Computational structural genomics: identifying protein targets for structural studies

Igor V. Grigoriev,
Department of Chemistry, University of California, Berkeley, CA 94720 USA
E-mail: ivgrigoriev@lbl.gov

ABSTRACT
The explosion of available gene sequences and advances in the methods for determining protein structures have motivated structural genomics efforts to determine all the different types of protein structures existing in nature in order to infer protein molecular functions. Since solving a structure is time-consuming and expensive, choosing the right targets for structural studies becomes important. The focus of this work is the comprehensive evaluation of potential targets for maximizing the chances of solving new folds with a minimum number of experiments. An automatic procedure for target selection has been designed and applied to the bacterial genome of *Mycoplasma genitalium*.

Keywords
Structural genomics, *Mycoplasma genitalium*, target selection, protein structure.

1. INTRODUCTION
Genomic sequences of many organisms have been completed in the past few years [1]. However, functions are inferred for only about half of the proteins encoded by the genes in the completely sequenced genomes. For the remaining proteins knowledge of a three-dimensional (3D) structure may help to understand their molecular function [2]. However, the number of known protein structures is very limited [3] and '*ab initio*' structure prediction does not provide sufficiently accurate structural models.

Existence of remote homologues, proteins with substantially different sequences but very similar structures, indicates that structure is more conserved during evolution than protein sequence. This observation provides a basis for threading and other methods that can assign a known fold (type of protein structure) to a protein if the sequence satisfies the structural constraints of the fold [4-7]. Fold recognition that employs these methods, despite being widely used in complete genomes annotation [7-13], is limited by the number of known folds [14]. Although the total number of different protein folds in nature is unknown and varies widely in its estimates [15, 16], a large fraction of protein folds remains undiscovered.

Structural genomics intends to understand the whole variety of protein structures that exists in nature by solving protein structures in a high-throughput manner, primarily by means of X-ray crystallography and NMR spectroscopy. Solving a protein structure involves a series of different experimental steps: cloning and expression of a gene in order to produce a sufficient amount of the protein, which then has to be purified and crystallized (Figure 1). 3D coordinates of the protein atoms are decoded from X-ray diffraction data of the crystallized protein (otherwise, from NMR spectra) using intensive computations (not discussed in this paper). Computations are also required for choosing the 'right' targets, i.e. new and potentially solvable protein folds, in the onset of structural genomics projects. This work addresses the question of identifying the protein targets for structural genomics of *Mycoplasma genitalium* (MG), a human pathogen with the smallest sequenced genome of any free-living organisms [17,18].

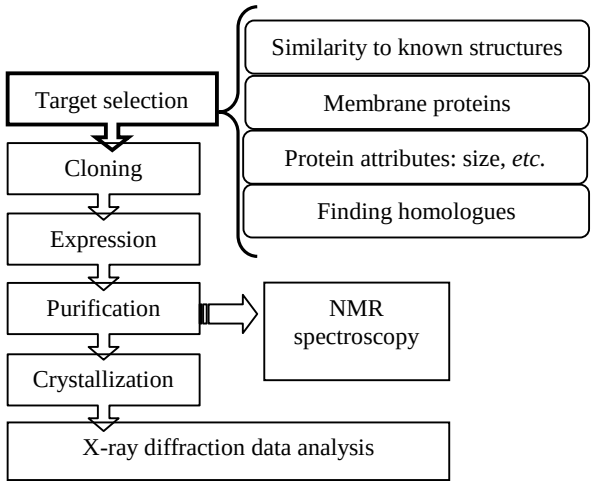

Figure 1. Structural genomics pipeline: target selection and experimental steps of structure determination.

2. TARGET SELECTION
High-throughput technology requires an automated procedure for target selection. In this work we focus on two basic components of such procedure. First, in order to work with new folds and avoid repetitive determination of similar structures, we have to exclude MG proteins with structures resembling the known 3D folds [3]. We combine two different sequence comparison methods for sequence-based structure assignment.

Secondly, the scale of structural studies requires optimization of the protein target set in order to minimize the number of experiments and prioritize the most prospective and easy targets. Here, we make the first attempt to choose the protein attributes that indicate potential delays (or advances) on different stages of structure determination experiments. A summary of such attributes will allow us to evaluate the chances of a protein to be solved. To increase these chances, some MG targets can be

Permission to make digital or hard copies of part or all of this work or personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.
SAC 2001, Las Vegas, NV
© 2001 ACM 1-58113-287-5/01/02…$5.00

substituted by their homologues from other organisms. Further in the text we discuss each of these components in details.

2.1 Similarity to known structures

We derive information about 3D structure of MG proteins combining two different approaches: (i) finding close homologues for all proteins with known 3D structure and (ii) detecting remote homologues using a representative set of protein folds.

In both cases each of the MG sequences (queries) is compared with sequences of proteins with known structures (templates) to establish their structural relatedness. However, each approach uses its own measure of reliability of protein similarity and different libraries of templates.

2.1.1 Close homologues

Close homology between protein sequences can be established using traditional sequence comparison methods [20-22]. We run a BLAST program [21] for each of MG protein sequences versus ~12,000 protein sequences from Protein Data Bank (PDB [3]). All MG proteins with at least partial similarity to known structures (e-value cutoff=10^{-10}) have been withdrawn from the first round of structural studies.

2.1.2 Remote homologues

In order to detect remote homologues we use our proximity-weighted dual profile method (PWDPM). It originated from the proximity correlation method that we developed earlier [22] and demonstrated good performance, as described elsewhere [23].

The basic concept of the PWDPM is integrating two novel aspects of protein sequence information into sequence alignment: we compare (a) evolutionary profiles instead of single amino acid sequences and (b) local segments of the sequences rather than individual residues. The evolutionary profiles corresponding to position-dependent mutation rate of amino acids have been derived from multiple alignments of close homologues of each protein sequence, using the PSI-BLAST [24]. For templates we have used a collection of protein domains representing the different folds [14, 25]. The alignments have been assessed with a heuristic scheme introduced earlier [22].

2.2 Target prioritization

There are no computational tools for predicting whether a protein will be expressed, crystallized, *etc*. Here we summarize some factors from general structural biology experience in order to determine protein attributes that often indicate certain problems or advantages in protein structure solution. A summary of such attributes allows us to set priorities for protein targets and properly plan the experiments.

2.2.1 Membrane proteins

The most difficult proteins for structure determination are membrane proteins. So far only a few of their structures have been solved because the proteins are insoluble and difficult for experimental analysis. Therefore, we excluded proteins with two or more predicted transmembrane (TM) segments from the list of potential targets. The predictions of TM segments have been provided by PEDANT [26].

2.2.2 Codon usage

The major problem with *Mycoplasmas* is that their DNA sequences code tryptophan residues by UGA nucleotide triplet, which normally terminates the protein synthesis in other organisms such as *E.coli* (used as expression systems) and produces a premature protein. The presence of several UGAs requires protein modification for successful expression.

2.2.3 Protein size

Protein size is the best hint in predicting the result of the structural experiment. Usually, smaller proteins are easier targets for structural studies. The optimal size of a protein for X-ray crystallography is between 100 and 400 residues, and less than 150 residues for NMR spectroscopy.

2.2.4 Protein sequence

Several methionine residues (Met) are usually required for decoding the X-ray diffraction data (optimally, 1 Met for every 30 amino acid residues in a sequence). If the only Met in a protein is the N-terminal one, it may be not resolved in the structure because of flexibility of the protein chain termini. In this case, disulfide-bonds formed by cysteine residues (Cys) provide the alternative for structure solution. On the other hand, exposed Cys on protein surface can cause protein aggregation as well as abundance of internal Cys-Cys may require chaperones for correct protein folding.

2.3 Amplification of target pool

Since UGA codon usage makes some MG proteins undesirable targets for structural analysis, we can substitute such proteins by homologues from other genomes. Then, the structures of the homologues, if solved, can be assigned to the respective MG genes. Such substitution is important when the protein size, number of Met residues or any other attributes discussed above are more favorable in the homologues than in the corresponding MG gene. Using homologues from thermophilic organisms provides additional advantage because their proteins are usually more stable and easier for experimental studies. Additionally, using several homologues from different genomes increases the chances to solve a structure because even subtle variations in protein sequences can be crucial for successful structure determination. Therefore, we increased the pool of targets with MG-gene homologues detected in seven completely sequences organisms (Table 1) using the program BLAST.

Table 1. Completely sequenced genomes used for amplification of MG targets

Genome	*Number of genes*
Archaeoglobus fulgidus	*2407*
Deinococcus radiodurans	*3101*
Methanococcus jannaschii	*1750*
Methanobacterium thermoautotrophicum	*1918*
Pyrococcus horikoshii	*1979*
Pyrococcus abyssi	*1765*
Thermotoga maritima	*1877*

3. RESULTS

We have applied our procedure to the complete MG genome. Earlier studies showed that ~2/3 of all MG proteins have the inferred function but structures have not been determined in most cases; the remaining ~1/3 of proteins are hypothetical, i.e. with no functional or structural information [18]. The results of our analysis of the 480 MG open reading frames (ORF) are summarized on Figure 2.

3.1 Excluded proteins

We have detected at least one close homologue with known 3D structure for 138 MG proteins, using BLAST program. Some of them are the native structures of MG proteins while the others show strong similarity to the known structures so that the reliable structural models of MG proteins can be built [27]. For an additional 113 MG proteins, the PWDP method has detected

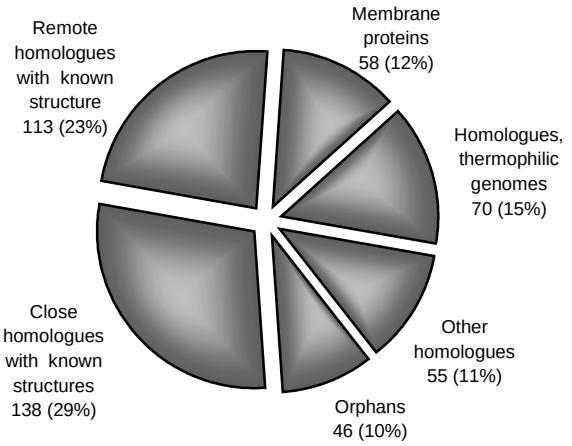

Figure 2. Characterization of MG proteins

remote homology to known structural domains. Combined, 251 out of 480 MG genes are excluded from the candidate list as being similar to known structures. Combining the methods we detect more structural homologues for MG than other methods [28].

58 membrane proteins with three or more TM segments have been excluded from the candidate list. All proteins with two TM have been found among those with known structures and excluded earlier. Proteins with single terminal TM segments have not been excluded.

Among the remaining 171 MG proteins four pairs of genes appears to be close homologues. Only one gene from each pair has been selected as candidate for structural studies on basis of the priorities discussed further in the text (e.g., MG3844861 with no UGA vs. 5 UGAs in MG3844859, or MG3844959 with twice as many Met compared with MG3844787).

3.2 MG Targets

167 MG protein are selected as potential candidates for structural studies and listed according to our prioritization scheme (Table 2). 28 proteins with TM segments have the lowest priority. Even though the terminal TM segments can be easily eliminated on the stage of cloning, it is difficult to determine the exact boundaries of the foldable protein fragment.

UGA usage in MG is the second critical factor affecting successful protein expression. Among the candidates, 60 ORFs do not use UGA codon, in 55 others, with 1-2 mainly terminal UGA, the codon can be mutated. For the remaining 52 MG proteins, considered as difficult targets, finding the homologues from other organisms becomes critical for structure solution. Therefore, the number of homologues is the next key attribute of targets, after protein size.

The distribution of the protein sizes is shown on Figure 3. Approximately 90% of the 167 candidates are within the size range optimal for structure determination. 57 proteins can be analyzed by means of NMR spectroscopy. As shown on Figure 3, most of the large MG proteins are not among the selected candidates. Such multi-domain proteins with at least one TM domain or structurally known domain will be selected for the next round of structural studies.

70 MG candidates have strong homologues in the organisms listed in Table 1 (e-value<10^{-10}). This number can be increased by lowering the desirable level of confidence in structural similarity between the homologues (BLAST e-value cutoff) or comparing with all known protein sequences (although, not all are available for experimental analysis). Still, about 46 MG genes are *Mycoplasma*-specific genes (orphans), i.e. are not homologues to any available protein sequence (e-value<10^{-5}).

One or several homologues can substitute for the remaining MG genes, because of higher priorities as potential targets than the original MG genes. For example, using the homologues almost doubles the number of targets without codon usage problem (Figure 4). Among the MG orphans 10 have three or more UGA codons: their expression will require modifications of the genes.

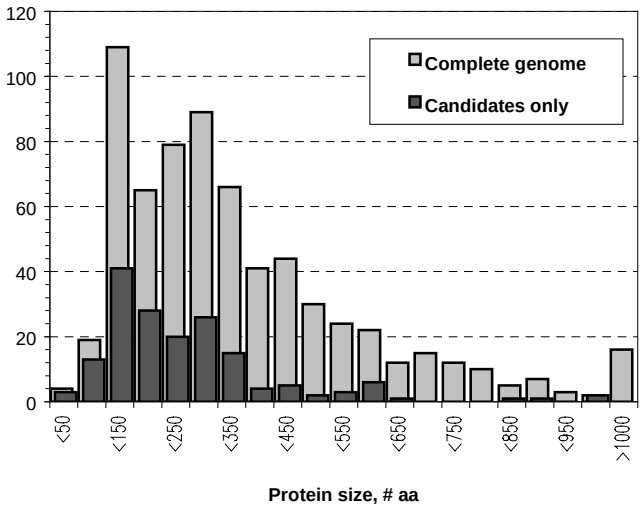

Figure 3. Distribution of MG protein sizes: candidates for structural studies vs. the complete genome

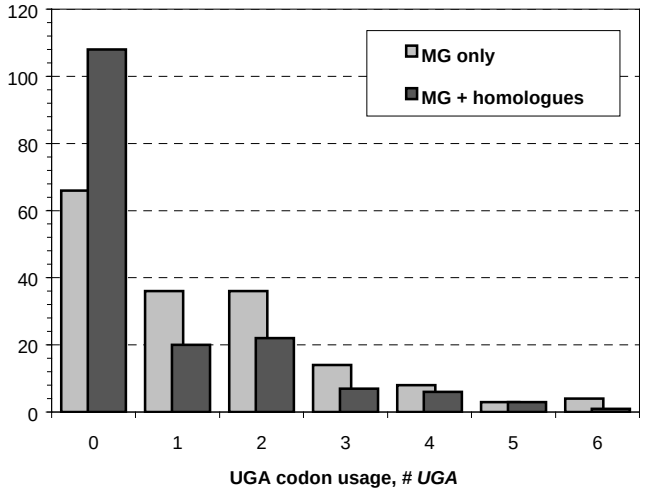

Figure 4. Effective reduction of UGA codon usage in the target pool by using MG gene homologues

Table 2. List of MG candidates for structure determination.

*GeneID	TM	U	Len	H	Me	GeneID	TM	U	Len	H	Me
3845061	-	-	48	1	3	3844770	-	-	37	0	1
3844949	-	-	57	-	2	3844904	-	-	87	-	3
3844902	-	-	53	0	1	3844794	-	-	59	0	1
3845019	-	-	64	0	1	3844950	-	-	88	0	1
1045949	-	-	97	0	1	3844838	-	-	104	2	3
3845037	-	-	119	2	3	3844744	-	-	106	2	2
3844795	-	-	124	2	2	3845009	-	-	132	2	2
3844674	-	-	139	2	2	3844679	-	-	105	1	3
3844724	-	-	100	0	5	3844747	-	-	106	0	3
3844854	-	-	119	-	3	3844836	-	-	100	0	2
3844736	-	-	112	0	2	3844761	-	-	115	0	2
3844875	-	-	129	-	2	3844919	-	-	138	-	2
3844876	-	-	138	0	2	3844718	-	-	145	0	2
3844821	-	-	148	-	2	3844756	-	-	108	0	1
3844799	-	-	122	-	1	3844840	-	-	140	-	1
3844636	-	-	190	2	6	3844763	-	-	150	2	1
3844947	-	-	162	1	5	3844824	-	-	154	1	3
3844705	-	-	157	1	3	3844988	-	-	152	1	2
1045936	-	-	165	0	4	3844994	-	-	176	0	4
3844639	-	-	186	0	4	3844806	-	-	196	0	4
1045686	-	-	176	-	3	3844833	-	-	153	0	1
3844796	-	-	167	0	1	3844968	-	-	192	0	1
3844900	-	-	227	2	5	3845027	-	-	243	2	3
3844746	-	-	211	2	2	3844912	-	-	239	2	2
3844814	-	-	207	1	2	3845002	-	-	225	0	3
3845047	-	-	292	7	3	3844719	-	-	259	2	3
3844861	-	-	278	1	6	3844857	-	-	284	1	4
3844841	-	-	294	-	4	3844807	-	-	308	2	5
3844955	-	-	328	1	7	1045963	-	-	340	0	3
3845066	-	-	437	2	6	3844688	-	-	477	5	10
3844657	-	1	48	0	1	3845039	-	1	89	1	2
3844758	-	1	61	1	1	3844772	-	1	131	4	4
3844774	-	1	123	2	3	3844771	-	1	124	2	1
1045733	-	1	145	1	2	3845020	-	1	141	-	3
3844811	-	1	147	-	2	3844757	-	1	180	6	5
3845041	-	1	150	2	3	3844805	-	1	167	0	5
3844936	-	1	166	0	3	3844965	-	1	193	0	3
3844637	-	1	151	-	2	1045685	-	1	168	-	1
3845053	-	1	169	-	1	3845038	-	1	231	2	4
3844969	-	1	218	-	3	3844689	-	1	222	-	3
3844745	-	1	257	2	4	3844706	-	1	251	-	5
3844716	-	1	285	0	5	3845033	-	1	258	-	2
3844655	-	1	315	6	4	1045727	-	1	316	0	5
3844820	-	1	346	-	3	3844941	-	1	393	2	7
3844804	-	1	597	3	7	3844956	-	1	557	2	12
3844752	-	2	138	2	7	3845010	-	2	146	2	6
3844823	-	2	102	-	3	3844964	-	2	104	-	3
3844631	-	2	145	0	4	1045980	-	2	132	-	3
3844977	-	2	101	0	2	3844819	-	2	114	0	1
3845001	-	2	157	2	1	1045731	-	2	178	0	3
3845021	-	2	171	-	2	3844644	-	2	213	2	2
3844872	-	2	244	2	2	3844707	-	2	214	-	5
3844974	-	2	236	0	3	3844753	-	2	200	0	2
3844751	-	2	268	5	7	1045746	-	2	284	2	5
3844954	-	2	262	1	6	3844635	-	2	298	0	6
3844892	-	2	297	0	4	3844958	-	2	323	2	4
3844878	-	2	320	1	7	3845029	-	2	383	0	6
3844734	-	2	531	2	5	3844929	-	3	166	1	4
3845043	-	3	237	0	2	3844699	-	3	278	1	6
3844844	-	3	292	1	4	3844691	-	3	280	0	2
3844959	-	3	324	2	10	1045964	-	3	336	2	3
3844931	-	3	322	-	2	3845056	-	3	425	6	8
3844944	-	3	411	0	3	3844647	-	3	450	1	11
3844894	-	3	599	0	3	3844850	-	4	242	3	5
3844809	-	4	219	-	4	3844622	-	4	218	0	1
3844962	-	4	274	-	5	3844961	-	4	281	0	5
1045981	-	4	347	-	5	3845011	-	4	954	3	11
3844933	-	5	311	-	10	3844986	-	5	524	-	4
3844940	-	6	280	-	2	3844887	-	6	410	3	2
3844627	-	6	600	2	6	3844916	1	-	73	0	6
3844837	1	-	99	0	1	3844996	1	-	102	0	2
3844938	1	-	137	-	2	3845048	1	-	155	0	4
3844672	1	-	311	0	3	3844694	1	1	200	1	2
3844812	1	1	268	1	4	3844743	1	1	281	-	8
3844906	1	1	295	0	3	3844803	1	1	343	0	4
1045977	1	1	556	-	5	3844980	1	2	127	-	3
1045728	1	2	123	-	2	3845034	1	2	136	-	1
1045750	1	2	137	-	1	1045982	1	2	196	-	3
3844951	1	2	224	0	4	3845032	1	2	272	-	3
3844687	1	2	477	5	4	3844684	1	2	650	-	7
1045748	1	2	806	1	21	1045996	1	2	982	3	19
3844911	1	3	212	-	3	3844783	1	3	250	-	3
1045712	1	4	591	0	7	1045986	1	5	368	-	4
1045954	1	6	874	2	10						

* MG ORF is identified by a unique GI number - *GeneID*; *TM*, *U, and Me* stands for the number of TM segments, UGA codons, and Met residues, respectively; *Len* is a number of amino acid residues in protein sequence, *H* indicates the number of genomes listed in Table 1 with homologues genes: dash marks *Mycoplasma*-specific genes, zero means possible homology to genomes not listed in Table 1.

3.3 Next step

Although we have already begun the structural analysis of the selected MG candidates and their homologues, the list of targets should be considered as preliminary for the initial round of our structural genomics project. It will be revised in a course of the project.

First, multi-domain proteins, including those with TM domains or partially solved structures, often require structural studies of the individual domains instead of the whole protein. This rises the question of accurate determination of boundaries between the structural domains.

Second, domains, which help each other to fold properly, should be co-expressed. But prediction of interactions between the domains encoded in different genes still remains a challenge, although new approaches have been developed recently [29].

So far, target prioritization has been based on simple attributes. If protein solubility or toxicity for a cell can be predicted, it will facilitate protein expression step. In turn, prediction of crystal contacts will dramatically accelerate the slowest part of protein structure solution – crystallization.

4. CONCLUSION

Structural genomics projects aims to solve as many as possible different types of protein structures in order to understand a complete structural picture of the protein universe. Because of the large scale of the structural studies, it is important to minimize the efforts required for solving a set of protein folds. In this work we describe a computational procedure for selection of protein targets for structure determination. It consists of two main parts: (i) excluding proteins with known structures in order to avoid their structure determination and focus on previously unknown folds, and (ii) predicting performance of the targets at different stages of structure determination.

We have been looking for the targets in the complete bacterial genome of *Mycoplasma genitalium*. Out of 480 genes in this

genome, 171 do not show similarity to any known structure or membrane protein and will require structure determination. In order to evaluate them as targets, we have determined and analyzed certain attributes of their protein sequences, such as length and presence of specific amino acid residues or TM segments. According to this analysis, structures of a substantial number of MG proteins can not be solved without protein modifications. However, some of these targets can be replaced by the homologues from other organisms and hence increase the number of 'easy' targets (from ~60 to ~100). Nevertheless, ~20-30 MG proteins are expected to be very difficult for structure determination because of the size, codon usage problems and absence of homologues. The list of MG targets will be revised and extended as more information and experience comes from the pilot structural genomics projects.

Different approaches to target selection were proposed earlier. In 1997, Fischer and Eisenberg listed 38 MG proteins with new folds and many homologues in sequence database as 'attractive targets for structural studies' [9]. Later, Cort *et al.* further developed the idea of structure determination for single representatives of the largest clusters of homologous proteins into a phylogenetic approach to target selection [30]. Other groups propose to focus on either hypothetical proteins in order to elucidate their functions [31,32], or organism-specific protein orphans [33]. We are integrating all these approaches by solving structures of all proteins in one complete genome. Small size of the genome allows us to optimize both computational techniques for target selection and experimental strategy. More importantly, the structural complement of the genome known as minimal set of genes required for life [34] will provide, to some extent, the completeness of structural studies.

5. ACKNOWLEDGEMENTS

The author thanks Prof. Sung-Hou Kim, Dr Chao Zhang and Hisao Yokota for very helpful discussions and reading the manuscript. Structural genomics project of *Mycoplasma* is conducted in the Berkeley Structural Genomics Center supported by NIH/NIGMS (#P50 GM62412). This work was supported by the Director, Office of Science, Office of Biological and Environmental Research under U.S. Department of Energy Contract No DE-AC03-76SF00098. The resources of the National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory, Berkeley, CA were used in this project.

6. REFERENCES

[1] Kyprides, N.C. (1999) *Bioinformatics* **15**, 773-774.

[2] Zarembinski, T.I., Hung, L.-W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R., Kim, S.-H. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 15189-15193.

[3] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) *Nucleic Acids Res.* **28**, 235–242

[4] Bowie, J.U., Luthy, R. & Eisenberg, D. (1991) *Science* **253**, 164-170.

[5] Fischer, D. & Eisenberg, D. (1996) *Protein Sci.* **5**, 947-955.

[6] Rychlewski, L., Zhang, B. & Godzik, A. (1998) *Fold. Des.* **3,** 229–238.

[7] Jones, D.T. (1999) *J. Mol. Biol.* **287,** 797–815.

[8] Frishman, D. & Mewes, H.-W. (1997) *Nat. Struct. Biol.* **4**: 626–628.

[9] Fischer, D. & Eisenberg, D. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 11929-11934.

[10] Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y.P. & Bork, P. (1998) *J. Mol. Biol*, **280**, 323-326.

[11] Gerstein, M. (1998) *Proteins* **33**, 518–534.

[12] Teichmann, S.A., Park, J. & Chothia, C. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14658-14663

[13] Wolf, Y.I., Brenner, S.E., Bash, P.A. & Koonin, E.V. (1999) *Genome Res*. **9**, 17-6.

[14] Murzin, A.G., Brenner, S.E., Hubbard, T.J.P. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536-540.

[15] Chothia, C. (1992) *Nature* **357,** 543–544.

[16] Orengo, C.A., Michie, A.D., Jones. S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997) *Structure*. **5**, 1093-1108.

[17] Kim, S.-H. (1999) *Curr. Opin. Struct. Biol.* **10**, 380-383.

[18] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., *et al*. (1995) *Science* **270**, 397-403.

[19] Smith, T.F. & Waterman, M.S. (1981) *J. Mol Biol.* **147**, 195-195.

[20] Pearson, W.R. & Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448.

[21] Altschul, S.F., Gish, W, Miller, Myers, E. W. & Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403-410.

[22] Grigoriev, I.V. & Kim, S.-H. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 14318-14323.

[23] Grigoriev, I.V., Zhang, C. & Kim, S.-H. (2000) *Prot. Eng. (submitted)*.

[24] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) *Nucleic Acids Res.* **25**, 3389-3402.

[25] Brenner, S.E., Koehl, P. & Levitt, M. (2000) *Nucleic Acids Res*. **28**, 254-256

[26] Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. & Frishman, D. (1999) *Nucleic Acids Res.* **27**, 44-48.

[27] Sanchez, R. & Sali, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13597–13602.

[28] Teichmann, S.A., Chothia, C. & Gerstein, M. (1999) *Curr. Opin. Struct. Biol.* **9**, 390-399

[29] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. (1999) *Nature.* **402,** 83-86.

[30] Cort, J.R., Koonin, E.V., Bash, P.A. & Kennedy, M.A. (1999) *Nucleic Acids Res.* **27,** 4018-4027.

[31] Balasubramanian, S., Schneider, T., Gerstein, M. & Regan, L. (2000) *Nucleic Acids Res.* **28**, 3075-3082.

[32] Eisenstein, E., Gilliland, G.L., Herzberg, O., Moult, J., Orban, J., Poljak, R.J., Banerjei, L., Richardson, D. & Howard, A.J. (2000) *Curr Opin Biotechnol.* **11,** 25-30

[33] Fischer, D. & Eisenberg, D. (1999) *Bioinformatics*. **15,** 759-762.

[34] Mushegian, A. (1999) *Curr. Opin. Genet. Dev*. **9**, 709-714.