Journal of Structural and Functional Genomics **6**: 63–70, 2005.
DOI 10.1007/s10969-005-2651-9

Structural genomics of minimal organisms and protein fold space

Sung-Hou Kim[1,2,*], Dong Hae Shin[2], Jinyu Liu[2], Vaheh Oganesyan[2], Shengfeng Chen[1], Qian Steven Xu[2], Jeong-Sun Kim[1], Debanu Das[2], Ursula Schulze-Gahmen[2], Stephen R. Holbrook[2], Elizabeth L. Holbrook[2], Bruno A. Martinez[2], Natalia Oganesyan[2], Andy DeGiovanni[2], Yun Lou[2], Marlene Henriquez[2], Candice Huang[2], Jaru Jancarik[1], Ramona Pufan[1], In-Geol Choi[1], John-Marc Chandonia[2], Jingtong Hou[2], Barbara Gold[2], Hisao Yokota[2], Steven E. Brenner[2], Paul D. Adams[2] & Rosalind Kim[2]

[1]*Department of Chemistry, University of California, Berkeley, California 94720-5230, USA;* [2]*Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; *Author for correspondence (e-mail: SHKim@cchem.berkeley.edu; Fax: +1-510-486-5272)*

Received 24 June 2004; accepted in revised form 15 February 2005

Key words: Berkeley Structural Genomics Center, minimal organisms, molecular function, protein fold space, structural genomics

Abstract

The initial aim of the Berkeley Structural Genomics Center is to obtain a near-complete structural complement of two minimal organisms, closely related pathogens *Mycoplasma genitalium* and *M. pneumoniae*. The former has fewer than 500 genes and the latter fewer than 700 genes. To achieve this goal, the current protein targets have been selected starting with those predicted to be most tractable and likely to yield new structural and functional information. During the past 3 years, the semi-automated structural genomics pipeline has been set up from cloning, expression, purification, and ultimately to structural determination. The results from the pipeline substantially increased the coverage of the protein fold space of *M. pneumoniae* and *M. genitalium*. Furthermore, about 1/2 of the structures of 'unique' protein sequences revealed new and novel folds, and over 2/3 of the structures of previously annotated 'hypothetical proteins' inferred their molecular functions.

The goal of obtaining protein structures on a genomic scale has motivated the development of high throughput technologies and protocols for macromolecular structure determination, and these technologies have begun to produce structures at a greater rate than previously possible [1, 2]. This structural genomics approach has also turned out to be a powerful method to infer the molecular functions of an increasing number of functionally unknown proteins, known as hypothetical proteins [1]. The initial objective of Berkeley Structural Genomics Center (BSGC) has focused on obtaining a near-complete structural complement of proteins of *Mycoplasma genitalium* (*MG*) and *Mycoplasma pneumoniae* (*MP*), two of the smallest pathogenic microbes (*MG* can be considered as a subset of *MP* in that the homologues of all *MG* genes are found in *MP*). This pilot project is designed to: (1) develop high throughput methods and protocols for all steps from cloning to structure determination; (2) identify the categories of proteins of different difficulties for structure determination and estimate the size of each category; (3) discover new protein folds; (4) discover molecular functions of hypothetical proteins; and (5) 'map' the protein fold space in terms of its distribution pattern and molecular functions. Such information may also provide a comprehensive view of the structural proteome of a small organism, which, in turn, can serve as a platform for understanding more complex organisms. We briefly summarize some methods and results in four different topics: (1) target selection, (2) semi-automation of cloning, purification and crystallization, (3) structure-based functional inference, and (4) a global mapping of the protein fold space of one organism, *MP*.

Target selection

In general, all rounds of target selection during the pilot period have involved multiple steps (Figure 1). In the first step, *MP* (and *MG*) proteins that have similar sequences to proteins of known structure are removed from further consideration. Our criteria for similarity varied between rounds of target selection, and will be described in detail in a future paper. In general, we used Pfam [3] and PSI-BLAST [4] with fairly permissive thresholds to remove potential targets. Next, proteins predicted to be experimentally intractable or difficult to do in a high throughput mode are put aside for the post-pilot period. These include proteins containing a region or regions of low-complexity, coiled coils, and transmembrane domains. The 'seg' program [5, version dated 5/24/2000] was run on all sequences to identify putative low complexity regions. The 'ccp' program [6, version dated 6/14/1998] was used to predict coiled coil regions in all sequences. To identify transmembrane regions, TMHMM 2.0a [7] and PHDhtm [8, version 2.1, dated 10/98] were used. Finally, specific targets were chosen amongst the remaining proteins and their homologues from other bacteria. The full length target sequences were cloned, even in cases where the homologues were longer than the original *Mycoplasma* protein.

Since we are only seeking to solve structures of proteins for which the structure cannot be reliably predicted *via* sequence comparison methods, it is necessary to frequently check whether structures that have similar sequences to our targets have been solved by others, those are then deleted from our target list. The target selection and de-selection procedures described above are time consuming to perform manually. Therefore, all steps from *MP*/*MG* target selection and identification of the target homologues in other organisms to target de-selection based on recent structures and their sequence homologues are automated (unpublished results). The automated procedures are designed for maximum sensitivity, but, since the automatic decisions can lead to false positives, the final decision on whether to stop any target is made manually.

Semi-automation of cloning, purification and crystallization

High-throughput methods to obtain well-expressing and highly soluble proteins have been

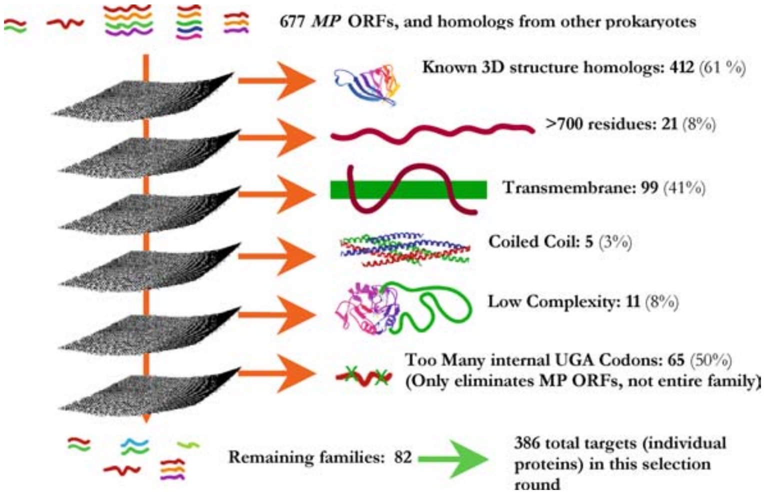
677 *MP* ORFs, and homologs from other prokaryotes
Known 3D structure homologs: 412 (61 %)
>700 residues: 21 (8%)
Transmembrane: 99 (41%)
Coiled Coil: 5 (3%)
Low Complexity: 11 (8%)
Too Many internal UGA Codons: 65 (50%)
(Only eliminates MP ORFs, not entire family)
Remaining families: 82
386 total targets (individual proteins) in this selection round

Figure 1. A simplified flowchart for target selection procedure. An example of a particular target selection round is presented. The number of ORFs eliminated by each filter is shown, and also expressed as a percentage of the number of targets entering the filter. The final filter for UGA codons eliminated only the *M. pneumoniae* ORFs but not their sequence homologues in other organisms. The target de-selection procedure performed weekly is not shown.

developed. The entire process of cloning and expression from polymerase chain reaction to mini-expression assay are subdivided into 9 steps, and each step is robotized for the Beckman Biomek 2000 (Beckman Coulter, Inc., Fullerton, CA). This cloning/expression scheme is based on T7 promoter-driven ligation independent cloning (LIC) vectors [9] we developed based on pET21a (Novagen, Madison, WI) (unpublished results). The steps are as follows: (1) PCR amplification, (2) PCR product analysis, (3) PCR purification, (4) PCR quantitation, (5) insert preparation, (6) LIC cloning reaction, (7) expression host (BL21(DE3)); transformation, (8) plasmid screening, and (9) mini-expression screening.

A rapid procedure to express recombinant proteins in an *E. coli* cell-free system using a 96-well format was also developed [10]. Since all our recombinant proteins harbor N-terminal His$_6$-tag or His$_6$-MBP-tag, the identification of soluble proteins is performed by the Dot Blot procedure using an anti-His tag antibody.

On-column refolding

A parallel process is used for protein purification. Installation of the AKTAexplorer (GE Healthcare, Piscataway, NJ), an automated protein purification system, minimizes preparation, running time, and repetitive manual tasks. It has the capacity to purify up to six different His$_6$- or His$_6$-MBP-tagged proteins per day and can produce mg amounts of protein for structural studies. However, insoluble expressed proteins cannot be purified by the above mentioned semi-automated procedures. To rescue these insoluble proteins we have developed an on-column chemical refolding method, which has achieved about 50% success rate in the production of refolded proteins out of tested insoluble targets [11]. Briefly, inclusion bodies solubilized in urea are first bound to a Ni-NTA column (Qiagen, Valencia, CA) and exposed to a detergent wash to prevent misfolding. This is followed by a β-cyclodextrin wash which removes detergent and promotes correct folding. The target protein is eluted with imidazole, goes through further purification steps (ion exchange and/or size exclusion chromatography), and is evaluated by dynamic light scattering (DLS) (DynaPro 99, Wyatt Technology Corp., Santa Barbara, CA) and mass spectrometry.

Optimum solubility (OS) screen

An initial crystallization screen is performed using the sparse matrix sampling method [12] with the Hydra-Plus One system (Matrix Technologies, Hudson, NH). This consists of 288 conditions, INDEX, SCREEN I & II, SALT (Hampton Research, Aliso Viejo, CA), and WIZARD (deCODE genetics, Bainbridge Island, WA) screens, at two temperatures: 4 °C and 22 °C. The sitting drop vapor diffusion method in Corning Crystal EX Conical Flat Bottom Plates (Corning Incorp., Corning, NY) is used for these screens.

An auto-imaging unit has been set up to handle a large number of crystallization plates (Discovery Partners International, San Diego, CA). Here again, a large fraction of purified proteins do not yield crystals. To improve the crystallization of these difficult proteins, we have developed a screen where a panel of buffers, pHs, and additives are tested in order to obtain the most suitable solution conditions that may favor crystallization [13]. After monitoring precipitation, the conditions leading to clear drops are selected for DLS characterization. The DLS results are used to select a new buffer for the protein sample before setting-up new screens. This method produced quite a number of target

Table 1. Progress summary: number of targets required to accomplish each stage in the BSGC experimental pipeline (as of June 14, 2004).

Experimental stage	Number of targets
Full length genes selected	423
Full length genes cloned	318
Expression tested	273
Soluble proteins	261
Insoluble proteins	12
Purified	191
Crystallized	84
Crystal structure	66*
NMR structure	3

This table contains all full-length targets selected using the target selection procedure described briefly in the paper. Of the 423 targets, 139 are *MP* genes, 33 are *MG* genes, and 251 are from other prokaryotes.

*These include 12 ligand-complexed structures for functional studies.

protein crystals from samples that had originally not been suitable for crystallization trials. Table I shows the results from BSGC from cloning of the unique genes (the genes with amino acid sequences that have no sequence homologues in the Protein Data Bank) to structures determined.

Structure-based functional inference in structural genomics

Structural genomics is emerging as a powerful approach not only to discover new protein folds, but also to annotate the function of hypothetical proteins. Structure-based inference for molecular function has been divided into five different categories [1]: I. 'Remote homologue' proteins; II. Proteins with unexpected bound ligands; III. Proteins in a 'twilight zone' of sequence and structural similarity; IV. Proteins with a new molecular function for a known cellular function; and V. Proteins with still unknown function. Here we present some examples to show the two most frequent categories where structure-based discovery of molecular function is possible. The efficiency by which function is deduced from a structure can be further improved by integrating other information from bioinformatics and experimental screening for enzymatic activity or ligand binding.

'Remote homologue' proteins

The most common way molecular function can be inferred from the structure of a hypothetical protein is when the structure turns out to be a remote homologue of one or more protein structures whose functions are known, i.e., the new structure is a structural homologue of one or more known structures with known functions despite the remoteness of its sequence similarity.

Methanococcus jannaschii MJ0936 (gi number 1499771), a sequence homologue of an *MP*/*MG* protein, was a hypothetical protein of unknown function with over 50 sequence homologues found in many bacteria and archaea. Its crystal structure, determined at 2.4 Å resolution (Figure 2), revealed structural homology to nucleases, phosphatases, or nucleotidases [14] with a Dali [15] Z-score higher than 6. A series of biochemical screens for catalytic activity was performed to test the biochemical activities suggested by the remote homologues. These assays revealed a novel phosphodiesterase activity with an absolute requirement for divalent metal ions, Ni^{2+} and Mn^{2+}. Thus, over 50 sequence homologues of this protein can be inferred to have a similar function.

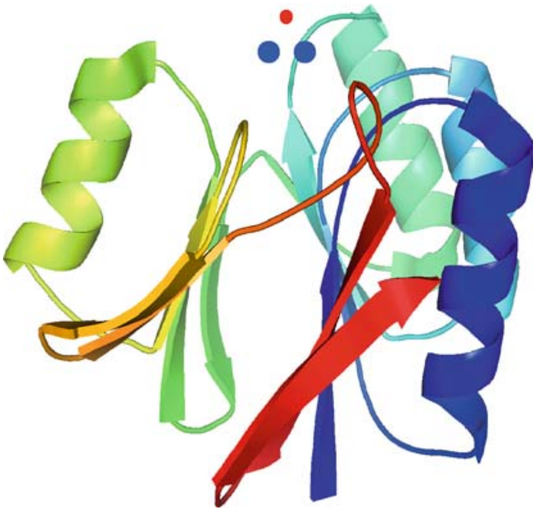

Figure 2. Overall structure of manganese-complexed MJ0936. Manganese ions are represented as blue spheres (PDB ID: 1S3N). A water molecule bridging the manganese ions in MJ0936 is represented as a small red sphere.

MG027 (gi 3844637) is one of the targets from *MG* which was also annotated as a conserved hypothetical protein. We have determined the crystal structure of the protein (Figure 3) and found that it is structurally homologous to the N-utilizing substance B protein (two members in the family) despite a low sequence identity between them [16]. The sequence alignment results also indicate that some highly conserved and functionally important residues in NusB are also well conserved in MG027, such as residues R14 and R18, so-called arginine-rich RNA-binding motif (ARM), which interact with the rRNA BoxA. Residue F26 involved in protein–protein or protein–RNA interactions is also conserved. Therefore, based on the structural and conserved sequence data between MG027 and NusB protein, we proposed that MG027 is a member of the NusB family despite the absence of sequence similarity.

SP_1288 (gi 15675166) from *Streptococcus pyogenes*, another sequence homologue of *MP*/

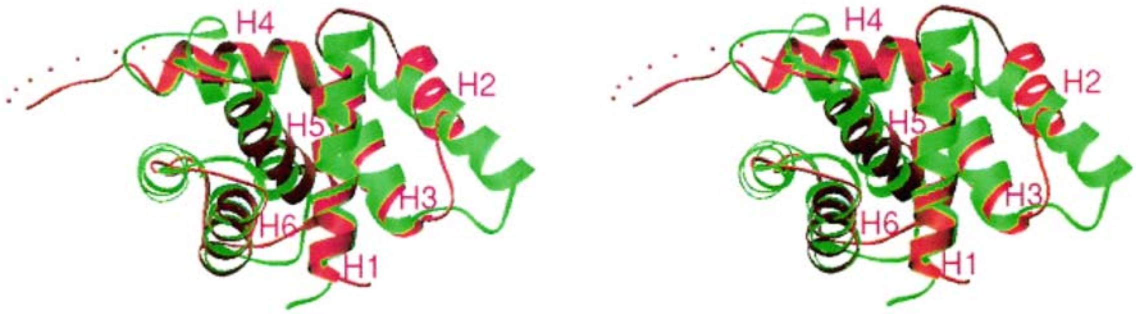

Figure 3. Stereoribbon diagrams of the Cα atoms superposition between MG027 and NusB from *Mycobacterium tuberculosis*. The MG027 structure was drawn in red (1Q8C) and the MT-NusB structure (1EYV) in green. H represents the α-helices.

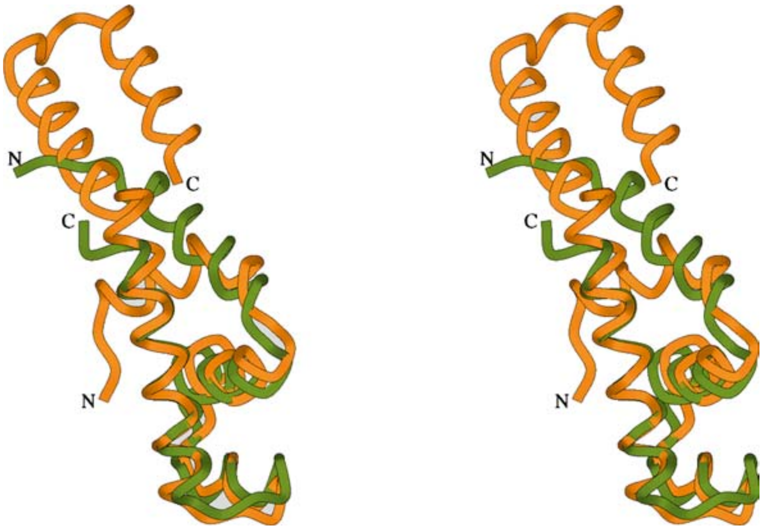

Figure 4. The SP_1288 structure superimposed with a σ factor, the closest structural homologue of known structure and function of SP_1288. Superimposition was performed with the *lsqkab* program from CCP4 suite using matrices that were suggested by the DALI search engine. The SP_1288 model (1S7O) is shown in dark orange. The C-terminal domain of alternative σ factor, σ^E (PDB ID: 1OR7, shown in olive) has the highest homology with SP_1288. The superimposition was done over Cα atoms. Fifty-nine corresponding residues from each structure yielded a Z-score of 7.8 and root-mean-square deviation of 2.4 Å.

MG and annotated as a putative DNA binding protein, is another example [17]. It belongs to the uncharacterized protein family UPF0122 (accession No. PF04297 in Pfam [3]) referred to as ‘putative helix-turn-helix proteins’ of which the genes from the members of this family are often part of operons that encode components of the signal recognition particle (SRP), which in turn is involved in translation. After the structure of SP_1288 was solved to 2.3 Å resolution (Figure 4), the structure homology search using Dali [15] revealed that 75% of the structure comprising the N-terminal 80 residues had good resemblance to domain 4 of RNA polymerase σ subunit (PDB accession codes 1or7, 1ku3 and ku7) with a Z-score of about 7.8. This suggests possible involvement of SP_1288 in the biochemical function of transcription initiation, which includes interaction with DNA. Thus, the function for all 26 members of UPF0122 can be inferred.

Proteins with unexpected bound ligands

The next frequent category covers the cases where the unexpected presence of a ligand in the structure of a hypothetical protein helps to infer its biochemical function. This was the case with protein TM1717 from *Thermotoga maritima* (gi 4982294), an *MP*/*MG* sequence homologue

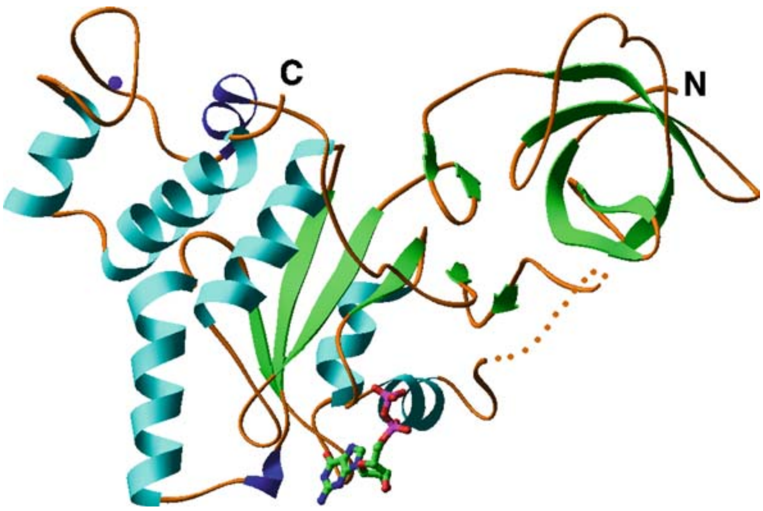

Figure 5. Crystal structure of TM1717 (1U0L). The structure of TM1717 is presented with bound GDP (ball and stick model) and zinc ion (purple).

and one of 83 members of a protein family. The crystal structure of this protein revealed that GDP was bound to the protein (Figure 5), immediately suggesting a possible role of the protein in GTP hydrolysis (unpublished results). Crystal structure and sequence analyses strongly indicated that TM1717 might be involved in translation because of the presence of an OB-fold domain known to bind to RNA and a zinc finger motif known to function in DNA recognition and RNA packing.

Another example is hypothetical protein AF2373 (gi 2650718) from *Archaeoglobus fulgidus*, another homologue of an *MP/MG* protein. The crystal structure solved to 2.5 Å resolution (Figure 6) revealed a bound NADP near three conserved motifs, GXXG, GGDGXXT, and TXXGSTXY(X)$_4$GG (unpublished results). Based on the crystal structure and sequence analyses, subsequent biochemical assays showed that AF2373 had an ATP-NAD kinase activity. Therefore, the functions for all 148 members of this family can be inferred.

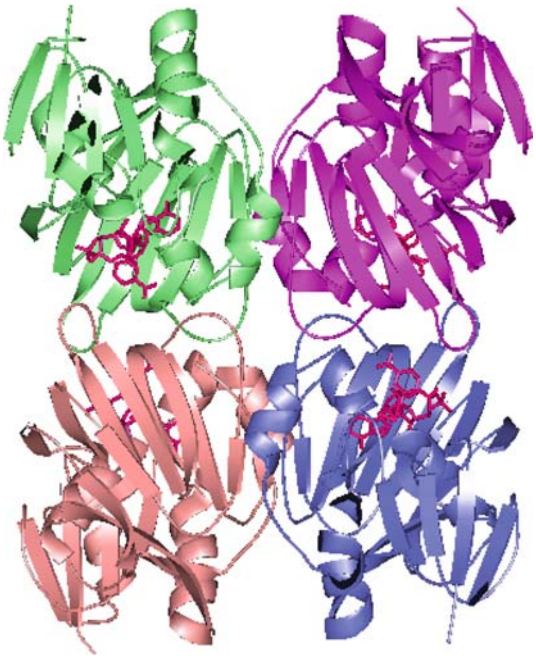

Figure 6. Crystal structure of tetrameric AF2372 (1SUW). AF2372 with bound NADP (red) is shown.

A global map of the protein structure universe

One of the principal goals of the structural genomics initiative is to identify the total repertoire of protein folds and obtain a global view of the 'protein structure universe' [18]. It is estimated that there are more than 10 million species of living organisms on earth and as many as a trillion different proteins among them. However, most of the proteins are composed of one or more structural domains (architectural units, also called 'folds'), and the diversity of protein structures arises from the combinatorial assembly and variation of a much smaller number of unique protein structural domains or protein folds.

We have developed a way to represent a three-dimensional map showing the distribution of the folds in the protein structure universe in which structurally related folds are represented by spatially adjacent points. In Figure 7, a map constructed using 1898 non-redundant protein structures reveals a highly non-uniform distribution of protein folds in the structure space and a segregation of four 'classes' of protein folds (α, β, α/β, and $\alpha + \beta$) into four elongated regions. Such a representation reveals a high level of organization of the protein structure universe that is intuitively interpretable in terms of the demography of protein fold type, the structural relationship among different proteins, and the evolution of protein structures. The *M. pneumoniae* proteome mapped on the protein universe space shows a high proportion of α/β class proteins (Figure 7).

In summary, we have learned:

1. Many processes from gene to structure can be automated and parallelized to achieve the goal of structural genomics at BSGC, but there also are many steps that are difficult to automate. In addition, a large fraction of proteins require additional development of methods and technology to obtain their structures;
2. 'Low-lying fruit' proteins (proteins whose structures can be obtained 'easily' by a single path from a large number of clones of a whole or subset of one or more proteomes) are only a few percent of the clones. The rest, over 90% of proteins, require multiple paths as well as new technologies, protocols, and/or methods to obtain their structures;
3. More than 2/3 of the structures of hypothetical proteins are remote homologues of pro-

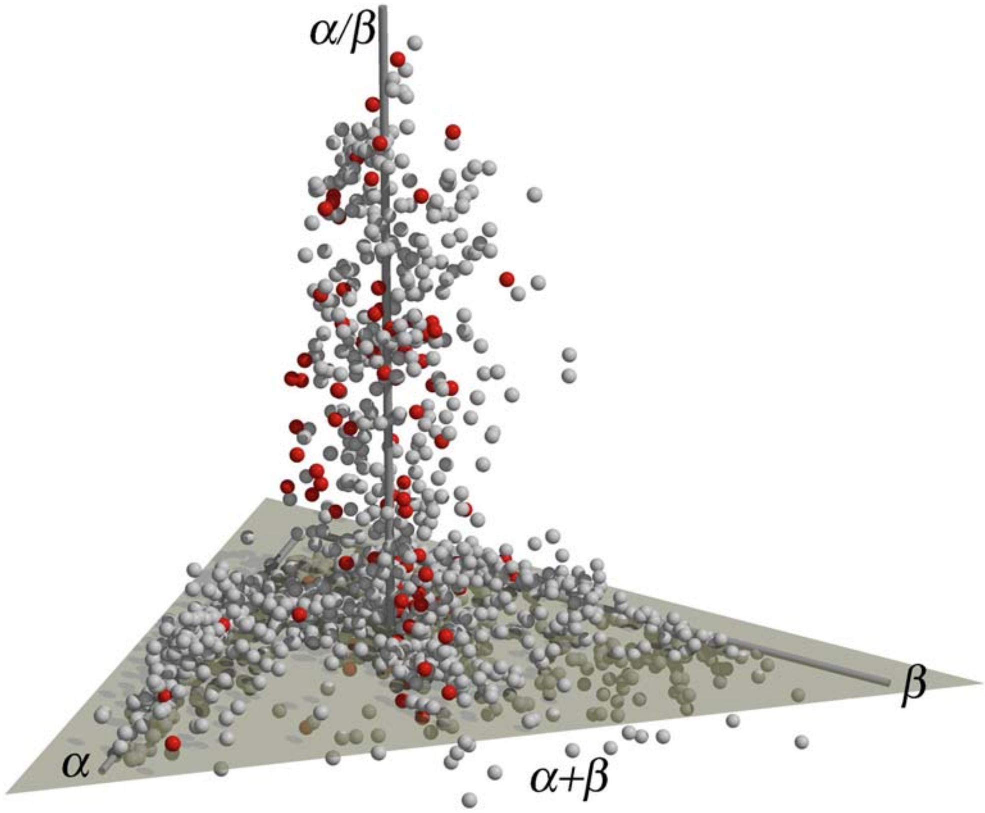

Figure 7. Known protein folds of *M. pneumoniae* in the protein structure space. To build the protein structure space (grey spheres plus red spheres), altogether 1898 non-redundant protein structures representing all PDB structures were selected. All pair-wise structural similarities were calculated and subsequently converted into dissimilarity scores (distances). The distances were then projected into three dimensions by using a multi-dimensional scaling procedure. Protein structures from α, β, α/β, and $\alpha + \beta$ classes are distributed around four axes, with the α, β and α/β axes shown in the map. Structures of *M. pneumoniae* (red spheres) or their homologues were found to be located mostly in the α/β region.

teins of known structures and functions, thus, structural genomics can provide a unique contribution in annotating molecular functions of many hypothetical proteins; and

4. About ~1/2 of proteins with no sequence homologues among the proteins of known structure reveal new folds.

It is also clear that the technologies and methods developed by the Protein Structure Initiative (http://www.nigms.nih.gov/psi) will have an important impact in the life sciences in general, particularly in structural biology.

Acknowledgements

We thank all the component members of BSGC for their efforts towards accomplishing the BSGC objectives. We gratefully acknowledge the support of NIH grant GM62412 for the structures cited in this article.

References

1. Kim, S.-H., Shin, D.H., Choi, I.G., Schulze-Gahmen, U., Chen, S. and Kim, R. (2003) *J. Struct. Funct. Genomics* **4**, 129–135.
2. Adams, M., Joachimiak, A., Kim, R., Montelione, G.T. and Norvell, J. (2004) *J. Struct. Funct. Genomics* **5**, 1–2.
3. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) *Nucleic Acid Res.* **28**, 263–266.
4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
5. Wootton, J.C. (1994) *Comput. Chem.* **18**, 269–285.
6. Lupas, A. (1996) *Methods Enzymol.* **266**, 513–525.
7. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) *J. Mol. Biol.* **305**, 567–580.
8. Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) *Protein Sci.* **4**, 521–533.
9. Aslanidis, C. and de Jong, P.J. (1990) *Nucleic Acids Res.* **20**, 6069–6074.
10. Busso, D., Kim, R. and Kim, S.-H. (2003) *J. Biochem. Biophys. Methods* **55**, 233–240.
11. Oganesyan, N., Kim, S.-H. and Kim, R. (2004) *PharmaGenomics* **7**, 22–26.
12. Jancarik, J. and Kim, S.-H. (1991) *J. Appl. Crystallogr.* **24**, 409–411.
13. Jancarik, J., Pufan, R., Hong, C., Kim, S.-H. and Kim, R. (2004) *Acta Crystallogr. D* **60**, 1670–1673.
14. Chen, S., Yakunin, A.F., Kuznetsova, E., Busso, D., Pufan, R., Proudfoot, M., Kim, R. and Kim, S.-H. (2004) *J. Biol. Chem.* **279**, 31854–31862.
15. Holm, L. and Sander, C. (1995) *Trends Biochem. Sci.* **20**, 478–480.
16. Liu, J., Yokota, H., Kim, R. and Kim, S.-H. (2004) *Proteins* **55**, 1082–1086.
17. Oganesyan, V., Pufan, R., DeGiovanni, A., Yokota, H., Kim, R. and Kim, S.-H. (2004) *Acta Crystallogr. D* **60**, 1266–1271.
18. Hou, J., Sims, G.E., Zhang, C. and Kim, S.-H. (2003) *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2386–2390.