOSUM 50 matrices reflected that the choice of the matrix has minimal effect on the results. It is also evident from Fig. 1 that the increase in number of relationships detected by the intermediate sequence based approach is much more than that of gapped blast alone.

The ability of intermediate sequence based approaches to detect remote relationships depends on the availability of intermediates and is family specific. Well-populated families fair better than poorly populated families of proteins. Given the benefits of such intermediates, Salamov *et al*[40] have demonstrated that using more than one intermediate is more effective than single intermediates in detecting relationships by repeatedly performing searches using true intermediates within the OWL database until a relationship to a remote target is detected.

It has become increasingly obvious, therefore, that even within a single group of proteins with common origin, a single pair of sequences can differ quite substantially in their composition. Using the properties and features of a set of related sequences in place of one related sequence, in a manner similar to Park *et al*[37] and Salamov *et al*[40], clearly performs better than direct pair-wise programmes that strongly depend on the properties of a single sequence. The use of evolutionary information to create a multiple sequence alignment helps to detect more remote relationships.

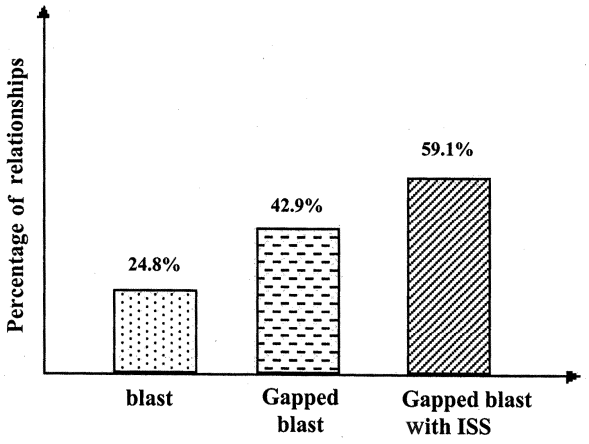

Fig. 1—Comparison of the number of relationships detected by blast, blast-pgp and blast-pgp with ISS. Graph derived from an assessment of sequence search procedures[40].

Using PSI-BLAST and Evolutionary Information to Detect Remote Relationships

There are several ways by which one might use evolutionary relatedness to detect remote homologies. A method, which has become increasingly popular in the last several years, is a flavour of gapped blast called PSI-BLAST[36].

PSI-BLAST is a powerful profile based sequence search procedure, which constructs a multiple alignment of all homologues detected in the first round of a gapped blast. A position-dependent weight-matrix is generated on the basis of the multiple alignments of all related homologues of a query and includes information from those sequences whose *E*-value is more significant than a given inclusion threshold. These weight matrices are then used in recurring steps of searching and profile building in the databases until no more sequences satisfying the inclusion thresholds are detected for inclusion into the profile. Thus, the searches made within the database are searches made with a family of proteins closely related to the query instead of using the sequence information of the query alone. Clearly, the power of the technique lies through such an iterative procedure and results in its ability to detect more distantly related proteins to the query.

The protein family database as exemplified by the PFAM[41,42] uses PSI-BLAST to organize protein sequences into families. Related proteins that share high sequence identities and common evolutionary origins are grouped together into families. New members of a family are detected by performing PSI-BLAST searches from an aligned family.

Muller *et al*[43] used a structural benchmark developed by Chothia and coworkers from the SCOP to assess the accuracy of homology based annotation of ORFs. They have benchmarked the coverage and error rate of genome annotation using PSI-BLAST. To briefly describe their assessment, PSI-BLAST was used to perform iterative searches against a non-redundant sequence database that includes every non-identical representative from the standard sequence databases together with the sequences of 1625 SCOP domains, which have less than 40% identity with respect to each other. Based on the assessment of the number of correct assignments, an *E*-value cut of 5 x 10^{-4} is considered stringent for correct assignment of function to the ORFs.

Hidden Markov Models and their Performance

Another profile-based method, which is very sensitive in its ability to detect remote homologues is the profile-based Hidden Markov Model[29-32,35,44-48]. There are essentially three steps in the construction of the Profile HMMs. In the first step, a multiple sequence alignment is made from known members of a given protein family. The quality of the alignment and the diversity of the sequences are crucial for the success of the profile. The multiple sequence alignment is used in the building of the profile HMM for that family of sequences. A model-scoring program then assigns a score for any sequence of interest with respect to the model. The sequence alignment and modelling (SAM-T98) method is another profile based HMM method that creates an initial HMM from a single given query sequence by iteratively finding homologues in a protein database[49]. A database of potential homologues is then created by searching a large database using a pair-wise sequence search method such as WU-BLAST[50] with a low score cut off. An iterative search for homologues of the query which have good local alignment scores to the HMM is made within this large database of potential homologues. The homologues, thus, detected at every iteration are used to refine the HMM and a new multiple alignment is generated for the query and all its homologues using sequence weighting and Dirichlect mixture priors[51]. The final HMM may then be used to search any database for homologies of the query sequence. A HMM trained on sequences, which are members of a protein family, can identify the positions of amino acids that describe conserved primary structure of the family. This HMM can be used to discriminate between family and non-family members in searches of sequence databases.

Assessment of ISS, PSI-BLAST and SAM-T98

Park *et al*[38] assessed the performance of the intermediate sequence based procedures, such as ISS, PSI-BLAST and SAM-T98, by determining the extent to which these procedures can detect evolutionary relationships between the members of the sequence database PDBD40-J. This database derived from SCOP contains the sequences of proteins of known structures whose sequence identities with each other are 40% or less. The evolutionary relationships that exist between these sequences with poor identities are found by examination of their structural and functional features. Of the 4,32,680 relationships, at a false positive rate of 1/50,000, SAM-T98 found 35% of the true homologous relationships, PSI-BLAST found 30% and ISS found 25%. This is twice the number of PDBD40-J relations that can be detected by pair-wise comparison procedures such as FASTA (17%) and GAP-BLAST (15%). Interestingly, for distantly related sequences whose identity is less than 30%, SAM-T98 and PSI-BLAST detected three times the number of relationships than pair-wise methods.

It is evident that the pair-wise search procedures for detecting close homologues and the profile based remote homology methods, such as HMM and PSI-BLAST, have different strengths. The direct search methods are the fastest with the runtime linear to the size of the database, while the iterative methods are slower by the number of iterations made. The utilities of these procedures depend upon the aim of the search procedure. Whereas, pair-wise search procedures detect closely related proteins, distant relationships are best detected by a linking method using multiple sequence information, such as the ISS procedure. The most distant homologies are detected effectively by search methods such as PSI-BLAST and HMMs.

Several genome wide analyses using these tools have added value to the genomic data. For example, extended genome wide analysis of *Plasmodium falciparum* has been recently reported[52]. Specific protein families have also been studied, such as G-proteins[53] and protein kinases[54,55]. Despite these developments, many previously unknown similarities between two or more proteins are detected only after the availability of 3-D structures determined using X-ray analysis or Nuclear Magnetic Resonance (NMR).

Similarities among Proteins in Three Dimensional Structures

Detection of relationship between amino acid sequences becomes very difficult if the sequence identity falls below 25%[56-58]. However, the proteins could exhibit the similarity in three dimensional structures inspite of low sequence identity[59,60]. Hence, knowledge about three dimensional structure of proteins can help in tracing their evolutionary relationships and in getting clues about their biochemical functions[61,62].

The information about the 3-D structure of proteins is translated from the amino acid sequence using the stereochemical codes that are not yet fully understood[63]. Similarity, in three dimensional structures, the absence of significant sequence similarities can be attributed to the stereochemical constraints in the three dimensional space[64-67]. A structural domain is a polypeptide chain or a part of the polypeptide chain that can fold into a compact and stable 3-D structure[68]. Domains are the fundamental units of tertiary structure and it can occur either as a single stretch along the polypeptide chain (continuous domain) or more than one stretch coming together in the three dimensional space (discontinuous domain).

The recent advances in the structure determination methods have enabled the exponential growth of protein structures (Fig. 2). Currently, the total number of structures in the Protein Data Bank exceeds 23000, however, the growth in the number of new folds identified so far is not at par with the growth in the total number of structures available (Fig. 3). This disparity can be ascribed not only to the extensive divergent evolution of protein domains but also to the convergent evolution, in which two evolutionarily unrelated proteins adopt the same fold and similar function[69]. The burial of hydrophobic residues, preferred packing of secondary structural elements coupled with specific topological connections, restrict the conformational space available to the protein domains[70-72]. The limit in the number of folds is thus explained by the convergent and divergent evolution of protein structures.

Divergent evolution often involves gene duplication and fusion resulting in multi-domain proteins. In a multi-domain protein, such as bovine liver rhodanese (Fig. 4), fusion of duplicated genes results in a single polypeptide chain that can fold into two similar domains. In some multi-domain proteins, such as

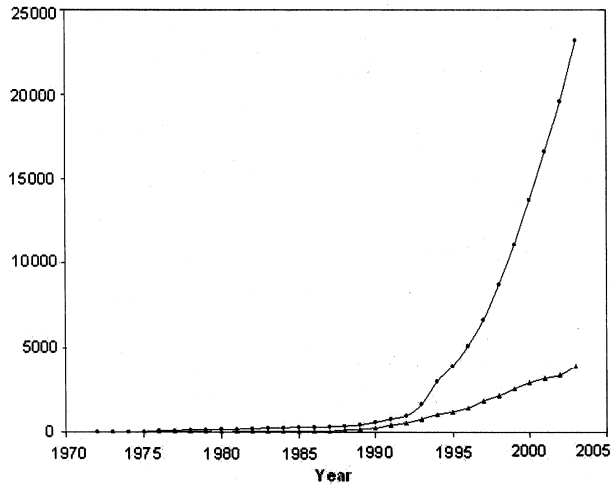

Fig. 2—The growth of structures deposited in PDB. Line connecing the bullets represent the total number of structures available in PDB at the current time and those with triangles depict the number of structures deposited for the year.

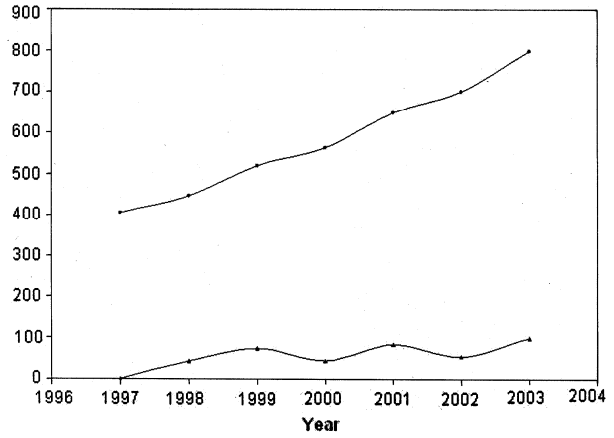

Fig. 3—The growth of newly discovered folds in SCOP. Line with bullets represent the total number of folds present in SCOP and those with triangles depict the number of new folds discovered in the year.

glyceraldehyde-3-phosphate dehyrogenase, the two domains are highly dissimilar.

The understanding and access to the voluminous structural information requires an effective organization and classification[59,60]. SCOP is one such database which provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose 3-D structures have been determined[10]. Here, the unit of classification is domain. The classification is hierarchical, symbolising the evolutionary and structural relationships. The proteins with high sequence identities and similarities in structures and

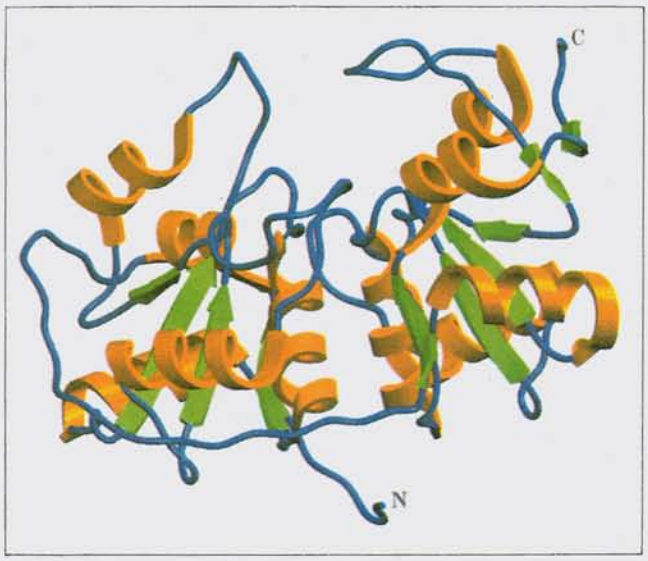

Fig. 4—The domain duplication and fusion observed in rhodanese. Helices, strands and loops are represented in different colours. This figure as well as Figs 5, 6 & 7 were prepared using SETOR[109].

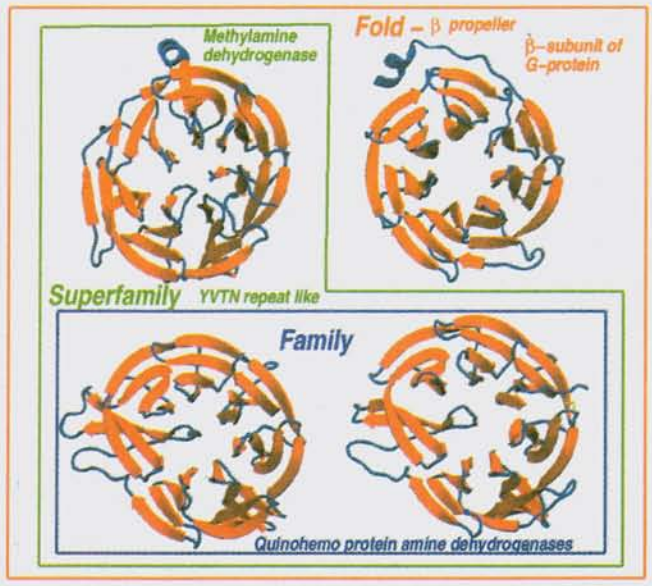

Fig. 5—The hierarchical classification of fold, superfamily and family in SCOP. The structures encircled in blue belong to same family and those that are encircled in green are connected by same superfamily relation. All the structures share a common fold (encircled in red) named β-propeller.

functions are grouped into families. Families whose structural and functional features suggest a probable common evolutionary origin form a group called superfamily. Families within the superfamily need not have significant sequence identities. Superfamilies and families having common secondary structural elements and topological connections are defined as members of the same fold. The members belonging to the same fold have a similar secondary structural arrangement but with very poor sequence identities and functional relationships (Fig. 5).

The different folds, based on the composition and sequential arrangement of secondary structures, are broadly classified into six catagories: (i) all α-proteins whose structures are primarily composed of α helices, (ii) all β-proteins whose structures are essentially formed of β sheets, (iii) α and β-proteins whose structures tend to alternate between α helices and β sheets forming β-α-β units, (iv) α plus β-proteins whose structures are formed by the segregation of α helices and β sheets, (v) small proteins-proteins with little or no regular secondary structures exhibiting several disulphide bridges or metal-cys interactions, and (vi) multi domain-proteins with domains of different fold for which no homologues are known at present.

The structural genomics initiative is expected to further the accumulation of structural data[73]. The number of proteins of known amino acid sequences outnumber the number of proteins of known 3-D structures[74]. This gap can be reduced considerably by establishing connections between the proteins of known 3-D structures and proteins of known amino acid sequences[75]. The comparison between the 3-D structures of homologous proteins can give a measure about its topological equivalences, spatial orientations and structural divergence[76,77]. This information can be extrapolated in modelling the proteins of unknown structures. The topological equivalence between the homologous protein structures can be found by the rigid body superposition of the Cα positions in the alignment[78]. The secondary structural elements are usually well conserved and it can be used to construct the framework for the model building[79]. The loop regions are generally highly variable and it can accommodate replacements, insertions and deletions. Hence, the modelling of loop regions is relatively difficult and it can be achieved by considering the various geometrical problems such as the clashing of loops with the rest of the protein[80]. The main chain and side chain conformers from the equivalent fragments of known protein structures can be extended to the model sequence[81,82].

The knowledge and understanding of the 3-D structure of proteins is crucial in identifying new folds and in predicting the structure from one dimensional

sequence. Since we are confined to the limited number of folds[62], the conformational search is considerably reduced. Hence, the methods, which use 3-D structure-1-D sequence match, attempt to recognise the protein folds[74].

Why Fold Recognition?

It is now well established that proteins could adopt similar 3-D structure even when no similarity at the level of sequences can be detected. This feature is often referred to as the proteins sharing the same fold[10,83]. There are a number of proteins whose homologues cannot be detected readily by sequence-based search methods because of absence of obvious sequence similarity. Moreover, lately there has been a surge in the number of completely sequenced genomes and this contributes to a glaring difference in the number of proteins with known structures and those for which the structures have to be determined experimentally. Thus, the procedures for protein structure prediction provide a much-needed tool for bridging this gap and also for detecting homologues in instances where even the powerful and sensitive sequence-based profile search methods like PSI-BLAST[36] prove ineffective. In these cases, it becomes imperative to look for structure-based methods for homologue detection so that the structure of a new protein of a particular gene product can be predicted.

Most of the fold recognition methods assess the fitting of a query sequence to a library of structural templates by considering a combination of factors, such as secondary structure, solvent accessibility or residue–residue contact preferences[84-88]. Fold recognition methods have achieved considerable improvement and progress in recent years[89] and, hence, in their ability to assign a given sequence to the correct fold. The following section describes the basic principle behind some of the most frequently used fold recognition methods, which perform fast fold predictions and are designed to automatically screen genomic databases[90]. Their limitations and application to genome-wide fold assignment are also discussed.

Methods for Fold Recognition

3D-PSSM (position-specific scoring matrix)[91] is a fold recognition method that combines the multiple sequence profiles with protein structural information in order to provide enhanced fold recognition for newly sequenced genomes. It uses structural alignment of homologous proteins of similar 3-D structure from SCOP[10] database to obtain structural equivalences of residues, which are then used to extend the multiply aligned sequences obtained by standard sequence search methods. The outcome is a large superfamily based multiple alignment, which is then converted into a PSSM. This, in addition to the secondary structural alignment and solvation potentials, generates 3D-PSSMs that are capable of rendering structural as well as functional assignment to proteins of unknown structures in various genomes.

GenTHREADER[87] is another fold assignment method that uses the sequence-structure alignment approach originally developed by Jones *et al*[92]. The fold library used here is derived from a unique set of the proteins found in Protein Data Bank[93,94]. There are three stages of alignment, viz. alignment of sequences, calculation of pair potential and solvation terms and evaluation of the alignment using neural network. The sequence-structure alignment of a target sequence and a template protein structure is generated using a sequence profile method. The profile for each template structure in fold library is generated by collecting related sequences, which is done by scanning the template sequence against the OWL non-redundant protein sequence data bank[39] using the program BLASTP[15]. Below a specific *E*-value cut-off the sequences matching the template sequence are extracted and a multiple sequence alignment is generated, from which a sequence profile is constructed using the BLOSUM 50 matrix[19]. Alternatively, instead of the template structure, a sequence profile is generated using the target protein sequence and this profile is scanned against the fold library. Of the two approaches, the one producing the highest scoring alignment is used for further evaluation. Evaluation of the sequence alignment with respect to the structural model is done by using the two threading potentials: pairwise-potentials and solvation potential. Finally, six scores are obtained, viz. initial sequence alignment score, number of aligned residues, length of target sequence, length of template protein sequence, pairwise energy sum and solvation energy sum. A neural network model is applied for optimizing the combination of scores and to arrive at a single measure of confidence.

UCLA-DOE fold server is based on a sequence-to-structure match method[74]. This is so since any given structure is encoded as a sequence of residue environment that is described in terms of secondary structure, solvent accessibility and extent of polar

environment around the residue. Each residue receives a score for occupying a particular type of environment. The sum of all residue scores is a gross measure of the fitting of the sequence to a fold. The UCLA-DOE Directional Atomic Solvation EnergY (DASEY) method for fold assignment to protein sequences uses this concept[95]. DASEY is an atom-based description of the environment of a residue position within a known 3-D protein structure. Further, the preference of each residue of the probe sequence is computed along with its secondary structure prediction for the alignment with the template environment. The database of aligned protein structures is obtained from domain database by creating domain pairs for each domain-fold family in CATH[11]. The Fold classification based on Structure-Structure alignment of Proteins (FSSP) database is used to verify that domain pairs share a common fold[96].

While the methods described above are early developments in fold recognition, strategies have been proposed recently to identify the fold by sophisticated sequence analysis[89,97]. These approaches use the knowledge of known common fold of amino acid sequences without using details of the structure.

Some Limitations of Automated Fold Prediction Methods

One obvious limitation of automated fold assignment methods is that the predictions are highly dependent on the database of known protein structures, i.e. new folds cannot be predicted. It is also imperative to realize that any fold assignment does not automatically mean that the protein concerned performs the same function as that of the matching fold. Fold assignment only suggests that there is a good probability that the particular protein performs the same function as that of the assigned fold, especially so if it also has a good measure of sequence match in the functional site. Functions can be directly extrapolated for homologous proteins that share a common ancestor, lesser so for analogous, that may not originate from the same ancestor. The reason for obtaining structures of proteins encoded in a genome is to understand the functions of proteins encoded and to understand further, the biology of that organism. Thus, it is possible to predict structures and test them with methods like mutagenesis experiments. Additionally, functional relationships can be deduced in cases where it cannot be detected using sequence alone[98]. Other than these, there are those inherent limitations, which are existent in any automated method like the score cut-off that most methods use. It is true that the link to a structure with a known fold should be at a high confidence for it to be plausible but that does not mean that those with lower scores are always wrong. In these cases, probably, the decision of whether the assignment is correct or not would depend on manual intervention and analysis of the results of many fold recognition methods, which is difficult for a large-scale genome fold assignment. Hence, a body for evaluation of structure prediction methods including the fold recognition method critical assessment of protein structure prediction (CASP) was formed in 1994[99]. CASP provides a benchmark as to what level of model accuracy can one expect from the structure prediction approaches, i.e. it assesses when a method works, when it fails and how it can be improved. Accordingly, an interesting observation was that good fold recognition depends on the template selection and alignment techniques to a large extent.

Examples of Genome-wide Fold Assignments

The usefulness of fold recognition procedures is most appreciated when a new genome is sequenced and it becomes desirable to predict the structure and function for the proteins in the genome. Thus, number of groups involved in development of fold recognition methods have done fold assignments for various genomes. *Mycoplasma genitalium* is one such genome for which fold assignment has been carried out[87,100]. *M. genitalium* constitutes one of the smallest genomes and it consists of 468-protein open reading frames (ORFs). About 40% of the proteins encoded have been associated with a protein of known structure, though in some cases only a single domain could be predicted. It is worth noting that some of the proteins with good sequence similarity could even be picked up by sequence comparison alone but a significant number of structure could not be assigned if the pairwise and solvation potentials are not used. Another observation suggests that the architecture of assigned folds is quite similar to that observed in the PDB, which gives rise to the suggestion that the folds in PDB might constitute a representative set and may not be biased towards particular folds and families as is a common belief[87]. The 3-D protein folds have also been assigned to the proteins encoded in the hyperthermophilic archaeon, *Pyrobaculum aerophilum.* Of about 2,681 ORFs predicted to code

for this organism, 916 matched a fold at 90% confidence level and 245 could be assigned at a 99% confidence level. Likewise, 286 proteins were predicted to have a previously unobserved fold with a 90% confidence level, and 14 at a 99% confidence level[101]. Consequently, the apparently novel folds present attractive targets for crystallographic or NMR structure determination. Sternberg and coworkers[102] have assigned structures to many of the human proteins and analyzed the functional features of those proteins involved in various diseases.

Mycobacterium tuberculosis H37Rv Genome Structural Assignments

Structural and functional assignments for proteins encoded in *M. tuberculosis* H37Rv genome have been performed by our group (Namboori *et al*, submitted for publication). For establishing structural relationships, integrated structure-sequence PALI[76] database has been used, which is a alignment database of homologous proteins of known structure that is largely derived from SCOP by using sensitive profile-based search methods, like IMPALA[103]. It was possible to link a number of proteins to specific folds and the details are available at the web site: http://hodgkin.mbu.iisc.ernet.in/~dots. For example, the hypothetical protein Rv0498 found a match (*E*-value of 7e-28 and sequence identity of 16%) with the PDB entry 1qtw[104], which is an endonuclease in the xylose isomerase superfamily belonging to TIM β/α-barrel fold (Fig. 6). Endonuclease IV is the archetype for a conserved apurinic/apyrimidinic (AP) endonuclease family that primes DNA repair synthesis by cleaving the DNA backbone 5′ of AP sites. The enzyme specifically cleaves the DNA backbone at AP sites and also removes 3′ DNA-blocking groups, such as 3′ phosphates, 3′ phosphoglycolates, and 3′α,β-unsaturated aldehydes that are formed as a result of oxidative base damage and the combined activity of glycosylase/lyase enzymes[105,106]. Since Rv0498 belongs to the same fold and superfamily, and functionally important residues are conserved, it could be postulated that hypothetical protein Rv0498 performs a similar function of DNA repair. Another hypothetical protein Rv3075c showed reasonable alignment (sequence identity 16%, *E*-value of 8e-20) with PDB entry 1dxe[107], which is a 2-dehydro-3-deoxy-galactarate (DDG) aldolase belonging to phosphoenolpyruvate/pyruvate domain superfamily in the TIM β/α-barrel fold (Fig. 7). DDG

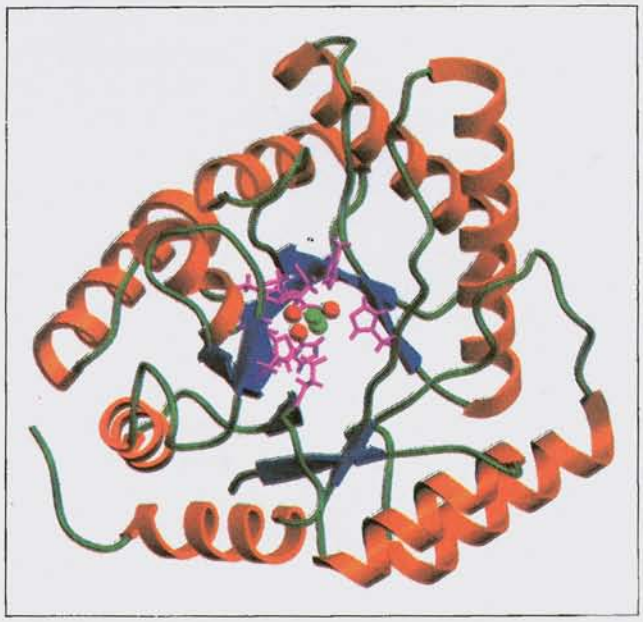

Fig. 6—Recognition of function of a hypothetical protein, Rv0498, from *M. tuberculosis*. Sequence of Rv0498 could be matched reliably with an endonuclease of known structure[104], which is shown. The two critical water molecules are shown as green spheres and the three zinc atoms are shown as red spheres. The sidechains of functionally important residues, which are conserved in the *M. tuberculosis* homologue, are shown in pink. The residues are: His69. His109. Glu145, Glu 261, Asp 179, His 216, His 182, His 231 and Asp 229. (PDB residue numbering is followed).

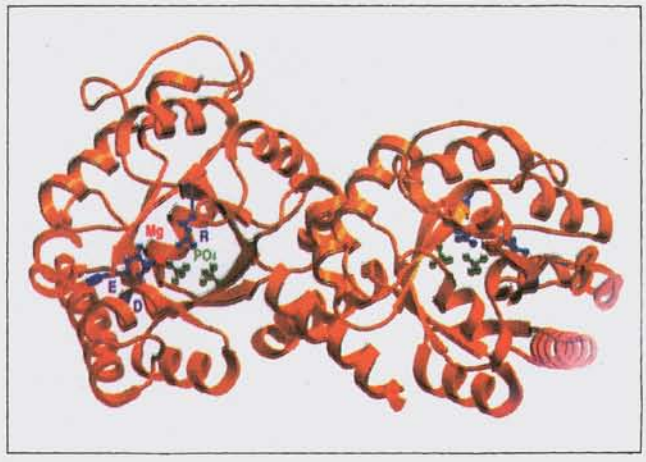

Fig. 7—Function recognition of a hypothetical protein Rv3075c from *M. tuberculosis*. Sequence of this hypothetical protein is found to be consistent with the fold of 2-Dehydro-3-Deoxy-Galactarate (DDG) Aldolase[107] which is shown in the ribbon representation. The critical functional residues of DDG that are conserved with Rv3075c are shown in blue. The phosphates and magnesium are shown in green and red respectively.

aldolase catalyses the reversible aldol cleavage of DDG to pyruvate and tartronic semialdehyde. The enzyme is part of the catabolic pathway for D-glutarate/galactarate utilisation in *Escherichia coli*[108]. Aldolases have been proved effective in biotransformations and synthesis of novel antibiotics. Significantly, the catalytically essential residues are conserved in hypothetical protein Rv3075c and, hence, it is possible that it could be involved in a similar biochemical function.

Outlook

Last few decades have witnessed spectacular development and growth of several laboratory techniques to characterize the functions of proteins. Myriad of proteins with known amino acid sequences has been studied. On the other hand, recent sensation of genome sequencing projects has resulted in explosion of sequence information for a large number of proteins, which are not even produced yet in a laboratory. Computational biology approach attempts to bridge this gap. The entire approach relies on our ability to detect similar sequence patterns between protein under study and the proteins that are well characterized. Sophisticated sequence analysis techniques perform this task ably. However, several non-trivial relationships become obvious only with the use of 3-D structures of proteins. With global structural genomics projects holding the promise of delivering huge number of protein structures to fill-in the gaps in protein fold space, use of 3-D structures in closing the gap between mere sequence data and structural data will continue to be attractive. However, a word of caution is appropriate because mere availability of 3-D structures without biochemical and biological characterization of proteins would not permit an effective use of 3-D structures. Traditional and modern approaches to characterize protein functions is a critical component in using 3-D structures to their full potential.

Much progress has already been made and planned towards obtaining a fair idea about repertoire of protein folds and functions. However, an exhaustive structural characterization of interactions between proteins still appears remote. Despite spectacular growth in the number of protein structures over the years, the number of macromolecular complex structures is only meager. Difficulties in obtaining stable crystals for X-ray studies, and too large a size for NMR studies, are the main reasons. Computational modelling approaches could make the best use of the small number of structures of protein-protein complexes and protein assemblies in continuing to add value to the genomic data.

Acknowledgement

We thank all the members of our group for their inputs. Research in NS group is primarily supported by the Wellcome Trust, London. NS is an International Senior Fellow of the Wellcome Trust. BA and SS are supported by the Wellcome Trust and SN is supported by the Department of Biotechnology, New Delhi in the computational genomics project.

References

1 Bork P & Koonin E V, Predicting functions from protein sequences—where are the bottlenecks? *Nat Genet*, 18 (1998) 313-318.

2 Chothia C & Lesk A M, The relation between the divergence of sequence and structure in proteins, *EMBO J*, 5 (1986) 823-826.

3 Hegyi H & Gerstein M, The relationship between protein structure and function: A comprehensive survey with application to the yeast genome, *J Mol Biol*, 288 (1999) 147-164.

4 Karp P D, What we do not know about sequence analysis and sequence databases, *Bioinformatics*, 14 (1998) 753-754.

5 Martin A C, Orengo C A, Hutchinson E G, Jones S, Karmirantzou M *et al*, Protein folds and functions, *Struct Fold Des*, 6 (1998) 875-884.

6 Russell R B, Saqi M A, Bates P A, Sayle R A & Sternberg M J, Recognition of analogous and homologous protein folds: Assessment of prediction success and associated alignment accuracy using empirical substitution matrices, *Protein Eng*, 11 (1998) 1-9.

7 Strong M, Graeber T G, Beeby M, Pellegrini M, Thompson M J *et al*, Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps, *Nucleic Acids Res*, 31 (2003) 7099-7109.

8 Zheng Y, Roberts R J & Kasif S, Genomic functional annotation using co-evolution profiles of gene clusters, *Genome Biol*, 3 (2002) R60.

9 Strong M, Mallick P, Pellegrini M, Thompson M J, Eisenberg D *et al*, Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: A combined computational approach, *Genome Biol*, 4 (2003) R59.

10 Murzin A G, Brenner S E, Hubbard T A & Chothia C, SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, 247 (1995) 536-540.

11 Orengo C A, Michie A D, Jones S, Jones D T, Swindells M B *et al*, CATH-a hierarchic classification of protein domain structures, *Structure*, 5 (1997) 1093-1108.

12 Holm L & Sander C, Protein structure comparison by alignment of distance matrices, *J Mol Biol*, 233 (1993) 123-138.

13 Johnson M S, Srinivasan N, Sowdhamini R & Blundell T L, Knowledge-based protein modelling, *Crit Rev Biochem Mol Biol*, 29 (1994) 1-68.

14 Sanchez R & Sali A, Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome, *Proc Natl Acad Sci USA*, 95 (1998) 13597-13602.

15 Altschul S F, Gish W, Miller W, Myers E W & Lipman D J, Basic local alignment search tool, *J Mol Biol*, 215 (1990) 403-410.

16 Pearson W R & Lipman D J, Improved tools for biological sequence comparison, *Proc Natl Acad Sci USA*, 85 (1988) 2444-2448.

17 Needleman S B & Wunsch C D, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J Mol Biol*, 48 (1970) 443-453.

18 Smith T F & Waterman M S, Identification of common molecular subsequences, *J Mol Biol*, 147 (1981) 195-197.

19 Henikoff S & Henikoff J G, Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci USA*, 89 (1992) 10915-10919.

20 Karlin S & Altschul S F, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc Natl Acad Sci USA*, 87 (1990) 2264-2268.

21 Brenner S E, Chothia C & Hubbard T J, Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc Natl Acad Sci USA*, 95 (1998) 6073-6078.

22 Taylor W R, Identification of protein sequence homology by consensus template alignment, *J Mol Biol*, 188 (1986) 233-258.

23 Bashford D, Chothia C & Lesk A M, Determinants of a protein fold: Unique features of the globin amino acid sequences, *J Mol Biol*, 19 (1987) 199-216.

24 Tatusov R L, Altschul S F, Koonin E V, Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks, *Proc Natl Acad Sci USA*, 9 (1994) 12091-12095.

25 Yi T M & Lander E S, Recognition of related proteins by iterative template refinement (ITR), *Protein Sci*, 3 (1994) 1315-1328.

26 Gribskov M, McLachlan A D & Eisenberg D**,** Profile analysis: Detection of distantly related proteins, *Proc Natl Acad Sci USA*, 84 (1987) 4355-4358.

27 Luthy R, Xenarios I & Bucher P, Improving the sensitivity of the sequence profile method, *Prot Sci*, 3 (1994) 139-146.

28 Thompson J D, Higgins D G & Gibson T J, Improved sensitivity of profile searches through the use of sequence weights and gap excision, *Comput Appl Biosci*, 10 (1994) 19-29.

29 Krogh A, Brown M, Mian I S, Sjolander K & Haussler D, Hidden Markov models in computational biology: Applications to protein modelling, *J Mol Biol*, 235 (1994) 1501-1531.

30 Baldi P, Chauvin Y, Hunkapiller T & McClure M A, Hidden Markov models of biological primary sequence information, *Proc Natl Acad Sci USA*, 91 (1994) 1059-1063.

31 Eddy S R, Multiple alignment using hidden Markov models, *Proc Int Conf Intell Syst Mol Biol*, 3 (1995) 114-20.

32 Eddy S R, Hidden Markov models, *Curr Opin Struct Biol*, 6 (1996) 361-365.

33 Eddy S R, Profile hidden Markov models, *Bioinformatics*, 14 (1998) 755-763.

34 Eddy S R, Mitchison G & Durbin R, Maximum discrimination hidden Markov models of sequence consensus, *J Comput Biol*, 2 (1995) 9-23.

35 Hughey R & Krogh A, Hidden Markov models for sequence analysis: Extension and analysis of the basic method, *Comput Appl Biosci*, 12 (1996) 95-107.

36 Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z *et al*, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res*, 25 (1997) 3389-3402.

37 Park J, Teichmann S A, Hubbard T & Chothia C, Intermediate sequences increase the detection of homology between sequences, *J Mol Biol*, 273 (1997) 349-354.

38 Park J, Karplus K, Barrett C, Hughey R, Haussler D *et al*, Sequence comparisons using multiple sequences detect three times as many remote homologues as pair-wise methods, *J Mol Biol*, 284 (1998) 1201-1210.

39 Bleasby A J, Akrigg D & Attwood T K, OWL—A non-redundant composite protein sequence database, *Nucleic Acids Res*, 22 (1994) 3574-3577.

40 Salamov A A, Suwa M, Orengo C A & Swindells M B, Combining sensitive database searches with multiple intermediates to detect distant homologues, *Protein Eng*, 12 (1999) 95-100.

41 Sonnhammer E L, Eddy S R & Durbin R, Pfam: A comprehensive database of protein domain families based on seed alignments, *Proteins*, 28 (1997) 405-420.

42 Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L *et al*, The Pfam protein families database, *Nucleic Acids Res*, 30 (2002) 276-280.

43 Muller A, MacCallum R M & Sternberg M J, Benchmarking PSI-BLAST in genome annotation, *J Mol Biol*, 293 (1999) 1257-1271.

44 Karplus K, Sjolander K, Barrett C, Cline M, Haussler D *et al*, Predicting protein structure using hidden Markov models, *Proteins Suppl*, 1 (1997) 134-139.

45 Karplus K, Barrett C & Hughey R, Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, 14 (1998) 846-856**.**

46 Bucher P, Karplus K, Moeri N & Hofmann K, A flexible motif search technique based on generalized profiles, *Comput Chem*, 20 (1996) 3-23.

47 Grundy W N, Bailey T L, Elkan C P & Baker M E, Meta-MEME: Motif-based hidden Markov models of protein families, *Comput Appl Biosci*, 13 (1997) 397-406.

48 Karchin R & Hughey R, Weighting hidden Markov models for maximum discrimination, *Bioinformatics*, 14 (1998) 772-782.

49 Karplus K, Barrett C & Hughey R, Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, 14 (1998) 846-856.

50 Lopez R, Silventoinen V, Robinson S, Kibria A & Gish W, WU-Blast2 server at the European Bioinformatics Institute, *Nucleic Acids Res*, 31 (2003) 3795-3798.

51 Sjolander K, Karplus K, Brown M, Hughey R, Krogh A *et al*, Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology, *Comput Appl Biosci*, 12 (1996) 327-345.

52 Aravind L, Iyer L M, Wellems T E & Miller L H, *Plasmodium* biology: Genomic gleanings, *Cell*, 115 (2003) 771-785.

53 Pandit S B & Srinivasan N, Survey for G-proteins in the prokaryotic genomes: Prediction of functional roles based on classification, *Proteins*, 52 (2003) 585-597.

54 Krupa A & Srinivasan N, The repertoire of protein kinases encoded in the draft version of the human genome: Atypical variations and uncommon domain combinations, *Genome Biol*, 3 (2002) 66.1-66.14.

55 Manning G, Whyte D B, Martinez R, Hunter T A & Sudarsanam S, The protein kinase complement of the human genome, *Science*, 298 (2002) 1912-1934.

56 Koehl P, Protein structure similarities, *Curr Opin Struct Biol*, 11 (2001) 348-353.

57 Doolittle R F, *Of urfs and Orfs: A primer on how to analyze derived amino acid sequences* (University Science Books, Mill Valley, CA) 1986.

58 Sander C & Schneider R, Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins*, 9 (1991) 56-68.

59 Levitt M & Chothia C, Structural patterns in globular proteins, *Nature (Lond)*, 261 (1976) 552-558.

60 Richardson J S, The anatomy and taxonomy of protein structure, *Adv Prot Chem*, 34 (1981)167-339.

61 Murzin A G & Chothia C, Protein architecture: New superfamilies, *Curr Opin Struct Biol*, 2 (1992) 895-903.

62 Chothia C, One thousand families for the molecular biologist, *Nature (Lond)*, 357 (1992) 543-544.

63 Anfinsen C B, Principles that govern the folding of protein chains, *Science*, 181 (1973) 223-230.

64 Chothia C, Levitt M & Richardson D, Structure of proteins: Packing of α helices and β sheets, *Proc Natl Acad Sci USA*, 74 (1977) 4130-4134.

65 Richardson J S, Handedness of crossover connections in β sheets, *Proc Natl Acad Sci USA*, 73 (1976) 2619-2623.

66 Richardson J S, β sheet topology and the relatedness of proteins, *Nature (Lond)*, 268 (1977) 495-500.

67 Sternberg M J E & Thornton J M, On the conformation of proteins: The handedness of the β strand-α helix-β strand unit, *J Mol Biol*, 105 (1976) 367-382.

68 Wetlaufer D B, Nucleation, rapid folding and globular intrachain regions in proteins, *Proc Natl Acad Sci USA*, 70 (1973) 697-701.

69 Blundell T L & Johnson M S, Catching a common fold, *Protein Sci*, 2 (1993) 877-883.

70 Russell R B & Barton G J, Structural features can be unconserved in proteins with similar folds: An analysis of side-chain to side-chain contacts secondary structure and accessibility, *J Mol Biol*, 244 (1994) 332-350.

71 Chothia C, Principles that determine the structure of proteins, *Annu Rev Biochem*, 53 (1984) 537-572.

72 Finkelstein A V & Ptitsyn O B, Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol*, 50 (1987) 171-190.

73 Kim S H, Shining a light on structural genomics, *Nat Struct Biol*, 5 (1998) 643-645.

74 Bowie J U, Luthy R & Eisenberg D, A method to identify protein sequences that fold into a known three dimensional structure, *Science*, 253 (1991) 164-170.

75 Sali A, Overington J P, Johnson M S & Blundell T L, From comparisons of protein sequences and structures to protein modelling and design, *Trends Biochem Sci*, 15 (1990) 235-240.

76 Balaji S, Sujatha S, Kumar S S & Srinivasan N, PALI-A database of Phylogeny and ALIgnment of homologous protein structures, *Nucleic Acids Res*, 29 (2001) 61-65.

77 Gowri V S, Pandit S B, Karthik P S, Srinivasan N & Balaji S, Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database, *Nucleic Acids Res*, 31 (2003) 486-478.

78 Sutcliffe M J, Haneef I, Carney D & Blundell T L, Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures, *Protein Eng*, 1 (1987) 377-384.

79 Blundell T L, Sibanda B L, Sternberg M J & Thornton J M**,** Knowledge-based prediction of protein structures and the design of novel molecules, *Nature (Lond)*, 326 (1987) 347-352.

80 Srinivasan N & Blundell T L, An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure, *Protein Eng*, 6 (1993) 501-512.

81 Jones T H & Thirup S, Using known substructures in protein model building and crystallography, *EMBO J*, 5 (1986) 819-822.

82 Levitt M, Accurate modelling of protein conformation by automatic segment matching, *J Mol Biol*, 226 (1992) 507-533.

83 Lindahl E & Eloffson A, Identification of related proteins on family, superfamily and fold level, *J Mol Biol*, 295 (2000) 613-625.

84 Sippl M J, The calculation of conformational ensembles from potentials of mean force: An approach to the knowledge based prediction of local structures in globular proteins, *J Mol Biol*, 213 (1990) 859-883.

85 Russell R B, Copley R R & Barton G J, Protein fold recognition by mapping predicted secondary structures, *J Mol Biol*, 259 (1996) 349-365.

86 Rost B, Schneider R & Sander C, Protein fold recognition by prediction based threading, *J Mol Biol*, 270 (1997) 470-480.

87 Jones D T, GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences, *J Mol Biol*, 287 (1999) 797-815.

88 Panchenko A, Marchler-Bauer A & Bryant S H, Threading with explicit models for evolutionary conservation of structure and sequence, *Proteins*, 3 (1999) 133-140.

89 Koretke K K, Russell R B & Lupas A N, Fold recognition without folds, *Protein Sci*, 11 (2002) 1575-1579.

90 McGuffin L J, Bryson K & Jones D T, What are the baselines for protein fold recognition? *Bioinformatics*, 17 (2001) 63-72.

91 Kelley L A, MacCallum R M & Sternberg M J, Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J Mol Biol*, 299 (2000) 499-520.

92 Jones D T, Taylor W R & Thornton J M, A new approach to protein fold recognition, *Nature (Lond)*, 358 (1992) 86-89.

93 Bernstein F C, Koetzle T F, Williams G J, Meyer E F Jr, Brice M D *et al*, The protein data bank: A computer-based

archival file for macro-molecular structures, *J Mol Biol*, 112 (1977) 535-542.

94 Berman H M, Bhat T N, Bourne P E, Feng Z, Gilliland G *et al*, The protein data bank and the challenge of structural genomics, *Nat Struct Biol, Suppl*, 7 (2000) 957-959.

95 Mallick P, Weiss R & Eisenberg D, The directional atomic solvation energy: An atom-based potential for the assignment of protein sequences to known folds, *Proc Natl Acad Sci USA*, 99 (2002) 16041-16046.

96 Holm L & Sander C, Dali/FSSP classification of three-dimensional protein folds, *Nucleic Acids Res*, 25 (1997) 231-234.

97 Sandhya S, Kishore S, Sowdhamini R & Srinivasan N, Effective detection of remote homologues by searching in sequence dataset of a protein domain fold, *FEBS Lett*, 552 (2003) 225-230.

98 Brenner S E & Levitt M, Expectations from structural genomics, *Protein Sci*, 9 (2000) 197-200.

99 Samudrala R & Levitt M, A comprehensive analysis of 40 blind protein structure predictions, *BMC Struct Biol*, 2 (2002) 3.

100 Fischer D & Eisenberg D, Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*, *Proc Natl Acad Sci USA*, 94 (1997) 11929-11934.

101 Mallick P, Goodwill K E, Fitz-Gibbon S, Miller J H & Eisenberg D, Selecting protein targets for structural genomics of *Pyrobaculum aerophilum*: Validating automated fold assignment methods by using binary hypothesis testing, *Proc Natl Acad Sci USA*, 14 (2000) 2450-2455.

102 Muller A, MacCallum R M & Sternberg M J, Structural characterization of the human proteome, *Genome Res*, 12 (2002) 1625-1641.

103 Schaffer A A, Wolf Y I, Ponting C P, Koonin E V, Aravind L *et al*, IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices, *Bioinformatics*, 12 (1999) 1000-1011.

104 Hosfield D J, Guan Y, Haas B J, Cunningham R P & Tainer J A, Structure of the DNA repair enzyme endonuclease Iv and its DNA complex: Double-nucleotide flipping at abasic sites and three-metal-ion catalysis, *Cell*, 98 (1999) 397-408.

105 Ramotar D, Popoff S C, Gralla E B & Demple B, Cellular role of yeast Apn1 apurinic endonuclease/3′-diesterase: Repair of oxidative and alkylation DNA damage and control of spontaneous mutation, *Mol Cell Biol*, 11 (1991) 4537-4544.

106 Demple B & Harrison L, Repair of oxidative damage to DNA: Enzymology and biology, *Annu Rev Biochem*, 63 (1994) 915-948.

107 Izard T & Blackwell N C, Crystal structures of the metal-dependent 2-dehydro-3-deoxy-galactarate aldolase suggest a novel reaction mechanism, *EMBO J*, 19 (2000) 3849-3856.

108 Hubbard B K, Koch M, Palmer D R, Babbitt P C & Gerlt J A**,** Evolution of enzymatic activities in the enolase superfamily: Characterization of the D-glucarate/galactarate catabolic pathway in *Escherichia coli*, *Biochemistry*, 37 (1998) 14369-14375.

109 Evans S V, SETOR: Hardware lighted three-dimensional solid model representations of macromolecules, *J Mol Graphics*, 11 (1993) 134-138.