



## Target space for structural genomics revisited

Jinfeng Liu<sup>1,2</sup> and Burkhard Rost<sup>2,3,\*</sup>

<sup>1</sup>Department of Pharmacology, Columbia University, 630 West 168th Street, New York, USA, <sup>2</sup>CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, NY 10032, New York, USA and <sup>3</sup>Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA

Received on June 5, 2001; revised on January 4, 2002; accepted on February 7, 2002

### ABSTRACT

**Motivation:** Structural genomics eventually aims at determining structures for all proteins. However, in the beginning experimentalists are likely to focus on globular proteins to achieve a rapid basic coverage of protein sequence space. How many proteins will structural genomics have to target? How many proteins will be excluded since we already have structural information for these or since they are not globular? We have to answer these questions in the context of our target selection for the North-East Structural Genomics Consortium (NESG).

**Results:** We estimated that structural information is available for about 6–38% of all proteins; 6% if we require high accuracy in comparative modelling, 38% if we are satisfied with having a rough idea about the fold. Excluding all regions that are not globular, we found that structural genomics may have to target about 48% of all proteins. This corresponded to a similar percentage of residues of the entire proteomes (52%). We explored a number of different strategies to cluster protein space in order to find the number of families representing these 48% of structurally unknown proteins. For the subset of all entirely sequenced eukaryotes, we found over 18 000 fragment clusters each of which may be a suitable target for structural genomics.

**Availability:** All data are available from the authors, most results are summarized at: [http://cubic.bioc.columbia.edu/genomes/RES/2002\\_bioinformatics/](http://cubic.bioc.columbia.edu/genomes/RES/2002_bioinformatics/)

**Contact:** E-mail: [rost@columbia.edu](mailto:rost@columbia.edu)

### INTRODUCTION

#### Structural genomics to determine all native protein structures

In 2000, the National Institutes of Health (NIH) in the USA began to finance pilot projects for large-scale protein structure determination (structural genomics). Two major objectives of structural genomics have often been given. First, experimentally determine one protein structure for

each natural protein (NIGMS, 2001). Second, determine one structure for all missing links in pathways and biological mechanisms (Blundell and Mizuguchi, 2000; Burley *et al.*, 1999; Christendat *et al.*, 2000; Gaasterland, 1998a,b; Lima *et al.*, 1997; Moulton and Melamud, 2000; Rost, 1998; Shapiro and Harris, 2000; Teichmann *et al.*, 1999; Thornton, 2001). These two objectives correspond to the two aspects of genome sequencing: (i) the mass of data, and (ii) the completeness of entirely sequenced organisms. One expected technical benefit from structural genomics is the development of techniques and protocols for large-scale expression, purification, crystallization and structure-determination. An important benefit for molecular biology may be the determination of the structural scaffolds for most basic functional elements. A considerable increase in the fraction of proteins for which we have some structural information may also advance the determination of function for single proteins or entire proteomes. It is commonly assumed that the scaffolds of protein folds constitute one of the ‘basic units’ for evolution. If so, structural genomics will also help to better understand evolution. Structural genomics focuses on structural modules or domains. However, isolated domains do not always suffice to understand function. Instead, understanding function often requires studying complexes composed of many proteins. The difficulty of determining structures for large complexes will be prohibitive for the first round of structural genomics.

#### Determine one structure for each family of closely related proteins

The safest strategy to go about determining structures for all native proteins is to simply express, purify, crystallize and x-ray all protein sequences one by one, just in the way large-scale genome sequencing operates. However, sequencing is technically much simpler than structure determination. None of the necessary steps—express, purify, crystallise, x-ray—has ever been accomplished on the scale of ‘all proteins in a proteome’. Consequently, we have to find a way of focusing on some representative

\*To whom correspondence should be addressed.

fraction of all proteins. Resources such as CATH (Orengo *et al.*, 1997), FSSP (Holm and Sander, 1999), HSSP (Schneider *et al.*, 1997), or SCOP (Lo Conte *et al.*, 2000) illustrate that fewer than 1000 folds and about 2500 families are representative for over 20 000 structures deposited in PDB (Berman *et al.*, 2000). Hence, the conceptually simple refinement of the selection strategy is to determine one structure for each unknown fold. Unfortunately, this straightforward concept hides a number of severe problems. The first is that the absolute majority of similar folds have less than 12% pairwise sequence identity (Rost, 1997, 1999; Yang and Honig, 2000b), i.e. populate the midnight zone of sequence comparisons in which we cannot detect the fold similarity from sequence alone. Hence, we would have to determine the fold to find the set of representative folds. One way around this vicious circle is to reformulate the goal: determine one structure for each family of proteins that are related by sequence. The levels of pairwise sequence similarity that imply similarity in structure are well established (Abagyan and Batalov, 1997; Brenner *et al.*, 1998; Muller *et al.*, 1999; Park *et al.*, 1997, 1998; Rost, 1999; Sander and Schneider, 1991; Yang and Honig, 2000b). Thus, it may seem that all bioinformatics has to do is to cluster all proteins into families of proteins with similar structures, exclude all clusters with known structures and define the remaining list as the target list for structural genomics (Vitkup *et al.*, 2001). In fact, this procedure describes the current *modus operandi* of structural genomics initiatives fairly well. Additionally, most groups exclude clusters that are particularly problematic due to the presence of membrane regions, and/or long regions of low-complexity.

### The age of structural genomics has begun

The currently active structural genomics groups differ in their focus. Most groups focus on particular organisms: *Mycoplasma genitalium* and *Mycoplasma pneumoniae* by BSGC (Kim, 2001), *Caenorhabditis elegans* by JCSG (Wilson, 2001), *Mycobacterium tuberculosis* by TBSGC (Terwilliger, 2001), *Caenorhabditis elegans* and *Pyrococcus furiosus* by SECSG (Wang, 2001), *Saccharomyces cerevisiae* by the YSG (YSG, 2001), *Thermus thermophilus* by SRG (Yokoyama and Kuramitsu, 2001), *Homo sapiens* by PSF (Umbach, 2001). Two groups focus on particular protein types (short proteins from eukaryotes by NESG; Montelione, 2001, disease related and 'easy' proteins by MCSG; Joachimiak, 2001), and one on particular functional types (enzymes by NYSGRC; Burley, 2001). The nine initiatives currently financed by the National Institute of Health (NIH) in the USA together intend to add about 2000 structures over the next four years. Given that almost 3000 new protein chains have been added to PDB (Berman *et al.*, 2000) over the last 12 months, this number may appear small. However, of

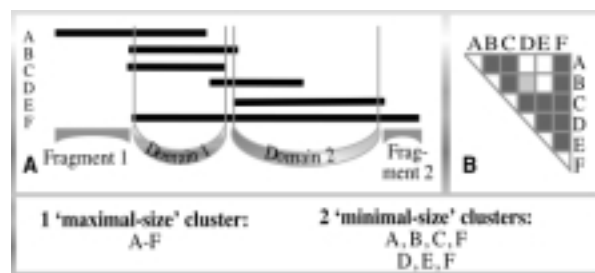
the 3000 structures added to PDB in 2001, only about 500 belonged to families of unknown structure (Berman *et al.*, 2000; Eyrich *et al.*, 2001a,b). Since the structural genomics initiatives set out to determine structures exclusively for such families of unknown structures their yield would double the number of families for which structures will be added until 2005. Another implicit goal of all structural genomics initiatives is to reduce the costs of determining a protein structure from its current value of about \$100K/protein. Interestingly, the US-based pilot groups receive about this amount from the NIH to determine their projected 2000 structures. However, the goal of the first round of structural genomics is not primarily to determine as many structures as possible, rather it is to pioneer the development of techniques that will be required for a cost-efficient large-scale structure determination.

### Existing methods that cluster sequence-space

Over the last years, a number of groups have presented different approaches to cluster sequence space. CATH (Orengo *et al.*, 1997), FSSP (Holm and Sander, 1999), and SCOP (Lo Conte *et al.*, 2000) group proteins of known structure according to their fold. These classifications can then be extended to homologous proteins for which we do not experimentally know structure—a concept pioneered in the HSSP database (Sander and Schneider, 1991). When we want to group proteins into families without knowing the structure of any protein in that family, the problem becomes how to define the boundaries of which proteins to include. For example, should proteins A and F in Figure 1 become part of the same family, or should we try to chop both A and F into domains and build a family labelled 'Domain 1' in Figure 1? PFAM (Bateman *et al.*, 2000; Sonnhammer *et al.*, 1997) is an expert annotated database of protein families that tries to build multiple alignments of regions in proteins that are believed to constitute domains. One limitation of PFAM is that not all known proteins are included yet. Even more limited in that respect is the similar approach toward listing all domains of secreted proteins in SMART (Ponting *et al.*, 1999). COG (Tatusov *et al.*, 2000, 1997) builds clusters of orthologous groups (COGs: proteins in different species that evolved from a common ancestral protein) or orthologous sets of paralogues (proteins from the same organism which are believed to be related by duplication) from at least three species. The authors try to split multi-domain clusters through pair relations. ProtoMap (Linial and Yona, 2000; Yona *et al.*, 1998, 1999, 2000) is an automatic, hierarchical classification of the entire SWISS-PROT database (Bairoch and Apweiler, 2000a) that is based on pairwise relations. The particular algorithm introduced by ProtoMap for merging and splitting groups of pairwise related proteins, yields an implicit

separation into clusters with single and multiple domains. An attempt at combining sequence-based and structure-based classifications is implemented in BioSpace that first clusters all proteins of known structures and then pulls in proteins of unknown structures in a way similar to the ProtoMap algorithm. Finding consensus motifs in alignments and then cutting according to some statistical criteria is the concept that leads to the automatic classification of all proteins in ProDom (Corpet *et al.*, 2000). The particular problem of ProDom is that the domains found tend to be shorter than those assigned from known protein structures. The basic idea of using boundaries in alignments to identify domains has also been implemented by other groups (Enright and Ouzounis, 2000; Marcotte *et al.*, 1999). In particular, the GeneRAGE (Enright and Ouzounis, 2000) algorithm appears to yield domains that resemble structural domains. ProClass classifies proteins into families based on PROSITE sequence motifs (Bairoch *et al.*, 1997; Hofmann *et al.*, 1999) and PIR super-families (Barker *et al.*, 2000). Domains are not explicitly detected by ProClass, rather they are taken from previous annotations (from experts, PFAM, or ProDom). Picasso (Heger and Holm, 2000) is another approach clustering protein space based on pairwise relations. It seems that Picasso splits domains in a way similar to the GeneRAGE algorithm. The idea of mapping the space of all proteins implies that we have some sort of metric that defines a distance between two groups. The problem with this concept is that we can only measure the similarity not the distance between two proteins. For example, assume proteins A and B are both 100 residues long. If they have 33 pairwise identical residues, we can infer that they have similar structures (Rost, 1999). If they only have 25 pairwise identical residues we know that the odds are one in ten that A and B have similar structure, however, these odds reflect our lack of knowledge of the relation between A and B rather than their actual structural similarity. In fact, A and B may structurally be more similar than a pair A'-B' with 33 identical residues. Furthermore, assume we have a globin, an immunoglobulin and a TIM-barrel. We know that the three are not similar, however, we cannot unambiguously define a distance relationship that concludes something such as the globin is more similar to a TIM-barrel than it is to the immunoglobulin. Amongst all the clustering attempts, ProtoMap appears to be the one that most successfully introduces a kind of distance metric (Linial and Yona, 2000).

Here, we re-evaluated earlier estimates (Liu and Rost, 2001; Teichmann *et al.*, 1999; Vitkup *et al.*, 2001) for the number of structural families to target by structural genomics efforts. We also presented two clustering strategies that illustrated problems with the simple concept of 'one structure per family'. In particular, our maximal-size clusters illustrated that we fail to cluster sequence-space if we



**Fig. 1.** Concepts of clustering and domain splitting. Regions in which the six proteins A–F have significant pairwise sequence similarity are marked as black lines (A). The particular pairs of 'significant similarity' are given in the matrix (B: grey boxes mark similar pairs). The six proteins group into two 'minimal-size' clusters, with protein F belonging to both. The first of the two clusters constitutes one HSSP file (Holm and Sander, 1999; Schneider *et al.*, 1997). The 'maximal-size' clustering assumes that we fail to dissect proteins into domains and want to ascertain that no two clusters have residual similarity. One way of dissecting proteins into domains is the simple triangular inequality:  $F = E$  (read 'similar to'),  $F = A$ , but  $A \neq E$  (read 'not similar to') that yields a split of F into two domains. Note that C is not split into two domains because its similarity to D is assumed to be on the borderline, i.e. below some given threshold (indicated by light grey in B).

do not dissect the sequences into structural domains before we start clustering. Our preliminary implementation of a domain-dissection approach suggested that structural genomics initiatives might have to target over 18 000 fragment clusters in eukaryotes alone. This estimate resulted from the proteins that we selected in our second round for the target selection of the North-East Structural genomics Consortium (NESG; <http://www.nesg.org>).

## METHODS

### Source of sequences

We obtained the sequences for the entire proteomes of the 30 organisms we analysed from the public domain. All ORFs were downloaded from <ftp://ncbi.nlm.nih.gov/genbank/genomes/>, except for *Homo sapiens* (from SWISS-PROT release 39 and TrEMBL database release 15), *Drosophila melanogaster* (from <http://www.fruitfly.org/>, release 2), and *Caenorhabditis elegans* (from [http://www.sanger.ac.uk/Projects/C\\_elegans/wormpep/](http://www.sanger.ac.uk/Projects/C_elegans/wormpep/), wormpep 65).

### Prediction methods

**Search for similar proteins.** We detected similar sequences in two ways. (1) Run PSI-BLAST (Altschul *et al.*, 1997) searches against all known sequences contained in SWISS-PROT (Bairoch and Apweiler, 2000b), TrEMBL (Bairoch and Apweiler, 2000b), and PDB

(Berman *et al.*, 2000). For simplicity, we refer to the combination of these three databases as the set BIG. We first searched against a filtered version of BIG and then used the final profile to search against the unfiltered BIG (Jones, 1999; Przybylski and Rost, 2002). We included all hits below a PSI-BLAST *E*-value of  $10^{-3}$ . We tested various thresholds for ‘significant sequence similarity to protein of known structure’. Firstly, we included all protein pairs with more than 50% pairwise identical residues (corresponding to ‘high accuracy in comparative modelling’). Secondly, we included all pairs above the refined HSSP-curve (medium accuracy in comparative modelling) relating the length of the alignment to the respective pairwise sequence identity/similarity (Rost, 1999; Sander and Schneider, 1991). Thirdly, we included all pairs with PSI-BLAST *E*-values below  $10^{-3}$  (for most of these, comparative modelling supposedly identifies the basic fold schematically).

*Predict membrane proteins.* We used only the filtered MaxHom alignments (Rost, 1999) for predicting membrane regions by the program PHDhtm (Rost, 1996; Rost *et al.*, 1995, 1996) using the default threshold of 0.8. We adjusted the total number of membrane proteins according to the false positive rate (1.6%) and false negative rate (3%) published in the original paper (Rost *et al.*, 1996):

$$n = \frac{1 - FP}{1 - FN - FP} \cdot n_{\text{pred}} - \frac{FP}{1 - FN - FP} \cdot n_{\text{total}} \quad (1)$$

where *n* was the final number of membrane proteins we reported, FP and FN were the false positive and false negative rates respectively, *n*<sub>pred</sub> was the number of predicted membrane proteins in the genome, and *n*<sub>total</sub> was the total number of proteins in the genome. Note: our notion of ‘membrane proteins’ is restricted to integral helical membrane proteins. In particular, we ignored proteins anchoring helices in the membrane or those inserting beta-strands (porins) since these classes of proteins cannot be identified from sequence information alone.

*Predicting signal peptides.* We predicted signal peptides using the program SignalP (Nielsen *et al.*, 1996, 1997). We considered a protein to contain a signal peptide if the ‘mean S’ value in the prediction was above the default threshold. The accuracy of SignalP was estimated to be around 90% (Emanuelsson *et al.*, 2000; Nielsen *et al.*, 1997). We excluded archaeobacteria from the analysis since SignalP was developed for prokaryotes and eukaryotes.

*Predicting coiled-coil helices.* We used the program COILS (Lupas, 1996; Lupas *et al.*, 1991) to predict coiled-coil regions, with the window-size set to 28 residues and the threshold for probability set to 0.9.

*Identifying regions of low-complexity (SEG).* We labelled regions of low-complexity using the program SEG (Wootton and Federhen, 1993, 1996) using the default parameters.

*Identifying regions with no regular secondary structure (NORS).* Using the filtered MaxHom alignments, we used PHDsec (Rost, 1996; Rost and Sander, 1993, 1994) to predict secondary structures. We considered stretches of more than 70 consecutive residues with less than 12% predicted helix or strand as ‘NORS’ (Liu *et al.*, 2002).

*Operational definition for removing fragments from the ‘to-do’ list.* Many proteins of known structure contain regions of low-complexity (Romero *et al.*, 1998; Saqi, 1995). However, proteins that contain almost no high-complexity regions constitute—at best—low-priority targets for structural genomics. We removed all proteins that had fewer than 50 residues in non-membrane, non-coiled, non-signal peptide, non-SEG, or non-NORS regions.

*Clustering sequence space.* In order to cluster sequence space for eukaryotes, we tested the following three approximations (Figure 1). (1) Maximal cluster size: merge all proteins that have some local similarity (BLAST score  $<10^{-3}$ ) to one another into one cluster; merge clusters as long as they have common members. (2) Minimal cluster size: given any two proteins A and B, group these into one cluster if the sequence similarity between the pair is above a threshold (BLAST score  $<10^{-3}$ ). While the maximal clustering is independent of the starting point, the final clusters resulting from the minimal clustering do differ. We followed the algorithm encoded in GeneRAGE (Enright and Ouzounis, 2000) by starting from single-domain proteins (Figure 1). Once we compiled the minimal-size clusters, we took the domains implied by the clustering and split those further.

## RESULTS

### We have some idea about structure for 6–38% of all proteins

We have explicit experimental information about structure for less than 0.3% of all entirely sequenced proteomes. The answer to the question for which fraction of entire proteomes we can predict structure by comparative modelling depends on the accuracy we require for the model. One extreme point is to model only proteins for which the respective experimental structure has more than 50% pairwise identical residues. At that level, models are typically very accurate ( $<3 \text{ \AA}$  C $^{\alpha}$ -rmsd) (Eyrich *et al.*, 2001a,b; Marti-Renom *et al.*, 2000, 2001). For all the 30 proteomes that we analysed (Appendix, Table W, we found that about 6% of the proteins can be modelled at

this level of accuracy (Figure 2, left panel, black bars). Next, we tested a level of average accuracy at which the models provide a good idea of the basic fold (around 5–6 Å C $\alpha$ -rmsd) (Eyrich *et al.*, 2001a,b; Marti-Renom *et al.*, 2000, 2001). At that ‘cartoon-level’ of model accuracy, we found similar structural regions for about 20% of all proteins (Figure 2, left panel, grey bars). Finally, we dropped the requirement for model accuracy entirely, and tested a threshold at which the model most often captures basic features of the respective structure. At that level, we found structurally known regions in 38% of all proteins (Figure 2, central panel, grey bars). Note in particular the extreme increase in coverage when using PSI-BLAST searches against the BIG database. The reason for this non-linear behaviour was that pairs of fairly diverged sequences dominated most structural families (Appendix, Figure W).

### 30–40% of all proteins contain non-globular regions

We found at least one membrane helix for about 22% of all proteins (Figure 2, right panel, black bars). About half of all predicted membrane proteins had more than five helices (Liu and Rost, 2001). While the percentage of helical membrane proteins was similar between all three kingdoms (archae, eukaryotes, and prokaryotes), we found significantly more proteins with coiled-coil regions in eukaryotes (eukaryotes > 10%; prokaryotes + archae < 5%, total about 8%; Figure 2, right panel, striped bars). Most coiled-coil proteins consisted of a single 28 residue coil (Liu and Rost, 2001). We also found that the percentage of long NORS regions (Methods) differed significantly between eukaryotes and the other two kingdoms: eukaryotes had about 25% NORS proteins, prokaryotes and archae only about 3%, bringing the total percentage to 16% (Liu *et al.*, 2002). Initially, structural genomics initiatives will discard all those proteins. The total percentage of proteins with membrane helices, coiled-coils, or NORS regions totalled to about 30–40% (Figure 2, central panel).

### About 48% of all proteins constitute targets for structural genomics

Even when avoiding membrane regions, experimentalists may still want to determine the structure for the globular region of a membrane protein. We assumed rather daringly that any region of more than 50 consecutive residues without: (i) membrane helices, (ii) coiled-coil helices, (iii) low-complexity stretches, (iv) similarity to a known structure, and (v) for which we predicted some regular secondary structure could be of interest to structural biology. After this reduction, we found about 48% of all proteins (slightly less for eukaryotes) to contain regions that could be of interest for structural genomics (Figure 2, centre). Remarkably, the respective number for the subset

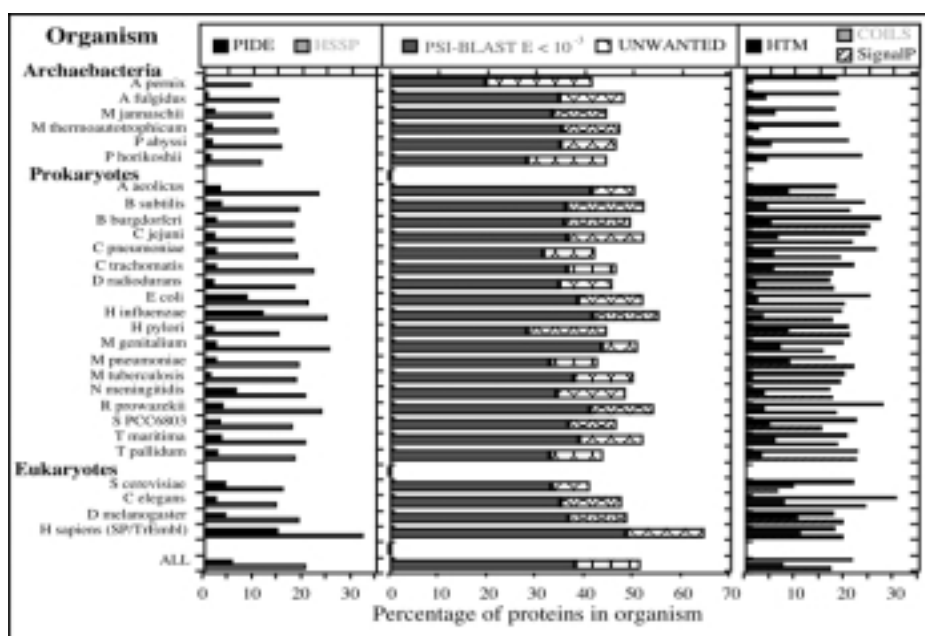
of all human proteins we used (31K) was only 35%.

### The immediate to-do list corresponds to about 54% of all residues

When estimating the percentage of proteins that structural genomics targets, we need to define arbitrary thresholds for when we consider the unwanted or structurally known regions to span enough of a protein to discard this from the to-do list. When estimating the percentage of residues in the entire proteomes that may become targets for structural genomics, we needed no assumptions about thresholds for ‘minimal globular regions’. Rather, we could simply count all residues in transmembrane helices, coiled-coil helices, low-complexity stretches, signal peptides, NORS regions, and regions for which comparative modelling could provide an idea about structure. We found that on average structural genomics will have to contribute to adding in structural information for about 54% of all residues (Figure 3). On the per-residue level, the subset of human (47%) did not differ as significantly from the average as for the protein level. This might suggest that the difference between human and others on the protein level has some reason other than that our subset was overly biased.

### Eukaryotes cluster into over 170 000 fragments

We did not have the CPU resources to cluster all proteomes. Instead, we only had results for *Methanococcus jannaschii*, *Saccharomyces cerevisiae*, and the results for all known eukaryotic proteomes (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, subset of *Homo sapiens*, and *Saccharomyces cerevisiae*). The 6357 *Saccharomyces cerevisiae* proteins fall into 3796 maximal-size and into 5448 minimal-size clusters (Table 1). The largest single maximal-size cluster contained 1351 proteins. The simple domain-splitting algorithm similar to GeneRAGE (Enright and Ouzounis, 2000) first separated and then grouped the minimal-size clusters into 3638–6867 clusters (Table 1). The data were similar for *Methanococcus jannaschii* (Table 1). When splitting ALL eukaryotes, the situation changed dramatically (Table 1): The 97K eukaryotic proteins fall into 22K maximal-size clusters with the largest single cluster containing almost half of all the proteins (46K). This result demonstrated that the maximal-size clustering was not reasonable. We were surprised by the separation of the 97K eukaryotic proteins into more than 170K fragments, i.e. by finding almost twice as many minimal-size fragment-clusters as proteins for the eukaryotes. The majority of these 170K fragments spanned over 80–150 residues (Figure 4). Overall, the length of the consensus region in each cluster corresponded to the length distribution of structural domains. The particular algorithm implemented in ProDom (Corpet *et al.*, 2000) that uses



**Fig. 2.** Estimate for the percentage of protein targets. Left panel: Percentages of proteins in respective proteome for which we found similarities to proteins of known structure above (1) pairwise sequence identities of 50% (PIDE), and (2) above the refined HSSP-threshold, e.g. given by ‘more than 33% pairwise identity over 100 residues aligned’ (Rost, 1999). Right panel: Percentages of proteins predicted with membrane helices (HTM), coiled-coil regions (COILS), and signal peptides (SignalP) in all proteomes. Centre: The lowest threshold for which we can somehow reliably predict aspects of structure through comparative modelling is an  $E$ -value in PSI-BLAST of  $10^{-3}$ . At this level, we found about 38% of all proteins to have similarity to known structures. To exclude all these proteins for target selection might be deemed highest priority. Next, we identified all the proteins without any globular region longer than 50 residues (UNWANTED). The sum over PSI-BLAST + UNWANTED marks the percentage of proteins that are certainly not interesting for target selection in the first round of structural genomics. For all proteomes this number added to about 52% leaving about 48% of all proteins as putative targets.

**Table 1.** Clustering and domain splitting of selected proteomes

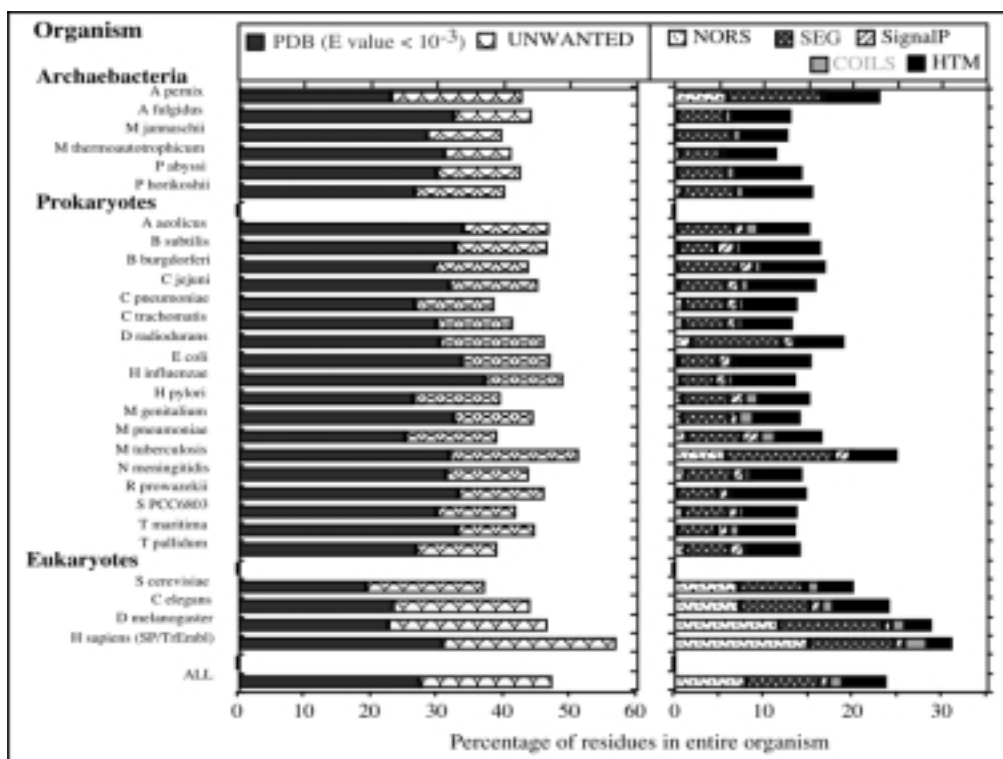
Set <sup>a</sup>	Nprot <sup>b</sup>	NminP <sup>c</sup>	NmergeP <sup>d</sup>	NmaxP <sup>e</sup>	Largest <sup>f</sup>	NminD <sup>g</sup>	NmaxD <sup>h</sup>
<i>Methanococcus jannaschii</i>	1735	1432	1070	1211	72	1459	1229
<i>Saccharomyces cerevisiae</i>	6357	5448	3337	3796	1351	6867	3638
<i>Eukaryotes</i>	97421	170186		22112	46318		
<i>Eukaryotic targets</i>						18127	15003

<sup>a</sup>Set: ‘Eukaryotes’: arabidopsis, worm, fly, yeast, and human, ‘Eukaryotic targets’ is the subset of clusters that may be targeted by structural genomics (at least one stretch of 50 residues without homologue of known structure, membrane regions, low-complexity residues, or NORS regions); <sup>b</sup>Nprot: the number of predicted proteins from the respective original publication; <sup>c</sup>NminP: the number of ‘minimal-size’ clusters; <sup>d</sup>NmergeP: the number of ‘minimal-size’ clusters after merging the clusters again with pairwise BLAST  $E$ -value of  $10^{-3}$ ; <sup>e</sup>NmaxP: the number of ‘maximal-size’ clusters; <sup>f</sup>Largest: number of proteins in largest single cluster; <sup>g</sup>NminD: the number of ‘minimal-size’ domain clusters; <sup>h</sup>NmaxD: the number of ‘maximal-size’ domain clusters.

evolutionary relations to split proteins into domains yields fragments that are too short. The differences between the fragments generated by the GenerAGE-type algorithm that we implemented and structural domains from PrISM (Yang and Honig, 1999, 2000a,b,c) indicated that the fragments we found were—on average—too long rather than too short.

### More than 16 000 targets for structural genomics were found in eukaryotes alone

The five eukaryotic proteomes corresponded to over 170K minimal-size clusters. Next, we extracted the consensus regions for all these 170K clusters, and removed all clusters that did not have at least one fragment of more than 50 consecutive residues without a homologue of known



**Fig. 3.** Estimate for the percentage of residues in putative targets. Right panel: Percentages of residues in transmembrane helices (HTM), coiled-coil helices (COILS), signal peptides (SignalP), low-complexity regions (SEG) and regions without regular secondary structure (NORS). Note: these numbers do not necessarily add up, since coiled-coil regions are occasionally detected by SEG. Left panel: Percentages of residues for which PSI-BLAST found similarities to known structures below an  $E$ -value of  $10^{-3}$ , and percentage of UNWANTED residues, i.e. those that have any of the regions listed on the right panel. These are unwanted in that they may seriously hamper a high-throughput structural genomics effort. Interestingly, the percentage of residues for putative targets was rather similar to the percentage of proteins (Figure 2).

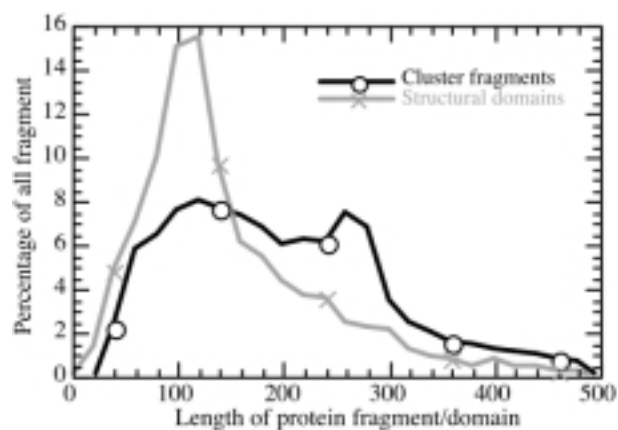
structure (according to PSI-BLAST), transmembrane- or coiled-coil helix, low-complexity or NORS region. This reduction yielded 107 410 eukaryotic fragments of potential interest to structural genomics (Table 1). An all-against-all for these 107 410 fragments resulted in 18 127 minimal-size clusters (Table 1), the largest of which contained 81 eukaryotic proteins. Finally, we mapped the 18 127 consensus regions to the Pfam-A database (Bateman *et al.*, 2000; Sonnhammer *et al.*, 1997). Most of our clusters did not correspond to any of the known 2267 Pfam-A families (82–85%, Table 2). The 3213 clusters for which HMMer found similarities of any protein in that cluster to Pfam, matched in 1208 distinct Pfam families. Most of these Pfam families (57%) matched exclusively in one of our target clusters, 77% (935) matched in at most two clusters (Table 2). At most 210 of the 3213 clusters that matched in Pfam matched

to more than one family. This number may provide an upper bound estimate for the error of our clustering if we assume that all Pfam families constitute structural domains. Thus, about 7% of our 18 127 clusters may be problematic. Consequently, we expect that we have about 17 000 targets for structural genomics in eukaryotes.

## DISCUSSION AND CONCLUSIONS

### About 48% of all proteins in the 30 proteomes constitute possible targets

Let us assume that structural genomics will have to experimentally determine structures for all proteins for which we do not have information about structure through experiments or through comparative modelling based on experimentally known homologues. We have explicit experimental information about structure for only a marginal fraction of all the proteins in currently



**Fig. 4.** Distribution of fragment lengths for eukaryotes. We found 170K clusters for the 97K eukaryotic proteins (Table 1). We suspected that this number was inappropriately high due to an oversplitting of the clustering algorithm applied (Enright and Ouzounis, 2000). However, we could not verify this suspicion when comparing the lengths of the 170 186 fragment clusters to that of structural domains from PrISM.

sequenced proteomes (<0.3%). Hence, the number of targets for structural genomics is not given by ‘all-structurally known’, rather it is given by ‘all-models’, i.e. by the number of proteins for which we can obtain structural information through comparative modelling. The size of structural families increases exponentially when lowering the threshold for detecting structural similarities (Appendix, Figure W). Lower thresholds imply lower accuracy in comparative modelling. Thus, the estimate for the number of targets for structural genomics is extremely sensitive to the accuracy we require in comparative modelling to remove a protein from the potential target list. While we have highly accurate information for only 6% of all proteins, we have low-accuracy information about structure for about 38%. In the first round of structural genomics, we may want to optimise the yield of ‘new structures’. Hence, the low-accuracy number (38%) appears to be a reasonable choice.

#### About half of all proteins constitute targets for the first round

Initially structural genomics may want to try avoiding experimental problems by targeting proteins that are as globular as possible. We found that about 48% of all the proteins contained fragments of over 50 residues that were not similar to known structures and did not contain problematic regions (membrane, coiled-coil, low-complexity, no regular secondary structure, or signal peptides, Figure 2). Interestingly, this fraction was significantly lower

**Table 2.** Eukaryotic target clusters and PFAM

#	Number of clusters		Percentage of clusters		Percentage of Pfam families	
	<i>BLAST</i>	<i>HMMer</i>	<i>BLAST</i>	<i>HMMer</i>	<i>BLAST</i>	<i>HMMer</i>
0	15 443	14 914	85.2	82.3		
1	2 565	3 003	14.2	16.6	56.6	57.0
2	107	191	0.6	1.1	23.0	20.4
3	11	19	0.1	0.1	8.7	9.9
4	1		0.0		3.7	3.6
≥5	0		0		4.7	4.8

#: Number of Pfam families that were matched by the same cluster (columns 2–5) or number of clusters matched by one Pfam family (columns 6–7). For each of the 107 410 potential eukaryotic target fragments that we grouped into 18 127 clusters (Table 1) we searched Pfam-A with two methods. (1) Align target fragments by pairwise BLAST (BLAST *E*-val of  $10^{-3}$ , columns labelled *BLAST*), and (2) align each Pfam family by HMMer to all target clusters (Pfam *E*-value of  $10^{-2}$ , columns labelled *HMMer*). Most target clusters (>82%) had no corresponding Pfam entry. Using BLAST, 2683 of our target clusters matched to one protein in 1141 distinct Pfam families; 56.6% of these Pfam families matched exclusively in one of our clusters. Using HMMer to find similarities, 3213 of our target clusters matched to one protein in 1208 distinct Pfam families; 57% of these Pfam families matched exclusively in one of our clusters.

(35%) for the subset of the 23 K human sequences that we analysed. Comparative modelling predicts structure only for the fragments that correspond to known structures. The average protein length in PDB is clearly lower than the average length of the proteins found in entire sequenced proteomes. Thus, we might expect that the percentage of all residues to target by structural genomics is significantly higher than the percentage of proteins. In fact, this expectation has recently been verified (Vitkup *et al.*, 2001). Surprisingly, we found that the 48% of all putative protein targets corresponded to about 52% of the entire residue mass of all proteomes (Figure 3). This significant difference between our results (Figure 2 and Figure 3) and the results published previously (Vitkup *et al.*, 2001) might have two reasons. Firstly, we used PSI-BLAST searches against the BIG database rather than pairwise BLAST searches against PDB (note that due to the small size of PDB, PSI-BLAST and BLAST searches against PDB basically yield the same results). Secondly, we marked all residues for which we predicted membrane or coiled-coil helices, and low-complexity or NORS regions. For all eukaryotic proteomes that we analysed, these regions added to almost half of the ‘residue mass’ excluded from the target list of structural genomics (Figure 3). Thus, our results suggested that about half of all the proteins in entire proteomes constitute potential targets for structural genomics.

### Clustering raised more questions than it answered

How to best cluster all known sequences depends on the reason for the clustering. In the context of structural genomics, the reason appears clear: find a representative set of targets. However, this seemingly straightforward concept hides a can of worms. The first problem is that of a hierarchy: The HSSP database that relates all known structures to known sequences (Sander and Schneider, 1991) implicitly treats the protein of experimentally known structure as the ‘master-representative’ of the structural family for that structure. If we use this concept, we find 4600 families in yeast, 1431 in *Methanococcus jannaschii* and about 30 000 families in all eukaryotes (data not shown). However, different structural genomics initiatives favour different organisms. Hence, we want to generate clusters without ‘master-representatives’. The obvious problem with this task is to find the basic unit for the clustering. If we assume that the ‘building blocks’ are structural domains, the problem becomes to dissect proteins of unknown structure into structural domains. Arguing that we cannot accomplish this, we identified the ‘maximal-size’ protein clusters; by construction there is no sequence similarity between any two of the single-linkage clusters. We found 1211 such clusters in *Methanococcus jannaschii* with the largest cluster containing 72 proteins (Table 1); for yeast the largest of the 3796 clusters contained over 1300 proteins. For all eukaryotes, the number of clusters appeared reasonable (22 112) but the largest cluster contained more than 46K proteins. These results suggested two conclusions. Firstly, sequence space appeared to be more continuous than we might have anticipated because almost half of all proteins are connected to one another by some local structural similarity. This may imply that domains were shuffled considerably during evolution (Apic *et al.*, 2001a,b) and/or that structural domains are not the appropriate ‘building blocks’. Secondly, the ‘maximal-size’ clustering obviously failed entirely to generate a reasonable map of sequence space when we did not split proteins into domains. Thus, we have to find some way to dissect proteins into domains. A particular way applicable to all protein sequences was suggested by Enright and Ouzounis (2000). For *Methanococcus jannaschii* this clustering/domain-splitting algorithm yielded about 1400 clusters (Table 1). The authors of GeneRAGE (Enright and Ouzounis, 2000) published a similar number, suggesting that their implementation of the major concept did not differ substantially from ours. The number of minimal-size clusters for yeast also appeared reasonable (Table 1). Interestingly, the numbers we obtained with and without explicitly starting from the already split domains did not differ very much (for yeast 3337 vs. 3638, Table 1). When we applied the algorithm to the 97K eukaryotic proteins, we obtained over 170K fragment clusters. Obviously,

the number appears rather high, suggesting that the algorithm might split proteins into regions that are too short. However, we found that the length distribution of the respective fragments was surprisingly similar to typical structural domains (Figure 4). Thus, the 170K fragments may indeed constitute the base for clustering eukaryotic sequences. We continued by excluding all the clusters that appeared of no interest to initial structural genomics approaches. Thus, we obtained 45051 fragment clusters containing 170K eukaryotic fragments. Next, we re-applied our minimal-size clustering by comparing all 170K against each other. This yielded 18 127 fragment clusters, the largest of which contained 81 proteins. Most of these clusters (82%) did not match to any Pfam family (Bateman *et al.*, 2000) (Table 2); 99% of all the clusters that matched in Pfam matched one or two Pfam families. Matches to more than two Pfam families might constitute errors in defining our clusters; the problem, in particular was that our domain-splitting approach missed many domains. Further splitting clusters is likely to increase the number of putative eukaryotic targets. A step missing from our analysis that works in the opposite direction is the attempted merging of some of the clusters through PSI-BLAST rather than pairwise relations.

### Structural genomics for eukaryotes may have to target 3000–17 000 protein fragments

We could not put up a firm conclusion as to the number of putative targets for structural genomics. One extreme answer was: less than 3000! This number based on the observation that the current PDB consists of about 2600 sequence-unique families which allow inferring low-resolution information about structure for about half of the proteins in all the proteomes we analysed. Assuming that a similar number of structures would fill in all unknowns, we need 2600 new structures to fill the white spaces. Another possible answer was: about 17 000 for eukaryotes alone! This number resulted when grouping the fragment clusters for eukaryotes that had more than 50 residues without known structure, membrane- or coiled-coil helices, and NORS- or low-complexity regions (Table 2). How many proteins will have to be added for prokaryotes and archae bacteria? To approach the answer to this question, we will first have to complete our clustering of all known proteomes. Clearly, our estimate puts the ball-park figure substantially higher than what was previously suggested (Vitkup *et al.*, 2001). While Vitkup and colleagues proposed a similar number (17 600 for all species), their estimate was valid for a level of modelling accuracy that covers less than 10% of all residues in current proteomes. In contrast, our estimate of 17 000 for eukaryotes was valid for an accuracy level at which over 45% of all residues were already covered. Furthermore, we excluded many fragments that were not excluded

by Vitkup et al. (NORS, coiled-coil, transmembrane helices, and signal peptides). Nevertheless, our results confirmed the work of Vitkup and colleagues in that structural genomics has a long way to go. If our estimates are correct, the first pilot phase of structural genomics will—at best—pave one fifth of the way by 2005.

## ACKNOWLEDGEMENTS

Thanks to Dariusz Przybylski (Columbia) and to Volker Eyrich (Columbia) for providing programs. We are grateful to our hard-working wet-lab colleagues from the North-East Structural Genomics Initiative (NESG), in particular to Guy Montelione (Rutgers) for his continued support and optimism. The work of JL and BR were supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last but not least, thanks to all those who deposit their experimental data in public databases, to those who maintain these databases, and to those heroes who will make structural genomics come true through their dedication and experiments.

## SUPPLEMENTARY MATERIAL

For Supplementary Material, please refer to *Bioinformatics* Online.

## REFERENCES

- Abagyan,R.A. and Batalov,S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.
- Altschul,S., Madden,T., Shaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Apic,G., Gough,J. and Teichmann,S.A. (2001a) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
- Apic,G., Gough,J. and Teichmann,S.A. (2001b) An insight into domain combinations. *Bioinformatics*, **17**, S83–89.
- Bairoch,A. and Apweiler,R. (2000a) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Bairoch,A. and Apweiler,R. (2000b) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.
- Barker,W.C. et al. (2000) The protein information resource (PIR). *Nucleic Acids Res.*, **28**, 41–44.
- Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
- Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Blundell,T.L. and Mizuguchi,K. (2000) Structural genomics: an overview. *Prog. Biophys. Mol. Biol.*, **73**, 289–295.
- Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
- Burley,S.K. (2001) *New York Structural Genomics Research Consortium*, <http://www.nysgrc.org/>, New York Structural Genomics Research Consortium (NYSGRC).
- Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) Structural genomics: beyond the human genome project. *Nature Genet.*, **23**, 151–157.
- Christendat,D. et al. (2000) Structural proteomics of an archaeon. *Nat. Struct. Biol.*, **7**, 903–909.
- Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
- Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
- Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.
- Eyrich,V., Martí-Renom,M.A., Przybylski,D., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001a) EVA: continuous automatic evaluation of protein structure prediction servers. WWW document (<http://cubic.bioc.columbia.edu/eva>) <http://cubic.bioc.columbia.edu/eva>, Columbia University.
- Eyrich,V., Martí-Renom,M.A., Przybylski,D., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001b) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.
- Gaasterland,T. (1998a) Structural genomics: bioinformatics in the driver's seat. *Nat. Biotechnol.*, **16**, 625–627.
- Gaasterland,T. (1998b) Structural genomics taking shape. *TIGS*, **14**, 135.
- Heger,A. and Holm,L. (2000) Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.*, **73**, 321–337.
- Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
- Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.
- Joachimiak,A. (2001) Midwest Center for Structural Genomics. <http://www.mcsg.anl.gov/>, Midwest Center for Structural Genomics (MCSG).
- Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
- Kim,S.-H. (2001) Berkeley Structural Genomics Center. <http://www.strgen.org/>, Berkeley Structural Genomics Center.
- Lima,C.D., Klein,M.G. and Hendrickson,W.A. (1997) Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science*, **278**, 286–290.
- Linial,M. and Yona,G. (2000) Methodologies for target selection in structural genomics. *Prog. Biophys. Mol. Biol.*, **73**, 297–320.

- Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.
- Liu, J., Tan, H. and Rost, B. (2002) Eukaryotes full of loopy proteins? *J. Mol. Biol.*, submitted
- Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
- Lupas, A. (1996) Prediction and analysis of coiled-coil structures. *Meth. Enzymol.*, **266**, 513–525.
- Lupas, A., Van Dyke, M. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
- Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
- Marti-Renom, M.A., Stuart, A., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000) Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, **29**, 291–325.
- Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A. and Sali, A. (2001) Accuracy of comparative modelling. <http://pipe.rockefeller.edu/~eva/cm/res/accuracy.html>, Rockefeller University.
- Montelione, G.T. (2001) Northeast Structural Genomics Consortium. <http://www.nesg.org/>, Northeast Structural Genomics Consortium (NESG).
- Moult, J. and Melamud, E. (2000) From fold to function. *Curr. Opin. Struct. Biol.*, **10**, 384–389.
- Muller, A., MacCallum, R.M. and Sternberg, M.J. (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.
- Nielsen, H., Engelbrecht, J., von Heijne, G. and Brunak, S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **24**, 165–177.
- Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
- NIGMS (2001) Structural genomics initiatives. [http://www.nigms.nih.gov/funding/psi/psi\\_research\\_centers.html](http://www.nigms.nih.gov/funding/psi/psi_research_centers.html), National Institute of General Medical Sciences (NIGMS).
- Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH—A hierarchic classification of protein domain structures. *Structures*, **5**, 1093–1108.
- Park, J., Teichmann, S.A., Hubbard, T. and Chothia, C. (1997) Intermediate sequences increase the detection of distant sequence homologies. *J. Mol. Biol.*, **273**, 349–354.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
- Ponting, C.P., Schultz, J., Milpetz, F. and Bork, P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.
- Przybylski, D. and Rost, B. (2002) Alignments grow, secondary structure prediction improves. *Proteins: Struct. Funct. Genet.*, **46**, 195–205.
- Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., Garner, E., Guillot, S. and Dunker, A.K. (1998) Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.*, **3**, 437–448.
- Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.*, **266**, 525–539.
- Rost, B. (1997) Protein structures sustain evolutionary drift. *Fold. & Des.*, **2**, S19–S24.
- Rost, B. (1998) Marrying structure and genomics. *Structure*, **6**, 259–263.
- Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
- Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
- Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.
- Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.*, **4**, 521–533.
- Rost, B., Casadio, R. and Fariselli, P. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.
- Sander, C. and Schneider, R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.*, **9**, 56–68.
- Saqi, M. (1995) An analysis of structural instances of low complexity sequence segments. *Prot. Eng.*, **8**, 1069–1073.
- Schneider, R., de Daruvar, A. and Sander, C. (1997) The HSSP database of protein structure–sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
- Shapiro, L. and Harris, T. (2000) Finding function through structural genomics. *Curr. Opin. Biotech.*, **11**, 31–35.
- Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.*, **28**, 405–420.
- Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
- Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
- Teichmann, S.A., Chothia, C. and Gerstein, M. (1999) Advances in structural genomics. *Curr. Opin. Struct. Biol.*, **9**, 390–399.
- Terwilliger, T. (2001) Mycobacterium tuberculosis (TB) Structural Genomics Consortium. <http://www.doe-mbi.ucla.edu/TB/>, Mycobacterium tuberculosis (TB) Structural Genomics Consortium.
- Thornton, J. (2001) Structural genomics takes off. *Trends Biochem. Sci.*, **26**, 88–89.
- Umbach, P. (2001) Protein Structure Factory. <http://www.rzpd.de/psf/>, Protein Structure Factory.
- Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.*, **8**, 559–566.
- Wang, B.-C. (2001) Southeast Collaboratory for Structural Genomics. <http://secsg.org/secs/default.html>, Southeast Collaboratory for Structural Genomics (SECSG).
- Wilson, I.A. (2001) Joint Center for Structural Genomics. <http://www.jcsg.org/>, Joint Center for Structural Genomics.

- Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
- Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.*, **266**, 554–571.
- Yang,A.S. and Honig,B. (1999) Sequence to structure alignment in comparative modeling using PrISM. *Proteins: Struct. Funct. Genet.*, **Suppl**, 66–72.
- Yang,A.S. and Honig,B. (2000a) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.*, **301**, 665–678.
- Yang,A.S. and Honig,B. (2000b) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.*, **301**, 679–689.
- Yang,A.S. and Honig,B. (2000c) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.*, **301**, 691–711.
- Yokoyama,S. and Kuramitsu,S. (2001) Structurome Research group, RIKEN. <http://www.riken.go.jp/engn/r-world/research/lab/harima/group-s/index.html>, Structurome Research group.
- Yona,G., Linial,N. and Linial,M. (1999) ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Struct. Funct. Genet.*, **37**, 360–378.
- Yona,G., Linial,N. and Linial,M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49–55.
- Yona,G., Linial,N., Tishby,N. and Linial,M. (1998) A map of the protein space—an automatic hierarchical classification of all protein sequences. *Ismb*, **6**, 212–221.
- YSG (2001) Yeast Structural genomics. <http://genomics.eu.org/>, Genoscope Evry Orsay-Gif-Saclay.