PERGAMON Progress in Biophysics & Molecular Biology 73 (2000) 297–320

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Review

Methodologies for target selection in structural genomics

Michal Linial[a,*], Golan Yona[b]

[a] *Department of Biological Chemistry, Institute of Life Sciences, Hebrew University, Jerusalem 91904, Israel*
[b] *Department of Structural Biology, Fairchild Building, D-109, Stanford University, CA 94305, USA*

Abstract

As the number of complete genomes that have been sequenced keeps growing, unknown areas of the protein space are revealed and new horizons open up. Most of this information will be fully appreciated only when the structural information about the encoded proteins becomes available. The goal of structural genomics is to direct large-scale efforts of protein structure determination, so as to increase the impact of these efforts. This review focuses on current approaches in structural genomics aimed at selecting representative proteins as targets for structure determination. We will discuss the concept of representative structures/folds, the current methodologies for identifying those proteins, and computational techniques for identifying proteins which are expected to adopt new structural folds.

Keywords: Protein fold; Computational method; Clustering, Protein families; Databases; Protein structure determination; Fold recognition; Protein sequences

Contents

*Corresponding author. Tel.: +972-2-658-5425; fax: +972-2-658-6448.
E-mail address: michall@leonardo.ls.huji.ac.il (M. Linial).

0079-6107/00/$ - see front matter
PII: S0079-6107(00)00011-0

1. Structural genomics — the challenge

The genomic era is characterized by the rapid accumulation of DNA sequence data of organisms from all domains of life. Already 24 genomes have been fully sequenced and this number is expected to increase rapidly within the coming years (see http://www.genome.ad.jp/kegg/ for updated list). The immediate outcome of the large-scale sequencing projects is the primary sequence of all the proteins encoded in each of these genomes. The challenge of the post-genomics era is to convey this information to predict protein functions and the role of these proteins in a cellular context. Current knowledge fails in predicting function from sequence. However, successful prediction of a protein function from its primary sequence is possible once a pre-knowledge of a closely homolog protein is available (Martin et al., 1998; Park et al., 1998). Protein sequences may be considered as a one-dimensional representation that exposes the fingerprint of evolution. In many cases, only a faint trace of the ancestor protein is retained in the sequence. For example, two proteins with the same function may share only very low-sequence identity (i.e., 10%), while proteins that are about 20% identical in their primary sequence may still carry different functions (Holm and Sander, 1997b; Zarembinski et al., 1998). A fundamental bridge between a protein sequence and its biological function is in the protein structural information. Indeed, the three-dimensional (3D) structure of a protein is much more conserved during evolution (Finkelstein and Ptitsyn, 1987). Consequently, the shape of a protein in space is an informative representation of its biological role. Inspection of hundreds of proteins whose structures were solved shows that, in general, proteins that carry the same structure also share similar functions (Thornton et al., 1991). High-resolution 3D structures of many globular proteins show that only a small subset of critical amino acids associate with the core of the protein's biological function (Hasson et al., 1998; Kasuya

and Thornton 1999). Moreover, this small set of amino acids may only be conserved in space, rather than in the primary sequence. Examples of structures in which few well-positioned amino acids are sufficient to construct the key functional elements (i.e., binding pocket, surface for docking, active site for catalysis) are very common (Russell, 1998; Russell et al., 1998; Jones and Thornton, 1997). Taken all together, successful prediction of function in large scale will be strongly dependent on structural information. This notion initiated the concept of Structural-Genomics.

Solving protein structures during the last 20 years shed light on fundamental aspects of biological processes including enzyme catalysis, protein specificity, protein interaction and recognition, gene regulation and signal transduction. Consequently, biomedical research towards developing new vaccines and drug design became structurally driven. In recent years we have witnessed an accelerated rate of structural determination. According to PDB records, the number of solved structures increased more than fourfold since 1993. This is due to improved technologies in the process of experimental structure determination — from protein expression and purification to advanced algorithms for data analysis. However, in most instances, the motivation for solving new structures is driven by the researchers interest and by current perceptions of what considered biologically important rather than by global considerations of the protein space. Moreover, these perceptions do not necessarily conform with what nature deems important. The outcome is that the structural information of many recently solved structures is redundant. Out of all structures that were solved during 1994 till 1997, over 50% were variations of existing PDB entries, and another 30–35% were clear homologs of proteins already in the PDB database. Only 10–15% were genuinely new structures, of which only 3–5% revealed new folds (Brenner et al., 1997).

The challenge taken by the "Structural-Genomics" approach is to provide 3D structures in large scale. These proteins should preferably represent the protein diversity in the biological universe. From experimental, economical and all practical aspects, solving the structure of hundreds of thousands known proteins is not feasible (Sali, 1998). The challenge then is to define a relatively small set of proteins that once solved, their structure will enrich our current knowledge of the available structural templates for protein folds. These templates will be used for a large-scale structural prediction survey including all proteins that can be modeled reliably based on homology with these new templates using techniques such as comparative modeling and fold recognition (Bryant, 1996; Lemer et al., 1995; Mirny and Shakhnovich, 1998). Such procedures are quite successful in predicting the 3D structure of new proteins as was demonstrated in CASP3 (Koehl and Levitt, 1999; Sternberg et al., 1999). Structural genomics approaches aim to optimize the large-scale structural determination effort, and maximize the biological impact based on the currently available genetic information (McKusick, 1997; Terwilliger et al., 1998; Gaasterland, 1998b; Kim, 1998; Koonin et al., 1998).

2. Essential steps in the "Structural-Genomics" approach

Several important objectives were set in the frame of "Structural-Genomics" initiative. They are artificially divided to an initial step of "which sequences should be structurally determined?" and

"how to optimize the production in structural determination?". The steps may be presented as a list of more specific goals:

- Selecting preferable targets for structure determination for maximal biological impact.
  - Identifying all protein families as structural representatives.
  - Selecting protein targets by pre-determined criteria.
  - Computational approaches for target selection.
- Developing new and advanced technologies for structural determination.
  - Improved protein expression and purification systems.
  - Developing methods to increase the success rate of crystal formation.
- Throughput analysis and increase speed of data accumulation.
  - Automation and parallel processing at all experimental stages.
  - Improving NMR spectroscopy and X-ray crystallography data accumulation and data processing.
- Developing a mechanism for co-operation and co-ordination among research groups to optimize the worldwide effort in the large-scale structural determination projects.

Within the scope of "Structural Genomics" all tasks mentioned above should be addressed in parallel. Thus, classical "bottle-neck" in solving 3D structures will be minimized and a throughput analysis will lead to solving thousands of new structures within a few years (Montelione and Anderson, 1999). Once the technological tasks will be completed we anticipate that the newly solved proteins would be true representatives of the protein universe. Achievements in the field of "Structural-Genomics" will eventually pave the way for the forthcoming "Functional-Genomics" phase. At that stage, methods to extract maximal knowledge from structures and methods for deducing function from experimental structures and from models will have to be developed (Bork and Eisenberg, 1998; Bork and Koonin, 1998).

In this article, we will concentrate only on the first aspect of "How to select target proteins for structural determination", with, emphasis on the computational approaches that aim to provide a list of proteins for structural determination which will have maximal impact, once they are solved.

3. Computational approaches for target selection

Structural genomics is naturally involved with the question of how many folds are expected to exist in nature. Currently, several hundred folds are known (e.g., see the structural classification in SCOP (Hubbard et al., 1999) and in CATH (Orengo et al., 1997), and the current estimates place the total number of different folds (known and unknown) between several hundreds and few thousands (Gonnet et al., 1992; Chothia, 1992; Green et al., 1993; Wang, 1996; Zhang and DeLisi, 1998). In view of these estimates, structural genomics efforts to map the structural repertoire of the protein universe seem feasible. The core of structural genomics is to identify those target sequences which will serve as templates of the unexplored area of the protein structure space, once their structure is determined, i.e., the structure of each of these sequences is expected to reveal a new unknown fold.

The classical approach of structural genomics relies on identifying known sequences which have unknown structure and are not homologous to sequences of known 3D structure. The target selection process is a multi-stage procedure. Given the set of all known protein sequences, the first and obvious step would be to eliminate all those sequences that already have a known structure. The second step of identifying and eliminating all sequences that are homologous to sequences of known structure, is more complicated. Clearly, all sequences which are highly similar to sequences of known 3D structure can be eliminated, since they are expected to adopt the same known fold. However, by this we have eliminated only small portion of the sequences which are actually expected to have a known fold. To detect as many as possible of these sequences, we should have reliable and efficient tools to detect homology. This is important, since, by definition, homologous proteins have evolved from the same ancestor protein, and almost always they have the same fold (Pearson, 1996; Sander and Schneider, 1991; Hilbert et al., 1993).

Although the common evolutionary origin of two proteins is almost never directly observed, we can deduce homology, with a high statistical confidence, given that the sequence similarity is significant. In principle, similarity does not necessarily imply homology, and similarity should be used carefully in attempting to deduce homology. It is generally accepted that two sequences with over 30% identity along much of the sequences, are probably homologous and are likely to have the same 3D structure or fold (Sander and Schneider, 1991; Flores et al., 1993; Hilbert et al., 1993; Brenner et al., 1998). Moreover, today sequence comparison algorithms are accompanied with statistical estimates which provide a measure of statistical significance of the observed sequence similarities. These estimates can further help in assessing the significance of the similarity, and in many cases can lead to deduction of homology. The confidence in the deduction clearly depends on the level of statistical significance.

Nevertheless, one encounters many cases of high similarity in fold, that is not reflected in sequence similarity (Brenner et al., 1998; Murzin, 1993; Pearson, 1997). In many cases sequences have diverged to the extent that their common origin is untraceable by a direct sequence comparison. In such cases more sophisticated methods must be applied. In general, the most powerful tools are those which are based on the identification of inter-media sequences or incorporate information from a group of related sequences. This strategy have led to the development of advanced and sensitive search tools such as PSI–BLAST (Altschul et al., 1997) and SAM-T98 (Karplus et al., 1998), and the compilation of databases of protein families and domains. These databases have become an important tool in the analysis of newly discovered protein sequences. They usually offer a lot of biologically valuable information about domains and the domain structure of proteins, through multiple alignments and schematic, representations of proteins, and can help to detect weak relationships between remote homologs.

In that view, selecting representatives for structural determination must be strongly correlated with studies on protein families. Not only these studies can help to discern those sequences which are not homologous to sequences of known structure (and hence are less likely to have a known fold), but also it provides a framework for selecting specific preferable representatives. In a simplified way one may say that solving the structures of representatives of protein families provides the structural templates of the protein universe. This simple view will define all potential protein targets as the minimal set derived from the identification of protein families in the protein

space. Therefore, a critical requirement for the selection of targets for structural genomics is the comprehensive organization of protein sequences into families. Once this organization is obtained, for each family where none of its members were structurally determined a "typical" (seed) sequence can be selected as a potential candidate for structural determination (in practice, the final decision of the representative sequence is influenced by more practical considerations which are discussed in Section 4).

Note that a complete set of representatives for all protein families will be redundant in terms of new folds since different families within the same superfamily usually have the same fold. Moreover, different superfamilies which share little or no significant sequence similarity may still adopt the same fold (on the average, each fold is adopted by two to three protein families (Brenner et al., 1997)). Some folds are occupied by large number of superfamilies with low or undetected sequence similarity and no shared function (e.g., Ferredoxin fold and α/β barrel). These folds are known as superfolds (Brenner et al., 1997) and they may evolve by convergent to a "preferable" fold during evolution. Such redundancy should be eliminated if we concentrate on finding exclusively new folds. One of the goals of structural genomics is to identify those sequences which are expected, with high probability, to adopt a new fold.

3.1. Target selection based on sequence family organization

Since the early 1990s a considerable effort has been made to organize the protein sequence space into rational groups of related sequences (e.g. protein families, domain families). Several large-scale studies which considered all or many of the known protein sequences were performed and different approaches have been tested. These studies are mainly divided into two categories: those focused on finding significant motifs, patterns and domains within protein sequences, and those which apply to complete proteins. Some of these classifications and the corresponding research groups took part in the Structural-Genomics initiative, in Washington, DC this year (http://www.structuralgenomics.org). This section is a short survey of these efforts, which concentrates on the differences in methodologies used for producing extensive collections of protein families.

3.1.1. Pfam (Bateman et al., 1999)

Pfam is a database of hidden Markov models (HMMs) for protein families. For each family, the process starts from a seed alignment (either a published multiple alignment or an alignment from other databases such as ProSite (Hofmann et al., 1999)) of a non-redundant representative set of known members. Pfam alignments represent complete domains. The alignment is checked manually (to verify that the conserved features are correctly aligned, and the alignment has enough information content to distinguish chance similarities from true relationships), and a HMM is built from the seed alignment. The HMM is then used to scan SwissProt and TrEMBL, in search for all the other members of the family. If, a true member is missed then it is added to the seed alignment, and the process is repeated. Finally, a full alignment is constructed for the family by aligning all members to the HMM. This full alignment is checked again (manually) and if it is not correct, the alignment method is modified

or the whole process starts with a new improved seed. The resulting collection of families is named Pfam-A. This database is supplemented by Pfam-B which is based on automatically generated alignments of sequence clusters in SWISSPROT and TrEMBL that are not part of Pfam-A. The last release of Pfam-A (release 4.3), contains 1815 families. Additional 39,506 clusters are defined in Pfam-B.

3.1.2. ProDom (Corpet et al., 1999)

ProDom is the result of an analysis of the modular organization of all sequences in the SwissProt + TrEMBL databases. This analysis has led to the creation of a database of protein domains. The analysis starts by building a profile for all seed alignments of Pfam families, and search the sequence databases using PSI–BLAST. Domains that are detected as members of these families are extracted from the databases, and the remaining entries are clustered into domain families using PSI–BLAST. Each time a domain family is found, the corresponding sub-sequences are extracted from the databases, and the process repeats with the remaining sequences, until no more similarities are detected by PSI–BLAST. Finally, multiple alignment and a consensus sequence are generated for each domain family (for fast homology searches). The last release of ProDom (release 99.2) contains 157,167 domains, of which 43,965 appear in at least two sequences, and 1382 domain families are associated with Pfam families.

3.1.3. ProtoMap (Yona et al., 2000)

The ProtoMap database contains an automatically generated hierarchical classification of protein sequences, that is based on the analysis of pairwise sequence similarities of all sequences in the SwissProt database.

The analysis starts from a very conservative classification, based on transitive closure of highly significant similarities (with expectation value below $1e^{-100}$), that consists of many classes. Subsequently, classes are merged to account for less significant similarities. Merging is performed via a two-phase algorithm. First, the algorithm identifies groups of possibly related clusters using a statistical test, and if there is a strong statistical evidence for a connection between clusters they enter the same group. Clusters within the same group are considered as candidates for merging. Then, a "global" test is applied to identify nuclei of strong relationships within these groups of clusters. Some of the clusters are merged, given that their connection is statistically significant, whereas others stay apart.

This process takes place at varying thresholds of statistical significance (confidence levels), where at each step the algorithm is applied on the classes of the previous classification, to obtain the next one, at the more permissive threshold. The analysis starts at the $1e^{-100}$ threshold. Subsequent runs are carried out at levels $1e^{-95}, 1e^{-90}, 1e^{-85}, \ldots, 1e^{-0}(= 1)$. Consequently, a hierarchical organization of all proteins is obtained. ProtoMap release 2.0 contains 13,354 clusters, of which 5869 contain at least two sequences.

3.1.4. COGs (Tatusov et al., 1997)

Comparison of proteins from eight complete genomes (six phylogenetic lineages), and a single-linkage clustering algorithm resulted in 864 clusters of orthologous groups (COGs). Each COG consists of orthologous proteins (genes in different species that evolved from a common ancestral

gene) or orthologous sets of paralogs (genes from the same genome, which are related by duplication) from at least three species, which typically have the same function. The COGs are created by starting from triangles of orthologous proteins from different species (the minimal COG), and then merging triangles which share a side. Consequently, the final COGs may contain paralogs as well. An additional step is carried out to split COGs which were incorrectly merged due to the existence of multi-domain proteins. Finally, COGs are merged to form superfamilies, using PSI–BLAST (Altschul et al., 1997) with the criteria that at least two proteins from the first COG hit members of the second COG.

3.1.5. PIR (Barker et al., 1996)

This study classifies all proteins in the PIR database (George et al., 1996) into families based on global similarities, and into homology domains based on local similarities. In this study sequences are classified into families and superfamilies based on similar overall architecture (same domains in the same order and over 50% sequence identity if the sequences belong to the same family; more flexibility is allowed if the sequences belong to different families within the same superfamily). Homology domains are defined using multiple alignments of homologous segments (identified based on local similarities). Both classifications depend on semi-automatic procedures and manual inspection. The last release of the resulting PROT–FAM database (associated with release 58.0 of PIR) contains 12,155 families and 365 homology domains.

3.1.6. Picasso (Holm, 1999)

This classification is derived from a non-redundant sequence database at 90% sequence identity level. Families are defined based on BLAST pairwise similarities. A Hierarchical procedure is applied to cluster families. By increasing BLAST *e*-values and considering weaker similarities, neighboring families are identified, and merged based on profile–profile comparison. The procedure resembles a single-linkage clustering algorithm with a threshold. The resulting families are closed, mutually disjoint sets of sequence domains. Most families can be represented by a single multiple alignment that contains all member sequences. Each multiple alignment is compiled around a seed sequence.

3.1.7. ProClass (Wu et al., 1999)

The ProClass database is a non-redundant protein database that classifies sequences into families based on ProSite patterns and PIR superfamilies. The current ProClass release (release 5.0) classifies 91,785 sequences, of which 58,925 are compiled into 1021 PROSITE-based families, 29,128 are compiled into 5150 PIR-based families (without PROSITE representatives). Another 3732 entries are classified, using a combination of several search and alignment tools, into motif families that are not detected by PIR and PROSITE.

3.1.8. Systers (Krause and Vingron, 1998)

The Systers database is based on an iterative method for database searching to cluster proteins in the SwissProt database and in the PIR database. For each seed protein, all blastp hits in a database search with *p*-value of 10^{-30} at the most are retained and the lowest scoring sequence is used as a query for the next search and clusters are extended accordingly. The process repeats

until no new sequences above the cut-off are found, or if the search has no sequence in common with the set of accepted hits from the first search. The current release of Systers (release 2) contains 12,659 non-singleton clusters.

3.2. Different collections of protein families — different targets

In general, the above studies can be divided into protein-based studies (COGs, ProtoMap, Systers) and domain-based studies (Pfam, ProDom, Picasso). PIR combines both. Additional extensive classifications that provide motif-based databases are ProSite dictionary, PRINTS, Blocks, DOMO, SMART and IDENTIFY.

The class of protein-based studies draw directly on pairwise comparison, and differ mainly in the details of the clustering procedure. The motif-based studies differ from each other in several aspects. Some are based on manual or semi-manual procedures (e.g. ProSite, PRINTS), others are generated semi-automatically (Pfam) and the rest are generated fully automatically (e.g. ProDom, Blocks, Domo). Some focus on short motifs (ProSite, PRINTS, Blocks, IDENTIFY) while others seek whole domains and try to define domain boundaries (Pfam, ProDom, Domo). Most databases also give the domain/motif structure of proteins. The methods used to represent motifs and domains vary, and among the common forms are the consensus patterns and the regular expressions (ProSite, PRINTS, IDENTIFY), the position-specific scoring matrices or profiles (Blocks) and the HMMs (Pfam). These forms differ in their mathematical complexity, as well as in their sensitivity/selectivity.

In the frame of the protein structure initiative (PSI) all seven groups provided their own classification for protein families. The collection can be accessed at the PSI web site (http://www.structuralgenomics.org). It is very hard to directly compare these family collections. Despite high overlap, there are marked differences among these family collections. This is due to the different approaches and methodologies that were employed, the differences in the databases that were analyzed, and different levels of automation. Consequently, the target lists provided by these groups were different. However, an overall view of these lists expose some clear trends, and a rough measure of the number of target sequences that need to be solved can be extracted. This number depends on the desired quality of models one wish to build for other proteins based on homology with these template structures. The confidence in the quality of the model clearly increases as the sequence identity between the modeled sequence and the template sequence increases. Obviously, at higher confidence levels, more family representatives need to be solved (see Fig. 1). The graph (kindly provided by D. Vitkup) can be used to derive the minimal number of family representatives as targets for structural determination at any desired quality. The broad band covers the range presented by all seven contributors. At a 35% level of identity between the modeled sequence and the template sequence only one out of 10–16 proteins may be solved. At this level, it is expected that all other neighbors will be still correctly modeled (with approximately 85% of their C-alpha atoms modeled within 3.5 Å of their correct position). Considering over 300,000 proteins in the non-redundant databases (e.g., SW — 80,000, TrEMBL — 200,000), this suggests solving the structure of about 18,000 proteins. This number of protein targets in not beyond reach in the next few years. Yet, additional considerations are required to rank these target proteins for the experimentalists. This issue is discussed in the following sections.

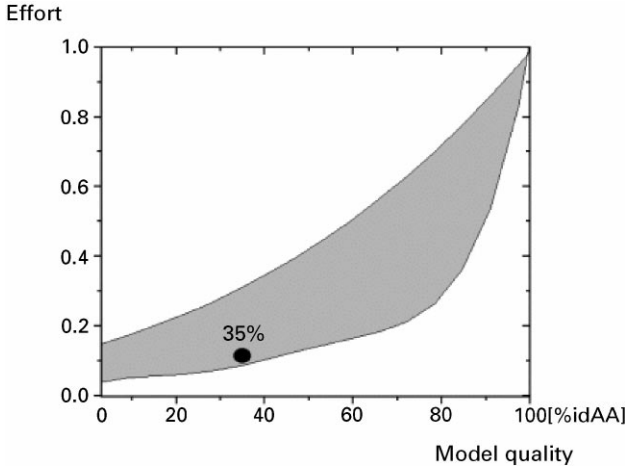

Fig. 1. How many target sequences should be solved to cover the protein universe? As the model quality increases the effort increases as well (i.e. the number of structures that need to be solved approaches the number of proteins with unknown structure). The figure is taken from the NIH report of the PSI meeting.

3.3. Pilot projects — selecting a subset from all protein families as targets

Structural genomics efforts which are based on protein family classifications aim to explore the entire protein universe. However, studies of protein classification cannot be immediately transformed to an experimental target list. Other considerations need to be addressed, such as ease of purification, (expected) stability, biological implication and importance and more (see Section 4). Moreover, the number of protein families which have unknown structure is still large, and few years of research will be required to define the structures of all family representatives. Therefore, in order to direct the structural-genomics efforts, additional criteria should be applied to provide a relatively small but oriented list of targets. General criteria in prioritizing families for structure determination include the family size, the expected functional insight (i.e. proteins for which no functional data is available are preferred), and the taxanomic diversity. For example, following the last criterion, the targets in the COGs database were selected as the most conserved proteins according to their appearance in at least three independent genomes. The rational behind this selection is that structural folds of proteins that are conserved throughout evolution may represent stable, fundamental and "successful" folds, of great biological importance. Indeed, many homologous proteins that are found in eukaryotes, prokaryotes and arche associate with fundamental cellular functions (Tatusov et al., 1997; Wolf et al., 1999). A similar approach is taken by Paul Bash (Northwestern University). Targets are selected from proteins with a wide phylogenetic distribution, to represent universal proteins. Other general considerations that are usually taken for narrowing down the potential targets to a reasonable number are:

3.3.1. Selecting target proteins driven by human-health aspects

The rational of selecting medical oriented proteins for structural determination is evident. Solving the structure of human proteins will benefit not only treatment of diseases but also our

ability to understand a large number of normal and pathogenic biological processes. The surprising homology between many of the yeast, *Drosophila* and *C. elegans* proteins to human proteins allows flexibility in choosing a human-related homolog protein as a target. Human health is tangled with other forms of life. Consequently, selection of proteins as structural representatives must include bacterial, pathogens, parasites and other infectious and virulent agents that affect the human body. Such a project is carried by Guy Montelione (Rutgers University). Targets are selected on the basis of relevance to human diseases, broad conserved metazoan genes, and genes of human pathogens.

3.3.2. Experimental and biophysical considerations

A different rational for target selection is based on experimental and biophysical arguments. In this line, the selection of small number of target proteins is based on the potential success of these proteins to be solved. Thermostable proteins, proteins predicted to be globular and those with a single, autonomous domain are all considered as candidates for NMR or X-ray structural determination.

3.3.3. Determining targets from specific model organism

An attractive approach for limiting the number of potential targets is to concentrate on a specific model system. Clearly, the complete set of genes of a single organism may not represent the whole protein universe (especially for lower organisms which are usually considered for this type of study). However, it provides a fairly good sampling of this space. Moreover, knowing the complete repertoire of protein structures in an organism may help in the understanding and modeling of cellular networks. Along this line, several structural genomics pilot projects were initiated. All these projects aim to proceed from the stage of "target selection" to the last phase of high-resolution structural determination. Among these projects are:

- Aerobic bacterium *Pyrobaculum aerophilum*, D. Eisenberg et al. (UCLA) — This project focuses on assigning folds to medically relevant proteins. Medical relevance is identified by homology with Human Mendelian Inheritance database.
- Hyperthermophilic archaeon *Methanococcus jannaschii*, S. H. Kim (UC Berkeley). The advantages of this organism is its position in the phylogenetic tree of life. Moreover, from experimental aspects, proteins of hyperthermophiles are relatively easy to purify and crystallize.
- *Saccharomyces cerevisiae*, B. Studier (Brookhaven National Laboratories) and BNL/Rockefeller. Yeast was chosen because of the high proportion of human-related proteins and the advantage of having large community of scientists working with this model organism.
- *Haemophilus influenzae*, John Moult (CARB, University of Maryland) and CARB/TIGR project. Targets are selected from unannotated open reading frames. The model organism represents the smallest free-living genome. Information about the function of "hypothetical" proteins may lead to discovery of novel biochemical processes in a simple cell.

Target selection by any of the above-mentioned considerations do not provide an unbiased sampling of the protein universe. These considerations are used only to help bridge between the thousands of potential targets that are provided by any of the family organization approaches and the limitations of experimentalists to solve that many protein structures.

3.4. Exhaustive target selection — fishing for new folds

The methods described in the previous section rely solely on sequence analysis. No structural considerations are used in the organization of protein sequences into domain families and protein families, nor in the process of target selection, aside of screening those families whose structure has been solved. Along the same lines, Elofsson and Sonnhammer (1999) studied the correspondence of the structural classes of SCOP and the sequence classes of Pfam to compile a list of target families which do not occur in SCOP (most of which are likely to be transmembrane proteins).

Incorporating structural information in the process is extremely important because structure is often conserved more than sequence (Levitt and Gerstein, 1998). The two approaches described next rely on the available structural information to predict which sequences are likely to have a new fold. The discovery of a novel fold may well contribute to the understanding of functional details of entire protein families, thus, a scheme for discovering those currently missing folds is desirable (Holm and Sander, 1997a; Murzin, 1996). The two approaches discussed below systematically analyze the protein space in an attempt to accelerate the pace of discovering new folds. The first method relies on a map which was constructed based solely on sequence properties and then incorporates structural information in the statistical analysis of the properties of this map (Portugaly and Linial, 2000). The second approach introduces structural considerations in the process of building the map (Yona and Levitt, 1999).

3.4.1. A statistical–computational approach for target selection (Portugaly and Linial, 2000)

This study is based on a statistical analysis of a map of the protein space as provided by ProtoMap. Contrary to the approaches discussed in Section 3.3 the underling principle is that no restriction or pre-determined criteria are set in the target selection procedure.

ProtoMap (Yona et al., 1999) organizes the protein space in a graph such that proteins that are located close in the graph are biologically related. Exploration of this protein map can reveal hidden biological information. In this study, it is used to extract structural information and predict which proteins have new, currently unknown structural fold. A detailed description of this procedure is given in Portugaly and Linial (2000).

The procedure is composed of four main steps:

1. Positioning each of the domains with a solved three-dimensional (3D) structure (from the SCOP database) into the map of the protein sequence space. The map of clusters is provided by ProtoMap, and each domain is mapped to a single cluster.
2. For each cluster, determine a ''representative fold'' based on the folds associated with all structural domains in that cluster.
3. Distances within the ProtoMap graph are computed from each representative fold to the neighboring folds. The distributions of these distances are used to create a statistical model for distances among those folds that are known and those that are yet to be discovered.
4. Statistical estimation is derived for the probability that any protein has a new, yet undetermined fold.

Proteins that score the highest probability to represent a new fold constitute the list of preferred target proteins for structural determination.

3.4.1.1. The ProtoMap graph. This computational approach builds on the representation of the protein sequence space as provided by ProtoMap. The underlying hypothesis is that distances in this graph can reflect hidden structural information. The semi-global properties of the ProtoMap graph leads to the definition of *Local maps* (Yona et al., 1999) which are used as "road maps" to tour the protein space. Each cluster (of sequences) in ProtoMap has a weighted list of related clusters (either directly or indirectly) that form connected components of varying sizes and connectivities. In such a representation, weights reflect relatedness among clusters. To illustrate the notion of ProtoMap graph, consider for example the family of small GTP binding proteins. Proteins of the Ras superfamily are all classified to cluster 9 (in ProtoMap release 2.0, level $1e^{-0}$). This cluster contains over 300 proteins that belong to the subfamilies of Ras, Ran, Rho, Rap, Rac and Rab. Among its related clusters in ProtoMap one can find the family of the alpha subunit of heterotrimeric G-proteins, ARF and Sar (cluster 31,177 proteins) and other GTP-binding proteins. A 2D representation of this cluster and its related clusters reveals a complex map of relations (Fig. 2). A detailed evaluation of this scheme suggests that evolutionary and biological features are captured in the graph, and specific cellular functions can be assigned to sub-territories in this local map of the Ras superfamily. For more details about functional local maps see Yona et al. (1999).

The ProtoMap graph can be considered at different levels of granularity. A key feature of this graph is that a numerical value is assigned to pairs of related clusters to quantify their degree of proximity. Thus, the graph of related clusters (as in Fig. 2) can be "clipped" at different thresholds, by eliminating all edges of significance below a given threshold. Each threshold yields a different scheme and thus, the protein universe is partitioned to connected components of different sizes and graph connectivities.

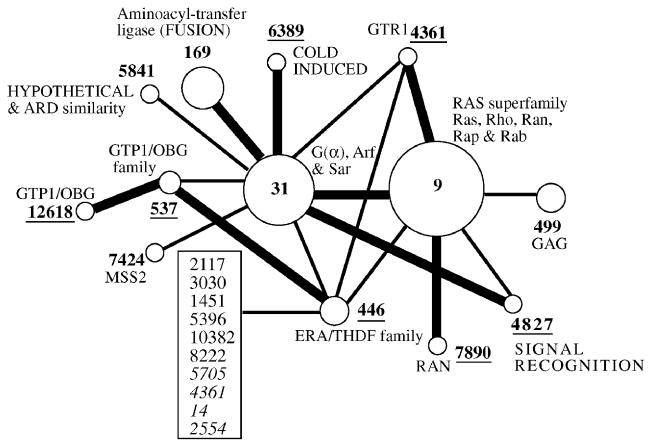

Fig. 2. A local map of the Ras superfamily. Each circle stands for a cluster. Circles' radii are proportionate to the clusters' sizes. Cluster 9 contains members of the Ras superfamily. The cluster is related to other clusters of small GTP-binding proteins. Edges represent relatedness between clusters and edge widths are proportionate to the quality of connection between the corresponding clusters. Edges with quality > 0.1 are marked with a thick line and edges with $0.01 <$ quality < 0.1 are marked with a thin line. Cluster numbers that are underlined indicate no additional neighbors. Some of the clusters (169 and 499) are connected to additional local maps that are not related to the GTP binding proteins (not shown).

3.4.1.2. Assigning representative folds to ProtoMap clusters. To embed the structural information within the ProtoMap graph, the ProtoMap classification is matched with the SCOP hierarchical organization of protein structures. ProtoMap (version 2.0) contains a classification of 72,623 proteins in the SwissProt (SP) database (Bairoch and Apweiler, 1999). At the coarsest level of granularity (level $1e^{-0}$) it consists of 13,354 clusters, of which 5869 contain at least two proteins. 1403 clusters have size 10 and above. SCOP is a hierarchical classification of all known protein structural domains (Hubbard et al., 1999). SCOP (release 1.37) comprises 11,748 solved structures giving 2,264 structural domains. These domains are derived by parsing PDB structures into their structural domains and grouping redundant entries within PDB. The 2264 domains are classified to 834 families, 593 super-families, 427 folds and 8 classes. Two additional classes — "*designed proteins*" and "*non-protein*" are not considered in this study.

To position the known structures in the ProtoMap graph, the information from the PDB database is matched with that of the SwissProt database. The correspondence among structural domains and SP-chains is bidirectional. An SP-chain is defined as occupied if it is mapped to at least one domain, and is vacant otherwise. Of the 72,623 SP-chains, 1688 are occupied. A cluster is defined as occupied if it contains at least one occupied SP-chain, and vacant otherwise. Of the 13,354 clusters in ProtoMap 756 are occupied. While 59% of the occupied clusters contain only one occupied SP-chain, more than 73% of the occupied SP-chains are in clusters with two or more occupied SP-chains. An occupied cluster is mapped to a specific fold if it contains an SP-chain that is mapped to that fold.

It is clearly desirable to assign a single representative fold to each ProtoMap cluster. However, apriori it is not clear that such a selection can be carried out. Many proteins are multi-domain, and an SP-chain may correspond to several domains, which usually have distinct folds (of the 1688 occupied SP-chains, 21% contain more than one domain). In practice, for each occupied cluster the best representative fold is defined as the most abundant one in the cluster. Of the 411 folds in SCOP 1.37 that are mapped to occupied SP-chains, 329 folds were chosen as cluster's representatives. Remarkably, only 80 occupied SP-chains are not mapped to the representative fold of their cluster (for more details see Portugaly and Linial, 2000). This matching suggests that ProtoMap is selective for SCOP folds. That is, a cluster gathers proteins of the same fold, though not necessarily all proteins of that fold.

3.4.1.3. Predicting a protein's probability to have a new fold. The statistical estimates are based on an analysis of distances in the ProtoMap graph. A threshold is set and the graph is "clipped" by eliminating edges which are less significant than the threshold. Two distinct probability distributions are defined: (i) Distances from new clusters to occupied clusters. (ii) Distances from old clusters to occupied clusters. Old and new here means clusters with an already solved structural fold and those without one, respectively. This step is repeated for various thresholds to extract maximum information from the graph.

Specifically, for each cluster in the clipped graph the *maximal vacant volume V* is measured. This is the number of non-occupied clusters in the maximal non-occupied sphere centered around the cluster (i.e. the maximal sphere that does not include occupied clusters). If there are no occupied clusters in the connected component then the maximal vacant volume is defined as *empty*. The measure reflects distances in the graph. The underlying assumption is that the vacant volume is strongly related to the probability that this cluster is associated with a new fold.

Intuitively, one may expect that cluster whose maximal vacant volume is small will be represented by a known fold, because of its proximity to known structures, whereas a cluster whose maximal vacant volume is large, will adopt a new fold. To quantify this property, all paths from occupied clusters (i.e. with a representative fold) to the clusters surrounding them in the graph are inspected. Reaching a cluster with the same representative fold as the cluster from which the tour originates simulates a tour between "old" folds, whereas reaching a different fold is considered as reaching a new fold. This process is repeated for all occupied clusters (756 clusters). Based on the information collected from touring the ProtoMap graph the distributions of vacant volumes are calculated to derive two conditional probability distributions $P_{\text{old}}(V)$ and $P_{\text{new}}(V)$. Many tours turnout to be non-informative as no other fold was found within that connected component in the graph. Such tours are aborted, and the corresponding maximal vacant volume is defined as empty.

Given a cluster with a specific vacant volume V, what is the probability $P(\text{new}/V)$ that the corresponding cluster will have a new fold? According to Bayes rule

$$P(\text{new}/V) = \frac{P(V/\text{new})P(\text{new})}{P(V)} = \frac{P_{\text{new}}(V)P(\text{new})}{P(V)}.$$

The prior probability of a new fold, $P(\text{new})$, is calculated based on the number of known folds (427, according to SCOP 1.37) and on the estimation of the total number of folds, for which a rather conservative estimate of 1000 is taken (Chothia, 1992). That is $P(\text{new}) = 1 - \frac{427}{1000} = 0.573$. The other term in that equation, $P(V)$, is given by the weighted sum over the two empirical probability distributions, i.e.

$$P(V) = P(\text{new})P_{\text{new}}(V) + (1 - P(\text{new}))P_{\text{old}}(V).$$

The above analysis places all clusters in five classes according to their vacant volume (the classes are defined so as to maximize the separation between the two empirical distributions). Fig. 3 shows the values calculated for the probability that a cluster will have a new fold for various

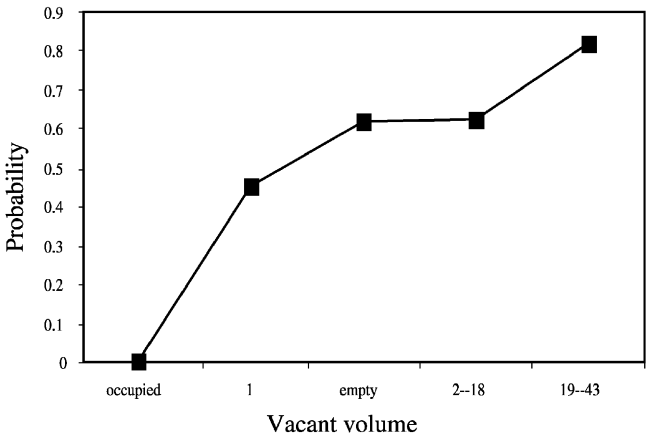

Fig. 3. Probability of a new fold for various vacant volumes. As the volume increases, the probability that a cluster with that vacant volume now increases as well. The probability was calculated from the ProtoMap graph, after all edges of quality <0.1 were eliminated (see text for details).

vacant volumes. As seen, the probability function increases monotonically with the volume. This supports the initial hypothesis that distances in the ProtoMap graph reflect structural relatedness. Unfortunately, for the many clusters to which no vacant volume can be assigned (denoted empty) this analysis provides little information. For these clusters, the probability values are slightly above the a priori value (0.573).

3.4.1.4. Evaluation of the predicted probability to have a new fold. To evaluate this prediction two test cases were considered. The first one is membranous proteins. So far, the structures of very few membranous protein have been solved (mostly classified in SCOP class 6). Therefore, clusters of membranous proteins are expected to have a high probability for being new. For this test over 1000 clusters (representing about 20% of the SP-chains) with proteins having multiple membrane spanning regions were considered. The occurrence of these membranous clusters in the top probability classes is 6.5 fold higher than the overall occurrence of clusters in that classes. This test confirms that the probability function indeed assigns higher probabilities to membranous clusters as hypothesized.

A more stringent evaluation is based on recent structural data that was not available during the statistical analysis. While the original analysis was performed using SCOP 1.37 (about 13,000 domains), the re-evaluation was performed against SCOP 1.39 (about 18,000 domains). Mapping structural domains to SP-chains using the records of SCOP 1.39, assigned new structures to 388 clusters of which 48 are new folds. Given the vacant volume of these clusters, it is possible to test how well the predictions match the new assignments. Since clusters with high vacant volume have high probability to adopt a new fold, we expect that most of these clusters are represented by new folds, i.e., the proportion of new clusters out of all clusters with the same vacant volume would increase as the vacant volume increases. The results are summarized in Fig. 4 and a strong correlation between the predicted probability of being new and the proportion of new folds among the recently released structures is found. Hence, the evaluation tests suggest that selecting targets from the top probability list will lead to accelerated pace of fold discovery.

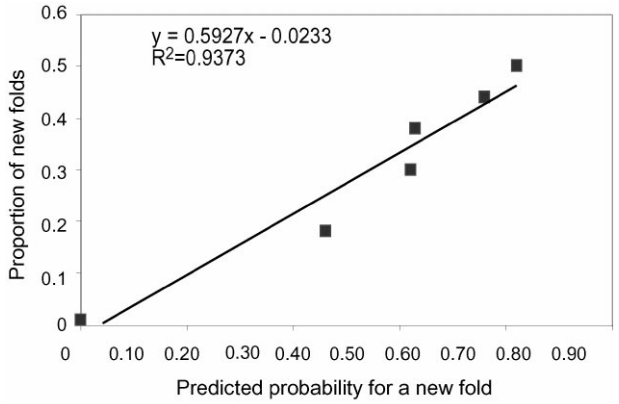

Fig. 4. Correlation of new folds with the predicted probability. Based on all clusters that were assigned new structures from SCOP 1.39, and the proportion of new folds of these clusters. *X*-axis — predicted probability to be new assigned to bin. *Y*-axis — proportion of new folds out of new structures that were assigned to bin.

3.4.1.5. The target list. The list of targets at the top probability score contains 713 clusters (5.3% of all clusters), which accounts for 8.2% of the SP-chains. Following subtraction of clusters of membrane proteins and considering clusters that have more than five SP-chains in each, the number or clusters in this target list is reduced to 125. Following the updated target list against the PDB (dated up to July 1999) a list of 101 clusters is suggested to the structural community. Preliminary results from this study were presented at the National Institute of General Medical Studies meeting (http://www.structuralgenomics.org). The complete list of target proteins can be accessed at (http://www.cs.huji.ac.il/~elonp/Targets). A sample list is given in Table 1.

A Structural Genomics project at Argonne National laboratory (supervised by Andrzej Joachimiak) was initiated for determining as many new folds as possible for X-ray crystallography (Shapiro and Lima, 1998). The 101 clusters that scored with top probability to have new fold are under study. For each cluster the best protein was selected following the considerations that are discussed in Section 4. Currently, already 80 proteins are at different stages of cloning, expression and purification.

Table 1
Selected list of target clusters

Cluster	P(new)[a]	Size	No. of species[b]	H	P	A	E	V[c]	ProSite[d] families (No. of sp-chains with this ProSite)
15	0.76	280	255	+	+	+	+	−	CYTOCHROME_B_HEME (256) CYTOCHROME_B_QO (254)
54	0.76	117	114	+	+	−	+	−	ATPASE_A (117)
82	0.76	98	100	+	+	−	+	−	COMPLEX1_ND1_1 (98) COMPLEX1_ND1_2 (98)
143	0.76	70	76	+	+	+	+	−	RIBOSOMAL_S4 (70)
146	0.82	69	74	+	+	−	+	−	
153	0.82	68	68	+	−	−	+	−	
180	0.76	58	55	+	+	+	+	−	RIBOSOMAL_S3 (58)
201	0.76	53	8	−	+	−	−	−	PROKAR_LIPOPROTEIN (2)
284	0.76	40	12	+	−	+	+	−	MCM_2 (40) MCM_1 (40)
292	0.82	39	9	+	−	−	+	−	UCH_2_2 (38) UCH_2_1 (38)
327	0.76	37	25	+	+	−	+	−	
354	0.82	35	15	−	+	−	−	−	HTH_LYSR_FAMILY (35)
390	0.76	32	20	−	+	−	−	−	HTH_MERR_FAMILY (26)
396	0.76	32	7	−	−	−	−	+	
406	0.76	31	31	−	+	−	−	−	DNAA (31)
411	0.76	31	30	−	+	−	+	−	SECA (13)
447	0.76	29	17	−	−	−	−	+	
480	0.76	27	20	−	+	−	+	−	
492	0.76	26	21	−	+	−	+	−	ISOCITRATE_LYASE (20) MICROBODIES_CTER (12)
516	0.76	25	20	−	+	+	+	−	ARGC (16)

[a] The actual probabilities are based on more elaborated arguments described in Portugaly and Linial (2000).

[b] As annotated by SwissProt. Some proteins may belong to more than one species.

[c] Cluster contains at least one protein from phyla. H, P, A, E, V — Human, Prokaryote, Archaebacteria, Eukaryote,Virus, respectively.

[d] See ProSite database (Hofmann et al., 1999).

3.4.2. The unified-map approach (BioSpace (Yona and Levitt, 2000))

The approach taken by Yona and Levitt (2000) focuses on establishing a consistent and unified framework for sequence and structure analysis. The protein space is mapped through a scheme that combines sequence-based metrics with structure-based metrics and considers domains as well as entire protein chains. By combining sequence information and structure information (when available) it is believed that a better and more accurate map can be created, insights about the geometry of the protein space that are missed by sequence analysis alone will emerge.

Obviously, structural information is available for only a small portion of the protein space. The process starts by defining classes that are "centered" around proteins with known structure. These classes enable large-scale modeling of protein sequences based on their similarity with the "center" sequences. Concurrently, a framework for unifying the structure-based metrics with the sequence-based metrics is established, since both are computable for these classes. Using these metrics the structure-based classes are compiled into a hierarchical organization, which resembles the organization of proteins into families, super-families, folds and fold classes. Once the structure-based clusters are defined, the yet unanalyzed areas of the protein space are subjected to sequence-analysis. This analysis provides sequence-based classes much in the same way as the structure-based classes. However, in the absence of structural information only sequence-based metrics are used. The classes of both types are then combined in a single map using higher-level measures of sequence similarity between classes, based on their correlation with structural similarities for structure-based classes.

Clusters which are located close in this unified map are expected to have similar folds. Therefore, only classes which are far from clusters of known folds are consider as possible targets for 3D structure determination.

3.4.2.1. Defining the first part of the map — the structure-based classes. Given a structural class (as in SCOP or CATH (Orengo et al., 1997) classifications), one can build a model based on either a multiple alignment, or iterative alignment procedure such as PSI–BLAST or SAM-T98. The resulting model is expected to be more accurate than sequence-based models since some of the members in structural classes have no apparent sequence similarity. Sometimes it is beneficial to split this structural class into subclasses so that the quality of the model is improved.

The analysis starts from the sequences of the SCOP domains. The domains in this database provide a natural definition of the basic building blocks of protein structures. Each of which is a well-defined part of a protein structure that can be assigned a structural or a functional role. The main use of this database is not the classification it provides, but the actual definition of these structural units. The sequences of these domains are clustered based on their sequence similarity, which is independent of the SCOP classification. These clusters are then used to identify as many as possible homologs in the protein space. Each sequence-based family is represented by a profile "centered" around a seed sequence, and this profile is used to search a non-redundant (NR) database composed of all major sequence databases, using PSI–BLAST, to detect a whole class of related protein sequences. Issues such as sensitivity vs. selectivity are addressed by applying different validity indices.

Each of the detected sequences is marked and a profile is built for the whole class of similar sequences. A 3D model is then built (using SegMod (Levitt, 1992)) for each sequence in the class

based on its (profile) alignment with the seed sequence whose structure is known. Classes based on SCOP domains are called type-I classes. In all, 1421 type-I classes are found to contain a total of 168,431 sequences (44.5% of the NR database) at significance level of $1e^{-5}$. The largest cluster contains 17,596 sequences, and 37 clusters contain more than 1000 sequences each. The results of this analysis as well as the models are available on the web, at http://biospace.stanford.edu.

3.4.2.2. Defining the second part of the map — the sequence-based classes. From the perspective of structural genomics, the structure-based clusters are not of immediate interest but they establish a framework for the identification and unification of structurally unexplored regions of the protein space. The analysis starts by removing all sequences that are homologs of proteins with known structure (all members of type-I clusters). The remaining sequences can be analyzed using any of the methods described above for sequence family organization. Specifically, the PSI–BLAST procedure is applied iteratively to sequences left over from the previous iteration, starting each time from a random query. The resulting clusters may overlap as the search is performed against the whole non-redundant database (including members of type-I clusters). These clusters are named type-II classes.

3.4.2.3. Combining the two parts. As discussed above, sequence families that have no structural representatives do not necessarily imply new folds. We expect that with some more sensitive measure of similarity we would be able to detect this relatedness. Some of this information can be revealed by combining the information from a group of related sequences (i.e. belonging to the same family/class/fold) using either classical methods such as multiple alignments, and the resulting profile representations, or more advanced statistical models such as HMM (Krogh et al., 1996) and PST (Bejerano and Yona, 1999). By comparing two such models which were obtained for two related families it may be possible to detect the relatedness (Park et al., 1998). Specifically, the profile representations obtained for type-I/-II clusters are used to provide a sensitive, powerful means of comparison between clusters. The acquired information about the sequence similarity between and within clusters, as well as their structural similarity (calculated using StructAl (Gerstein and Levitt, 1998; Levitt and Gerstein, 1998)) enables the development of a framework for unification of these two metrics, and the profile similarity scores are calibrated to the scores of the structural similarities available for type-I clusters.

Following this rational, type-I and II clusters are grouped using higher level measures of similarity. Those clusters with significant overlap in membership, are marked first. The clusters are then compared using either a structure metric (when known 3D structures are available) or sequence profile metric (when no structure available), and clustered into superfamilies and fold families.

The probability for having a new fold corresponds to distances in this map, and a list of target sequences which represent clusters that are far from known structures is reported. Since this organization is hierarchical, clusters that are mapped to the same vicinity are ranked and only one member of each super-class is chosen. At lower priority other cluster centroids within the same super-class can be selected. The full list of targets will be available at http://biospace.stanford.edu.

4. From theory to practice

One of the goals of selecting target proteins in the frame of structural genomics is to accelerate the pace of determining protein structures that will have maximal impact on bio-medical science. Following the PSI meeting, an interactive website was created (http://www.structuralgenomics.org/) that presents all current information on target selection. Another useful resource is the PRESAGE database (http://presage.stanford.edu) that was established a year ago at Stanford University (Brenner et al., 1999). This website aims to help structural biologists to select their preferred targets by providing the status of proteins that are currently being studied by other groups.

The question remains how to proceed from a target list to the laboratory bench? The current hurdle and rate determining step in structural biology is the preparation of the biological samples. All the following steps in structural determination depend on the availability of well-diffracting crystals and on well-behaved NMR samples. A fundamental decision that helps bypass most difficulties in expression and purification of proteins is to use thermophilic genomes as the biological source. Several of these genomes are already fully sequenced (*Pyrobaculum aerophilum*, *Methanococcus jannaschii*) and others will be completed soon (e.g., *Sulfolobus solfataricus*, *Methanobacterium thermoautotrophicum* and some *Pyrococcus genomes*). The success in solving structures from thermophilic genomes has been demonstrated by Kim et al. (1998) and Lim et al. (1997). An additional consideration is to choose proteins with a sequence composition that indicates a high rate of success in latter stages of data collecting. For example, proteins with high occurrence of methionine may be selected as recombinant selenomethionine proteins can be analyzed by multiwavelength anomalous diffraction (MAD) phasing (Ogata, 1998; Hendrickson et al., 1990). The next steps following expression and purification involve biochemical and biophysical criteria. Applying partial proteolysis, dynamic light scattering measurements and mass spectrometry help to determine the protein fragments that are most stable. Another rational decision is to favor solving the structure of globular proteins. At present, crystallization of membrane proteins (and a number of other classes) is very difficult and currently cannot be approached in high-throughput projects. In general, small proteins without transmembrane segments and without low-complexity segments (that may suggest the existence of non-globular domains or may correspond to flexible linkers between domains) tend to crystallize readily. Selecting a specific protein for structural determination from all other proteins within the cluster is based on a large set of criteria. Such decisions may be biased towards a specific organism, a preferable expression system, a suitable codon usage for protein expression and above all, a biological and a medical interest in that candidate protein.

5. Discussion and future directions

Structural genomics aims to direct large-scale structural determination efforts so as to maximize their biological impact and to extend substantially the known structural repertoire of the protein universe. Throughout this article we focused on the worldwide goal in solving all fold representatives and have discussed computational methods for deriving suitable targets for structural determination. These approaches provide selected lists of representative proteins that are likely to represent new structural folds.

The solved structures can serve as templates for building models for other proteins based on homology. Once these target sequences have been determined, the next stage is to use the information in these structural modeling, for the detection of structural similarities and new topologies. By structural comparison of these structures with known structures we may assign a putative function for proteins whose molecular function is unknown (Danchin, 1999; Zarembinski et al., 1998).

It should be noted that not all these structures will eventually reveal new folds. About 10% of all folds in SCOP are folds with multiple superfamilies. Extreme cases are the TIM barrel, Ferredoxin, Flavodoxin, etc. (Brenner et al., 1997). For many of these superfamilies, no ancestor protein is predicted, therefore, sharing the same fold is a result of convergent evolution. Consequently, we expect that some of the proteins that were selected as targets will turn out to be new superfamilies that belong to already known folds. Still, solving the structures of new proteins whose fold is already known may have a tremendous impact on biological and biomedical sciences. The assumption is that for drug design and in the case of many genetic diseases, the availability of many similar structures will be essential. The code of catalysis, affinity and specificity may be deciphered by high-resolution structures of closely related proteins.

While target selection is the first essential step in structural genomics, it would not have been possible to proceed without substantial advances in technology of high-resolution structure determination. This includes breakthrough in the use of synchrotron beamlines, new detectors and multiple-wavelength anomalous diffraction (Moffat and Ren, 1997). In addition, new NMR methods increased the size of proteins that can be solved by this technique. From computational aspects, advanced systems for rapid data collection, processing, and model building have become available. In general, we expect that computational driven target selection will help to achieve the goal set by the structural genomics foundation — a comprehensive structural view on the secrets of life.

Acknowledgements

We thank Nathan Linial for his mathematical advice and suggestions for the approach described in Section 3.4.1 and Steven Brenner and the SCOP team for their help with their files. We thank Michael Levitt and David McKay for critically reading this manuscript and for making many helpful comments. Fig. 1 was taken from NIH report. We thank D. Vitkup, E. Melamud, J. Moult and C. Sander for providing these unpublished results.

This study was partially supported by the Israeli Academy of Science (Initiatives in Res. in Sc. & Technology) and the Horowitz Fund. Golan Yona is supported by a Burroughs-Welcome Fellowship from the Program in Mathematics and Molecular Biology (PMMB).

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bairoch, A., Apweiler, R., 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. 27, 49–54.

Barker, W.C., Pfeiffer, F., George, D.G., 1996. Superfamily classification in PIR-international protein sequence database. Methods Enzymol. 266, 59–71.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., Sonnhammer, E.L., 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. 27, 260–262.

Bejerano, G., Yona, G., 1999. Modeling protein families using probabilistic suffix trees. The proceedings of RECOMB 99, pp. 15–24.

Bork, P., Eisenberg, D., 1998. Sequences and topology — Deriving biological knowledge from genomic sequences. Curr. Opin. Struct. Biol. 8, 331–332.

Bork, P., Koonin, E.V., 1998. Predicting functions from protein sequences — where are the bottlenecks. Nat. Genetics 18, 313–318.

Brenner, S.E., Chothia, C., Hubbard, T.J.P., 1997. Population statistics of protein structures: lessons from structural classifications. Curr. Opin. Struct. Biol. 7, 369–376.

Brenner, S.E., Chothia, C., Hubbard, T.J.P., 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. USA 95, 6073–6078.

Brenner, S.E., Barken, D., Levitt, M., 1999. The PRESAGE database for structural genomics. Nucleic Acids Res. 27, 251–253.

Bryant, S.H., 1996. Evaluation of threading specificity and accuracy. Proteins 26, 172–185.

Chothia, C., 1992. One thousand families for the molecular biologist. Nature 357, 543–544.

Corpet, F., Gouzy, J., Kahn, D., 1999. Recent improvements of the ProDom database of protein domain families. Nucleic Acids Res. 27, 263–267.

Danchin, A., 1999. From protein sequence to function. Curr. Opin. Struct. Biol. 9, 363–367.

Elofsson, A., Sonnhammer, E.L., 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. Bioinformatics 15, 480–500.

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

Flores, T.P., Orengo, C.A., Moss, D., Thoronton, J.M., 1993. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 2, 1811–1826.

Gaasterland, T., 1998. Structural genomics: bioinformatics in the driver's seat. Nat. Biotechnol. 16, 625–627.

George, D.G., Barker, W.C., Mewes, H.W., Pfeiffer, F., Tsugita, A., 1996. The PIR-International protein sequence database. Nucleic Acids Res. 24, 17–20.

Gerstein, M., Levitt, M., 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. Protein Sci. 7, 445–456.

Gonnet, G.H., Cohen, M.A., Benner, S.A., 1992. Exhaustive matching of the entire protein sequence database. Science 256, 1443–1445.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., Claverie, J.M., 1993. Ancient conserved regions in new gene sequences and the protein databases. Science 259, 1711–1716.

Hasson, M.S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G.L., Babbitt, P.C., Gerlt, J.A., Petsko, G.A., Ringe, D., 1998. Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. Proc. Natl. Acad. Sci. USA 95, 10396–10401.

Hendrickson, W.A., Horton, J.R., LeMaster, D.M., 1990. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J. 9, 1665–1672.

Hilbert, M., Bohm, G., Jaenicke, R., 1993. Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins 17, 138–151.

Hofmann, K., Bucher, P., Falquet, L., Bairoch, A., 1999. The PROSITE database, its status in 1999. Nucleic Acids Res. 27, 215–219.

Holm, 1999. Protein sequence space partitioning (PSSP) http://columba.ebi.ac.uk:8765/holm/pssp.

Holm, L., Sander, C., 1997a. New structure-novel fold? Structure 5, 165–171.

Holm, L., Sander, C., 1997b. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins 28, 72–82.

Hubbard, T.J., Ailey, B., Brenner, S.E., Murzin, A.G., Chothia, C., 1999. SCOP: a Structural Classification of Proteins database. Nucleic Acids Res. 27, 254–256.

Jones, S., Thornton, J.M., 1997. Prediction of protein-protein interaction sites using patch analysis. J. Mol. Biol. 272, 133–143.

Karplus, K., Barrett, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856.

Kasuya, A., Thornton, J.M., 1999. Three-dimensional structure analysis of PROSITE patterns. J. Mol. Biol. 286, 1673–1691.

Kim, K.K., Hung, L.W., Yokota, H., Kim, R., Kim, S.H., 1998. Crystal structures of eukaryotic translation initiation factor 5A from *Methanococcus jannaschii* at 1.8 Å resolution. Proc. Natl. Acad. Sci. USA 95, 10419–10424.

Kim, S.H., 1998. Shining a light on structural genomics. Nat. Struct. Biol. 5, 643–645.

Koehl, P., Levitt, M., 1999. A brighter future for protein structure prediction. Nat. Struct. Biol. 6, 108–111.

Koonin, E.V., Tatusov, R.L., Galperin, M.Y., 1998. Beyond complete genomes: from sequence to structure and function. Curr. Opin. Struct. Biol. 8, 355–363.

Krause, A., Vingron, M., 1998. A set-theoretic approach to database searching and clustering. Bioinformatics 14, 430–438.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., Haussler, D., 1996. Hidden Markov models in computational biology: application to protein modeling. J. Mol. Biol. 235, 1501–1531.

Lemer, C.M., Rooman, M.J., Wodak, S.J., 1995. Protein structure prediction by threading methods: evaluation of current techniques. Proteins 23, 337–355.

Levitt, M., 1992. Accurate modelling of protein conformation by automatic segment matching. J. Mol. Biol. 226, 507–533.

Levitt, M., Gerstein, M., 1998. A unified statistical framework for sequence comparison and structure comparison. Proc. Natl. Acad. Sci. USA 95, 5913–5920.

Lim, J.H., Yu, Y.G., Han, Y.S., Cho, S.j., Ahn, B.Y., Kim, S.H., Cho, Y., 1997. The crystal structure of an Fe-superoxide dismutase from the hyperthermophile Aquifex pyrophilus at 1.9 Å resolution: Structural basis for thermostability. J. Mol. Biol. 270, 259–274.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., armirantzou, M., Laskowski, R.A., Mitchell, J.B., Taroni, C., Thornton, J.M., 1998. Protein folds and functions. Structure 6, 875–884.

McKusick, V.A., 1997. Genomics: Structural and functional studies of genomes. Genomics 45, 244–249.

Mirny, L.A., Shakhnovich, E.I., 1998. Protein structure prediction by threading. Why it works and why it does not? J. Mol. Biol. 283, 507–526.

Moffat, K., Ren, Z., 1997. Synchroton radiation applications to macromolecular crystallography. Curr. Opin. Struct. Biol. 7, 689–696.

Montelione, G.T., Anderson, S., 1999. Structural genomics: keystone for a Human Proteome Project. Nat. Struct. Biol. 6, 11–12.

Murzin, A.G., 1993. OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. 12, 861–867.

Murzin, A.G., 1996. Structural classification of proteins: new superfamilies. Curr. Opin. Struct. Biol. 6, 386–394.

Ogata, C.M., 1998. MAD phasing grows up. Nature Struct. Biol. 5, 638–640.

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

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., Chothia, C., 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J. Mol. Biol. 284, 1201–1210.

Pearson, W.R., 1996. Effective protein sequence comparison. Methods Enzymol. 266, 227–258.

Pearson, W.R., 1997. Identifying distantly related protein sequences. Comp. Appl. Biosci. 13, 325–332.

Portugaly, E., Linial, M., 2000. Estimating the probability of a protein to have a new fold based on a map of all protein sequences. Unpublished results, presented in ISMB'99 poster session.

Russell, R.B., 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J. Mol. Biol. 279, 1211–1227.

Russell, R.B., Sasieni, P.D., Sternberg, M.J.E., 1998. Supersites within superfolds. Binding site similarity in the absence of homology. J. Mol. Biol. 282, 903–918.

Sali, A., 1998. 100,000 protein structures for the biologist. Nat. Struct. Biol. 5, 1029–1032.
Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9, 56–68.
Shapiro, L., Lima, C.D., 1998. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. Structure 6, 265–267.
Sternberg, M.J., Bates, P.A., Kelley, L.A., MacCallum, R.M., 1999. Progress in protein structure prediction: assessment of CASP3. Curr. Opin. Struct. Biol. 9, 368–373.
Tatusov, R.L., Eugene, V.K., David, J.L., 1997. A genomic perspective on protein families. Science 278, 631–637.
Terwilliger, T.C., Waldo, G., Peat, T.S., Newman, J.M., Chu, K., Berendzen, J., 1998. Class-directed structure determination: foundation for a protein structure initiative. Protein Sci. 7, 1851–1856.
Thornton, J.M., Flores, T.P., Jones, D.T., Swindells, M.B., 1991. Protein structure. Prediction of progress at last. Nature 354, 105–106.
Wang, Z., 1996. How many fold types of protein are there in nature? Proteins 26, 186–191.
Wolf, Y.I., Brenner, S.E., Bash, P.A., Koonin, E.V., 1999. Distribution of protein folds in the three superkingdoms of life. Genome Res. 9, 17–26.
Wu, C., Shivakumar, S., Huang, H., 1999. ProClass protein family database. Nucleic Acids Res. 27, 272–274.
Yona, G., Linial, N., Linial, M., 2000. ProtoMap: Automatic classification of protein sequences, and hierarchy of protein families, and local maps of the protein space. Nucleic Acids Res. 28, 49–55.
Yona, G., Levitt, M., 2000. A unified sequence-structure classification of protein sequences: combining sequence and structure in a map of protein space. The proceedings of RECOMB 00, pp. 308–317.
Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R., Kim, S.H., 1998. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc. Natl. Acad. Sci. USA 95, 15189–15193.
Zhang, C., DeLisi, C., 1998. Estimating the number of protein folds. J. Mol. Biol. 284, 1301–1305.