The Status of Structural Genomics Defined Through the Analysis of Current Targets and Structures

P.E. Bourne, C.K.J. Allerston, W. Krebs, W. Li, I.N. Shindyalov, A. Godzik, I. Friedberg, T. Liu, D. Wild, S. Hwang, Z. Ghahramani, L. Chen, and J. Westbrook

THE STATUS OF STRUCTURAL GENOMICS DEFINED THROUGH THE ANALYSIS OF CURRENT TARGETS AND STRUCTURES

P. E. BOURNE, C. K. J. ALLERSTON, W. KREBS, W. LI, and I. N. SHINDYALOV

The San Diego Supercomputer Center, The University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA

A. GODZIK, I. FRIEDBERG, and T. LIU

The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037 USA

D. WILD and S. HWANG

The Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711 USA

Z. GHAHRAMANI

Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London, WC1N 3AR, UK

L. CHEN and J. WESTBROOK

Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854 USA

Structural genomics – large-scale macromolecular 3-dimenional structure determination – is unique in that major participants report scientific progress on a weekly basis. The target database (TargetDB) maintained by the Protein Data Bank (http://targetdb.pdb.org) reports this progress through the status of each protein sequence (target) under consideration by the major structural genomics centers worldwide. Hence, TargetDB provides a unique opportunity to analyze the potential impact that this major initiative provides to scientists interested in the sequence-structure-function-disease paradigm. Here we report such an analysis with a focus on: (i) temporal characteristics - how is the project doing and what can we expect in the future? (ii) target characteristics - what are the predicted functions of the proteins targeted by structural genomics and how biased is the target set when compared to the PDB and to predictions across complete genomes? (iii) structures solved – what are the characteristics of structures solved thus far and what do they contribute? The analysis required a more extensive database of structure predictions using different methods integrated with data from other sources. This database, associated tools and related data sources are available from http://spam.sdsc.edu.

1 Introduction

Structural genomics has been heralded as the follow on to the human genome project. This is interpreted to mean a large-scale project, with scientific, engineering and technological components and with the potential to have a large impact on the life sciences. Whereas the goals of the human genome project were

relatively well defined – sequence the 3 billion nucleotides comprising the human genome and define all open reading frames – the goals advanced for structural genomics are more diverse (`http://www.nigms.nih.gov/news/meetings/-hinxton.html/`) [1]. For instance, some of the NIH P50 structural genomics centers have focused on all of the protein structures in a given genome – *A. thaliana, T. maritima* and *M. tuberculosis,* are examples under scrutiny. Other groups have focused on obtaining sufficient coverage of fold space [2] to facilitate accurate homology modeling of the majority of proteins of biological interest (see http://spam.sdsc.edu/sgtdb for a description of the focus of each center). Since structure has already taught us so much about biological function when undertaken as a functionally driven initiative, undertaking structure determination in a broader genomic sense will likely also bring significant new understanding of living systems. Further, it will likely lead to advances in the process of structure determination, whether by X-ray crystallography or NMR. With such diversity of deliverables and with some projects now well established, an obvious question is, how are we doing? This paper addresses this question.

The question has been addressed before in the context of new folds and functions and has proven has to be a somewhat controversial. An initial report in Science [3] implied that the number of structures produced as of November 2002 was minimal. A response from the US Northeast Structural Genomics Consortium (NESG) [4] indicated it was early in the process and that indeed that the absolute number of structures produced may not be the best measure, but rather the value of those structures is more to the point. NESG indicated that a structure containing a novel fold would indeed provide a new template from which many sequences could be related and hence was a significant contribution. It is not our intent here to join this argument but to simply point readers at some quantitative data and suggest how the process might proceed in the future and the challenges it provides to the bioinformatics community.

2 Methods

An important feature of structural genomics, laid out by the NIH as part of the awards made to the pilot centers engaged in this high throughput structure determination, was the importance of reporting their progress on a regular basis. The 16 pilot centers in the US and worldwide do this by way of weekly updates made available through their individual centers and collated by the Protein Data Bank (PDB) into what is known as the target database (TargetDB; http://targetdb.pdb.org) [5]. The contents of the target database are also available as an XML file. This file was used to create a local database from which the results presented here are derived. This database is available at http://spam.sdsc.edu/sgtdb.

Fold prediction is based on three existing methodologies, FFAS [6] iGAP [7] and Bayesian networks [8] which are fully described elsewhere. Prediction of all open reading frames from complete proteomes uses the iGAP methodology and is part of the Encyclopedia of Life (EOL; http://eol.sdsc.edu) project.

3 Results

3.1 Progress

In the past year (May 1, 2002 – May 31, 2003) 314 structures resulting from structure genomics were reported by TargetDB. During the same period, a total of 3324 structures were deposited with the PDB. Thus structure genomics is currently contributing approximately 10% of structures to the field of structural biology. The number of structures at each stage in the pipeline is shown in figure 1.
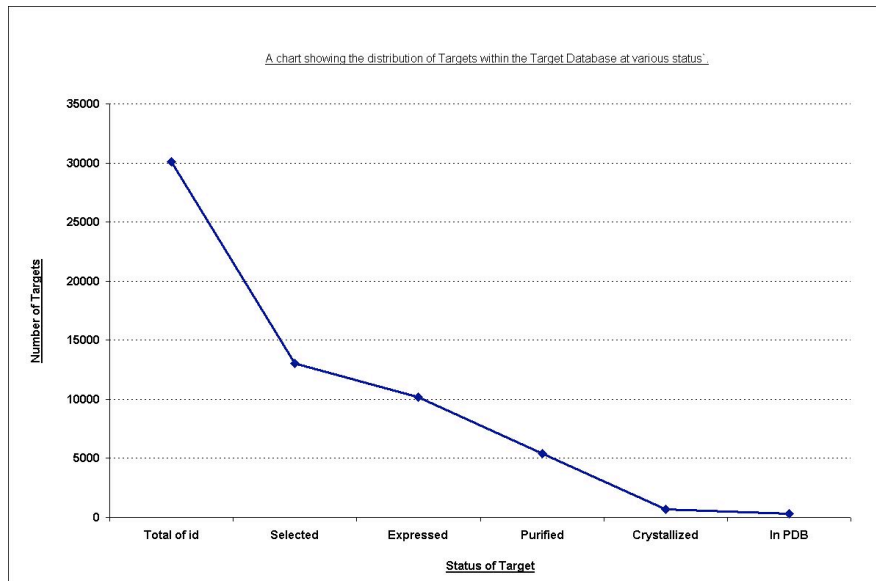

Figure 1 Structural Genomics Targets at Different Stages of Solution (April 1, 2003)

Slightly less than 50% of targets are selected for scrutiny. From these a high percentage can be expressed, but the number purified and crystallized drops off dramatically, indicating these steps continue to register low success rates and should be a focus of renewed efforts. Of those that crystallize, the majority find their way into the PDB.

Is the percentage of structures determined by structural genomics likely to increase in the near future? To address this question requires that we look for temporal trends in the data. This is possible since TargetDB is updated each week and the mean time that an active target spends at each step in the structure determination pipeline can be assessed. These results are shown in Figure 2. It

should be noted that not all of the centers reporting weekly status update their internal status tracking data with the same frequency. Consequently, the interval assessment here must be interpreted with care.
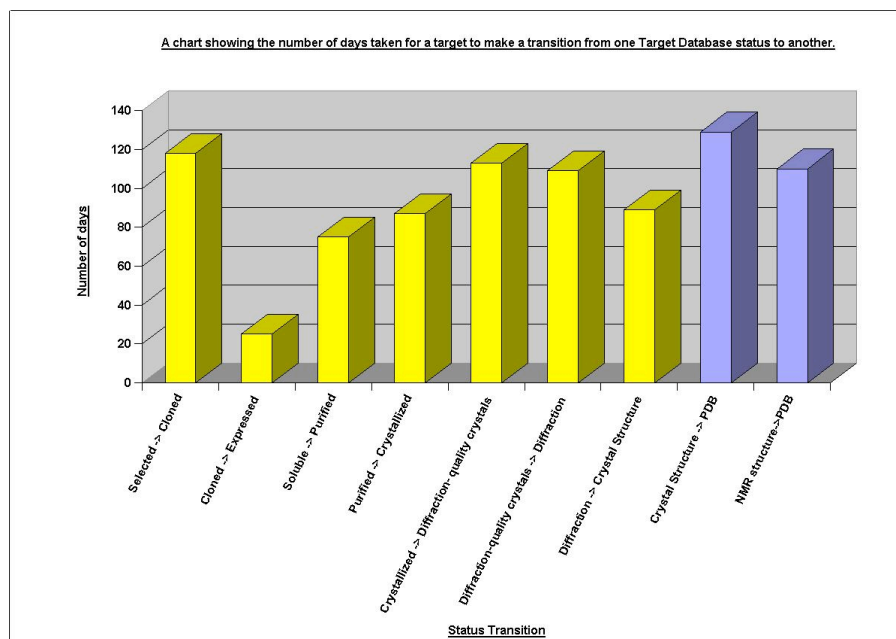

Figure 2 Mean Time of Targets at Each Structure Determination Step

For targets that make it to the next step, the data indicates that there is no specific bottleneck at this point, but rather a balance between the time taken at each structure determination step. Without a significant bottleneck the prospects for improving the rate of structure determination would seem good, particularly as the early stages of the project have included a significant engineering component for some projects. However, a final answer to the question will come from further review of TargetDB in the next two years.

3.2 Target characteristics

The characteristics of targets being attempted by individual structural genomics groups are highly variable (see http://spam.sdsc.edu/sgtdb for a synopsis of the activities of each individual group). Groups are focusing on one or more of the following: complete proteomes, pathways and diseases, new folds, new technologies and specific structures. Thus the relative number of active targets from

each group is meaningless and no attempt is made here to compare groups, rather the characteristics of the targets as a whole is considered.

A review of the over 30,000 targets in the database (April 1, 2003) indicates a 13% redundancy at the 100% sequence identity and 38% redundancy at the 30% sequence identity level. This implies that either individual groups are operating without regard for other groups, or there is interest in the same targets by different groups perhaps indicating some important functional significance for a particular target. This data could be probed further to ascertain (if possible from sequence alone) the functional significance of these hotly contested targets. It should be noted that there is a temporal aspect to these target data. When a target was selected, which may be up to three years ago, the level of redundancy with respect to NR may have been significantly different, so these data need to be interpreted with care

A review of each groups targets indicates that there is a significant level of redundancy within a groups targets (Figure 3). In some cases this is the nature of the redundancy in the complete proteome under study, in other cases perhaps a desire to attempt to solve multiple instances of an important structure that, based of sequence identity, are known to have the same fold.
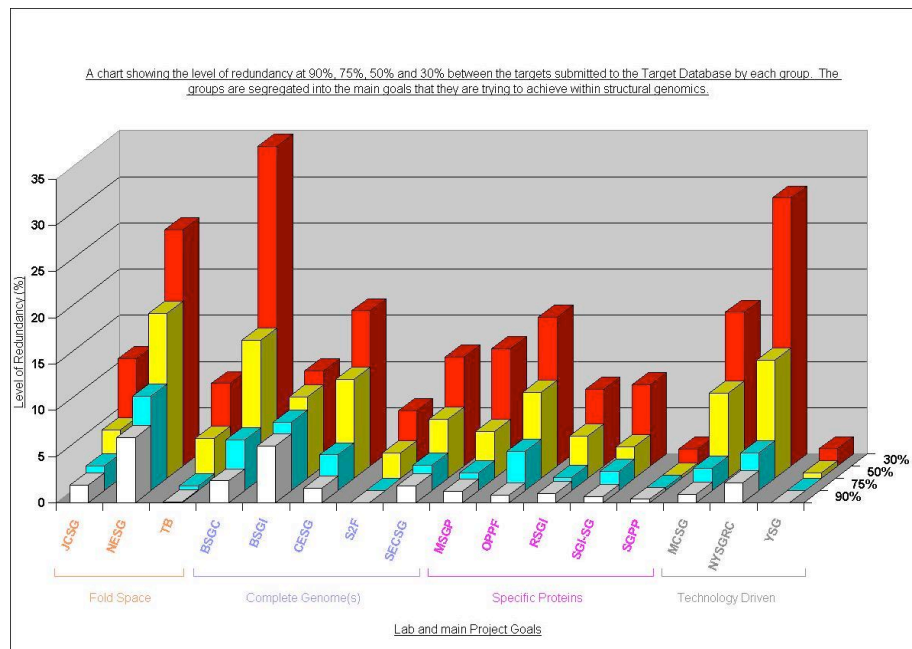

Figure 3 Sequence Redundancy within each Groups Targets

3.3 Structure characteristics

Are there any specific characteristics of the novel folds in the structures determined by the Structural Genomics Initiative? How do these differ from the general population in the PDB and why? In short, what is novel from the structures being determined by structural genomics and how do they aid us by increasing our understanding of living systems and/or aid more rapid structure determination or modeling? An analysis of the former is provided by [9]. Here we focus on the characteristics important to bioinformatics, specifically fold and function, which can be used in further analysis, for example, in homology modeling.

An analysis of the new folds as defined by SCOP is given in Table 1.

Table 1 New Folds Resulting from Structural Genomics

Period	Total New Folds	New Folds from Structure Genomics
Oct 2001 - Mar 2002	48	1. YchN-like (c.144) 2. Hypothetical Protein MTH777 (c.115) 3. alpha/beta knot (c.116) 4. Archaeosine tRNA-guanine transglycosylase, C-terminal additional domains (e.36) 5. YebC-like (e.39)
Apr 2002 - Sep 2002	27	1. DsrC, the gamma subunit of dissimilatory sulfite reductase (d.203) 2. Ribosome binding protein Y (d.204) 3. Hypothetical protein MTH637 (d.206) 4. Thymidylate synthase-completmenting protein Thy1 (d.207) 5. MTH1598-like (d.208)
Oct 2002 - Mar 2003	64	1. S13-like H2TH domain (a.156) 2. C-terminal domain of DFF45/ICAD (a.164) 3. BEACH domain (a.169) 4. Viral chemokine binding protein m3 (b.116) 5. Obg-fold (b.117) 6. N-terminal domain of MutM-like DNA repair proteins (b.113) 7. Pututive glycerate kinase (c.118) 8. DegV-like (c.119) 9. YbaB-like (d.222) 10. SufE (d.224) 11. Replication modualtor SeqA, C-terminal DNA-binding domain (d.228)

In the first reporting period the number of new folds reported by structural genomics was approximately 10% of the total number reported (5 out of 48), a result proportional to the percentage of structures coming from structural genomics. In the second and third periods this jumped to 18% (5 out of 27) and 17% (11 out of 64), respectively indicating that the goal of new fold discovery may be being met, given that only 10% of structures overall are coming from structural genomics. However, the sample of new folds is small and hence we will need to wait for additional time periods and review this trend again.

A review of the sequences of solved structures against the non-redundant protein sequence database (NR) ordered in bins of expectation value (E-value) is given in Figure 3.
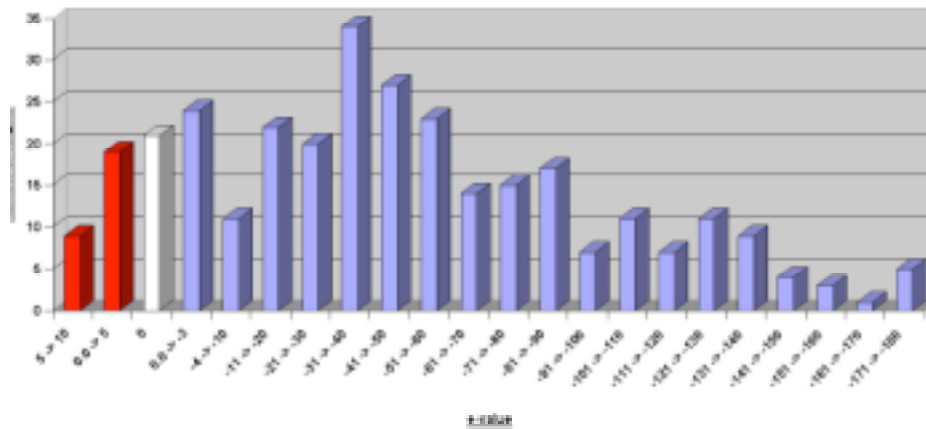
Figure 4 Likely Uniqueness of New Targets

Approximately 70 of a total of 314 structures have an E-value of 10-3 or higher and represent a group for which sequence homology is not guaranteed and hence represent possible new functions (assuming functions were correctly assigned to sequences in NR). Again the above is only an indicator of the situation. A better analysis would require comparison against NR at the time the structure was solved or released.

What of the overall distribution of folds represented by TargetDB? Figure 5 shows the distribution of folds derived by FFAS [6], iGAP [7] and Bayesian networks [8]. The level of reliability is not considered, only possible predictions are represented, both FFAS and iGAP provided predictions for the nearly all targets, Bayesian networks for about 10%, based on a smaller template library. Not only does this highlight internal consistency between the methods of prediction, it also indicates differences. The distribution of major folds seems consistent with the distribution of associated biological functions in living systems. For example, it is known that p-loop containing protein families are very prevalent in nature.
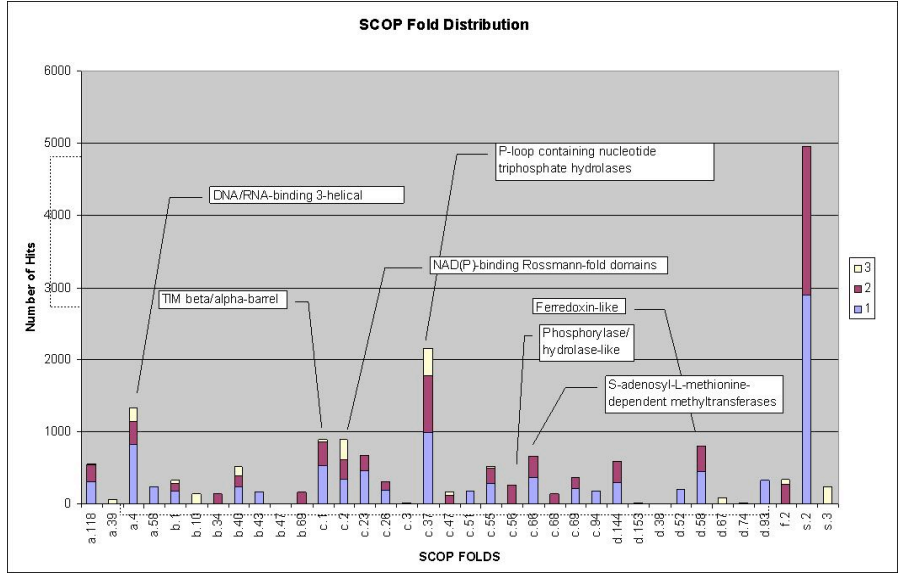
Figure 5 Predicted Folds from TargetDB: 1=FFAS; 2=iGAP; 3=Bayesian Networks

This relationship is probed further in figure 6. Fold predictions are made for all open reading frames is a variety of organisms as well as the PDB and TargetDB.
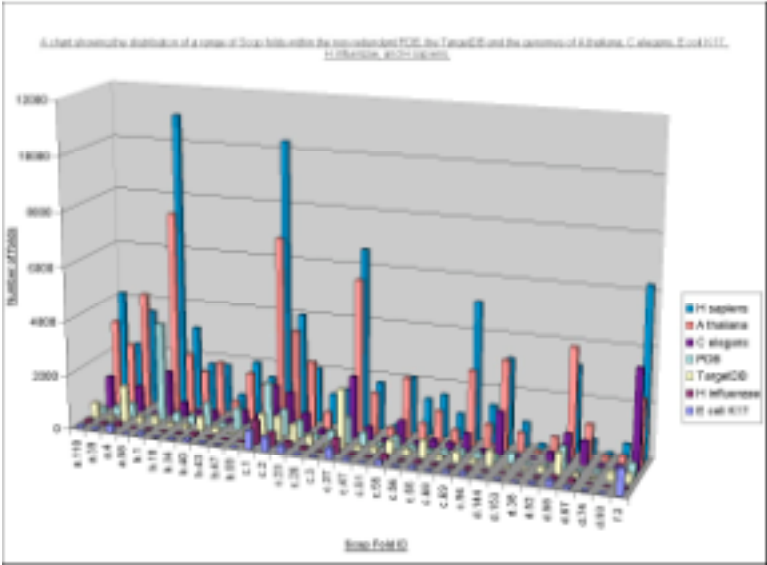
Figure 6 SCOP Fold Distributions in Several Model Organisms, PDB and TargetDB

A question that can be posed from these data is how biased are the distributions of folds in TargetDB relative to those from specific target organisms and the PDB? Intuitively one would expect the PDB to be biased towards proteins that are a) likely to be crystallized easily b) smaller proteins amenable to NMR or c) over represented by particular classes of proteins since they represent drug targets or functionally important proteins. Conversely, TargetDB would be somewhat closer to what is found in nature as whole genomes are being attempted. Having said that, it may be at this stage of structural genomics that projects are going for the low hanging fruit and hence it may be too early to make such a comparison.

It should also be noted that there is an undetermined bias in these data and hence they should be considered cautiously. The bias arises in that predictions are done with a mix of fold prediction and homology modeling. In both cases there is a bias towards known folds since, nevertheless expected trends do occur.

Immunoglobulin-like beta sandwiches (b1) are over represented in the PDB and under represented in TargetDB. This would suggest they have proven particularly amenable to crystallization and represent a sequence rich fold class which recognizes many of the targets and if new folds is an aim will likely discount a large number of targets, hence the under representation from TargetDB. The same argument can be made for tim barrels (c1). The empirical rule that emerges from these and other fold classes is that a class that is over represented in the PDB is under represented in TargetDB.

RNA/DNA binding 3 helical bundles (a4) appear to be over represented in TargetDB relative to what appears in the PDB and several model organisms. The same is true of P-loop containing nucleotide triphosphate hydrolases, perhaps a reflection of their role as drug targets. S-adenosyl-L-methionine-dependent methyltransferases also appear over represented in TargetDB.

4 Discussion

Structural genomics is a large science project involving multidisciplinary teams seeking to increase the number of macromolecular structures. From this process comes new understanding of living systems derived from functional inference from structure and improved methodologies. Improved methodologies range from new engineering practices which speed the structure determination process to an increased number of known folds that improves our ability to provide realistic models of proteins of unknown structure.

A unique aspect of structural genomics is a weekly report by all groups engaged in this activity. Thus for the first time we are in a position to monitor quantitatively the scientific progress of a major scientific project. This progress is in the form of the status in the structure determination process of protein sequence targets. This status terminates at the point the structure enters the PDB and hence structures completed by structural genomics can be compared against structures

derived from conventional functionally driven structure determination experiments. Targets which have not yet been solved can be predicted with a variety of existing structure prediction methods. Taking existing unsolved targets, solved structures and predicted structures of the targets a picture of the progress of structural genomics begins to emerge. Here we have reported on that picture.

The percent of structures being contributed by structural genomics is approximately 10% at this time. The time to solution ranges from three to eighteen months with a peak in the 8-10 month range (data not shown). Data are not available for how this compares to conventional structure determination but it is estimated to be of a similar order.

At this time structural genomics would seem to be contributing twice the number of new folds as conventional structure determination, but the numbers are two small to be considered statistically significant. An argument has been made that structure genomics might contribute less new folds that one might anticipate since the emphasis will be on determining the maximum number of structures. Numbers implies taking what crystallizes easily and this could be construed as being those structures that appear in a subset of folds most amenable to crystallization. Conversely, a functionally driven initiative on a single target might expend more time and energy performing experiments that would result in the crystallization of a less amenable fold not pursued by structural genomics. This type of conjecture will become more fact as the number of structures increases. We will continue to process TargetDB and report our finding through the Web site at http://spam.sdsc.edu/sgtdb.

Acknowledgments

References

1. S.E. Brenner SE, and M. Levitt. Expectations from Structural Genomics *Protein Sci* **9(1),** 197 (2000).
2. E. Portugaly and M. Linial. Estimating the Probability for a Protein to have a New Fold: A Statistical Computational Model. *Proc Natl Acad Sci U S A*. **97(10),** 5161 (2000)
3. R.F Service Tapping DNA for Structures Produces a Trickle. *Science* **298**, 948 (2002).
4. M. Gerstein *et al*. Structural Genomics: Current Progress. *Science* **299**, 1663 (2003).
5. J. Westbrook J. *et al*. The Protein Data Bank and Structural Genomics. *Nucleic Acids Research* **31(1)** 489 (2003).
6. L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik, A. Comparison of Sequence profiles. Strategies for Structural Predictions using Sequence Information. *Protein Science* **9** 232 (2000).

7. W.W. Li, G.B. Quinn, N.N. Alexandrov, P.E. Bourne and I.N. Shindyalov Proteins of Arabidopsis (PAT) database: A Resource for Comparative Proteomics. *Genome Biology* In Press (2003).
8. A. Raval, Z. Ghahramani and D.L. Wild A Bayesian Network Model for Protein Fold and Remote Homologue Recognition. *Bioinformatics* **18(6)** 788 (2002).
9. C. Zhang and S-H Kim Overview of Structural Genomics: From Structure to Function. *Current Opinions in Chemical Biology* **7** 28 (2003).