

# Characterization of Protein Structure and Function at Genome Scale Using a Computational Prediction Pipeline

Dong Xu<sup>1\*</sup>, Dongsup Kim<sup>1</sup>, Phuongan Dam<sup>1</sup>,  
Manesh Shah<sup>1</sup>, Edward C. Uberbacher<sup>1</sup>, and Ying Xu<sup>1,2</sup>

<sup>1</sup>Life Sciences Division, and <sup>2</sup>Computer Sciences and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

**Running title:** Studying Proteins Using a Computational Pipeline.

**Keywords:** protein structure prediction; fold recognition; threading; genome annotation; structure-function relationship; hypothetical protein; *cyanobacteria*; carboxysome.

---

\*Correspondence to Dong Xu, Protein Informatics Group, 1060 Commerce Park Drive, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6480. Tel: 865-574-8934. Fax: 865-241-1965. Email: [xud@ornl.gov](mailto:xud@ornl.gov).

# 1. Introduction

Recent advances in high-throughput production capabilities for biological data such as genomic sequence (Lander et al., 2001; Venter et al., 2001), large-scale gene expression data (DeRisi et al., 1997, Chu et al., 1997, Zhu et al., 2000), genome-scale protein-protein interactions (Fields & Song, 1989; Ho *et al.*, 2000), and protein structures (Chance et al., 2002), are revolutionizing the biological sciences. Essential to this new revolution are capabilities to computationally interpret large quantities of biological data generated under various experimental conditions and build mathematical models that fit these data. The combination of on-line bioinformatics tools and easy access to the high-speed Internet has made it generally possible to facilitate such computational steps and make biological discoveries *in silico* in a highly efficient manner. By utilizing various bioinformatics prediction, analysis and modeling tools, one can quickly generate hypotheses and theoretical models, which could then guide the design of experiments for further validation. The paradigm that links and integrates systematic data generation, computational data interpretation, and experimental validation is clearly providing a new and powerful way for conducting biological research. The focus of this paper is on (a) development of new computational tools for interpretation of large quantity of genomic sequence data for structural and functional inference and (b) example applications of these tools to studies of microbial genomes, particularly *cyanobacterial* genomes.

One of the key goals in bioinformatics in the post-genome era is to systematically derive functional information for the gene products (usually proteins) generated by the large-scale genome sequencing efforts. One of the popular approaches for achieving this is by recognition of homology using sequence comparison tools like BLAST and PSI-BLAST (Altschul et al., 1997). Though highly effective, the limitation of such an approach is also clear. The general observation has been that about 30-40% of genes in a newly sequenced genome cannot be detected to be significantly similar to proteins with known cellular roles or molecular functions. These unknown proteins may fall outside the limit of the current sequence-based techniques for homology detection. A more general class of computational methods for functionally characterizing unknown proteins is through prediction of three-dimensional (3D) structure. Existing prediction methods for protein structure have matured to a level such that useful information can be extracted about function, as demonstrated in the recent CASP contests (Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) (CASP, 1993, 1995, 1997, 2001, 2002). In many cases, the predicted protein 3D structures can reach the accuracy level better than 4 Å root mean square deviation (RMSD), which provides not only direct functional information about the proteins under study, but also highly useful guidance to experimentalists designing experiments for further investigation of protein

function. Structure-based functional inference provides a more general class of tools for functional characterization of proteins, as they use more information than sequence-based approaches. Even when a protein can be characterized through sequence-based comparison methods, a predicted structure can clearly provide additional information about the biochemical mechanism of the protein at atomic detail, as demonstrated in a large number of real life applications (Xu et al., 2001b).

Existing protein structure prediction methods fall into two main classes: (a) comparative modeling methods give predictions based on identified sequence-structure relationship with known protein structures (Sali and Blundell, 1993; Bowie et al, 1991); (b) *ab initio* methods (Li and Scheraga, 1987; Skolnick and Kolinski, 1991) give predictions directly from a protein sequence without using structure templates. Comparative modeling methods, when applicable, are generally faster and more accurate than *ab initio* methods. In particular, fully automated comparative modeling using automated computer servers is approaching the performance level of computer-assisted manual predictions on some classes of proteins, as demonstrated in CASP5 (CASP, 2002). Even in cases where the predicted structure may not be very accurate due to poor alignment, the established evolutionary relationship between the query protein and a protein with known structure can provide useful functional information. This is one of the advantages that comparative modeling has over *ab initio* methods, where such relationships are difficult to achieve. As more protein structures are experimentally solved, comparative-modeling methods will clearly become more applicable. Statistics from the PDB Web site (<http://www.rcsb.org/pdb/holdings.html>) show that about 90% of the proteins submitted to the PDB database (Westbrook et al., 2000) during 1997-2002 share similar folds to structures already in PDB. This suggests that these protein structures are potentially solvable by comparative modeling methods. Note that this does not necessarily indicate that comparative modeling may apply to 90% of all proteins, as the sampling (of protein structures) from the space of all proteins is certainly biased. In particular, membrane proteins are clearly under-represented in PDB. Nevertheless, it is generally believed that 60-70% of new proteins are potentially solvable using comparative modeling methods (Montelione and Anderson, 1999).

Comparative modeling can be generally divided into two classes of approaches: (1) sequence-sequence comparison-based approach (Karplus et al., 1998), and (2) sequence-structure comparison-based approach (threading) (Bowie et al., 1991; Jones et al., 1992; Xu and Xu, 2000). Threading makes a structural fold prediction from an amino-acid sequence by recognizing a structural template that represents the native-like fold of a query protein in a database of experimentally determined structures. Technically, fold recognition is achieved by finding an optimal alignment of residues of the query with residue positions of each structural template in the database, and by identifying those sequence-structure alignments that are statistically

significant. For each such alignment, the residues of the query sequence are predicted to have the coordinates of the aligned backbone positions in the template structure. Since protein threading uses structural information as well as sequence-based information, it is generally more effective than sequence-sequence comparison-based methods for identification of native-like folds.

Protein structure prediction is a multi-faceted and complex process with multiple steps. It generally involves several tools in addition to the tool used for building the three-dimensional model of the structure. Different classes of proteins, e.g., soluble *versus* membrane-associated proteins, may require different computational techniques for their structure predictions, due to their different physicochemical or other properties. A protein can have multiple structural domains. Prediction of a whole protein structure with multiple domains may not be directly possible, as there may not be a structural template for the whole protein in the PDB database. An observation has been that the folding of each structural domain of a protein, to a large degree, occurs independently of other domains, and hence each domain structure can be predicted independently, assuming the domains are represented in the database (Wetlaufer, 1978). One problem then becomes how to identify such domain boundaries in a protein sequence. Some protein sequences may contain signal peptides, which are not involved in folding the protein into its native structural conformation, and will eventually be cleaved out. Such complexity currently requires human expertise to guide a structure prediction process. In addition, each computer tool that addresses a particular issue often involves different adjustable parameters. Usually it takes a long time before a user can master each tool effectively. These difficulties often are the hurdles that prevent experimentalists from fully using protein structure prediction tools.

To integrate various computational analysis and prediction tools in an automated fashion, we have recently developed a computational pipeline (PROSPECT pipeline) for large-scale protein structure prediction. A distinguishing feature of this system is that it captures and incorporates expert knowledge from human predictors. It has been noted in the CASPs that one of the key reasons that computer-assisted human predictors have outperformed automated computer predictions is that human predictors can often refine computer predictions through better interpretation of the prediction results, using additional information and domain knowledge, integration of additional structural and functional information into the prediction process in an iterative manner, cross-validation of prediction results from different tools, and application of human intuition and judgment.

During previous CASPs, we developed an effective computer-assisted manual prediction procedure (Xu et al, 1999, Xu et al, 2001a), which involves a set of (human) decision-making and inference processes. These include tool selection criteria for different prediction conditions, integration of information from different

sources, cross-validation of prediction results from different tools, and intelligent interpretation of prediction results. A significant portion of this manual process has now been computationally implemented and incorporated into the PROSPECT prediction pipeline. Another unique feature of the pipeline is that it is accessible to the research community over the Internet (<http://compbio.ornl.gov/proteinpipeline/>). This is made possible, largely because of the availability of the powerful supercomputing resources available to us at the Oak Ridge National Laboratory. The pipeline has been implemented to run in a heterogeneous computational environment, consisting of Alpha, Solaris and Linux servers, a 64-node Linux cluster and a wide range of supercomputers as a client/server system with a web interface, which facilitates interactive communication between the pipeline and the user.

The rest of this paper is divided into four sections. Section 2 outlines the PROSPECT pipeline. Section 3 describes manual interpretation of the results obtained from the PROSPECT pipeline for protein structure and function analysis. Section 4 presents an example application of the PROSPECT pipeline in global analysis of three cyanobacteria genomes and an in-depth study of the carboxysomes common to all the three genomes. Section 5 summarizes the work presented.

## **2. Description of PROSPECT Pipeline**

In this section, we describe the components of the PROSPECT pipeline, which consists of a dozen prediction and analysis tools, built in-house or from third parties. The centerpiece of the pipeline is the PROSPECT threading-based protein structure prediction system (Xu and Xu, 2000).

### **2.1. Tool selection and key features**

The following nine prediction and analysis tools have been deployed to accomplish the required component functionality of the pipeline. More tools are being added to the pipeline. Each of these tools has a set of default parameters, suggested by the developers of these tools, which are used as the default values in the pipeline. The flow of the pipeline, as shown in Figure 1, is controlled by a set of rules which were derived from prediction experience gained through CASP and other prediction applications (Xu et al., 2000; Xu et al. 2002).

(1) **Signal peptide detection using SignalP** (Nielsen et al., 1997). SignalP predicts the signal peptide in the target protein sequence with very high accuracy (more than 90%). The PROSPECT pipeline cuts off the peptide at the identified cleavage site before running structure prediction tools.

(2) **Domain parsing using PRODOM** (Corpet et al., 2000). PRODOM identifies structural domains in a target protein sequence, by searching for the known protein domains in the PRODOM database. It saves computing time substantially and typically increases the threading accuracy by threading each predicted domain sequence against the structure template database.

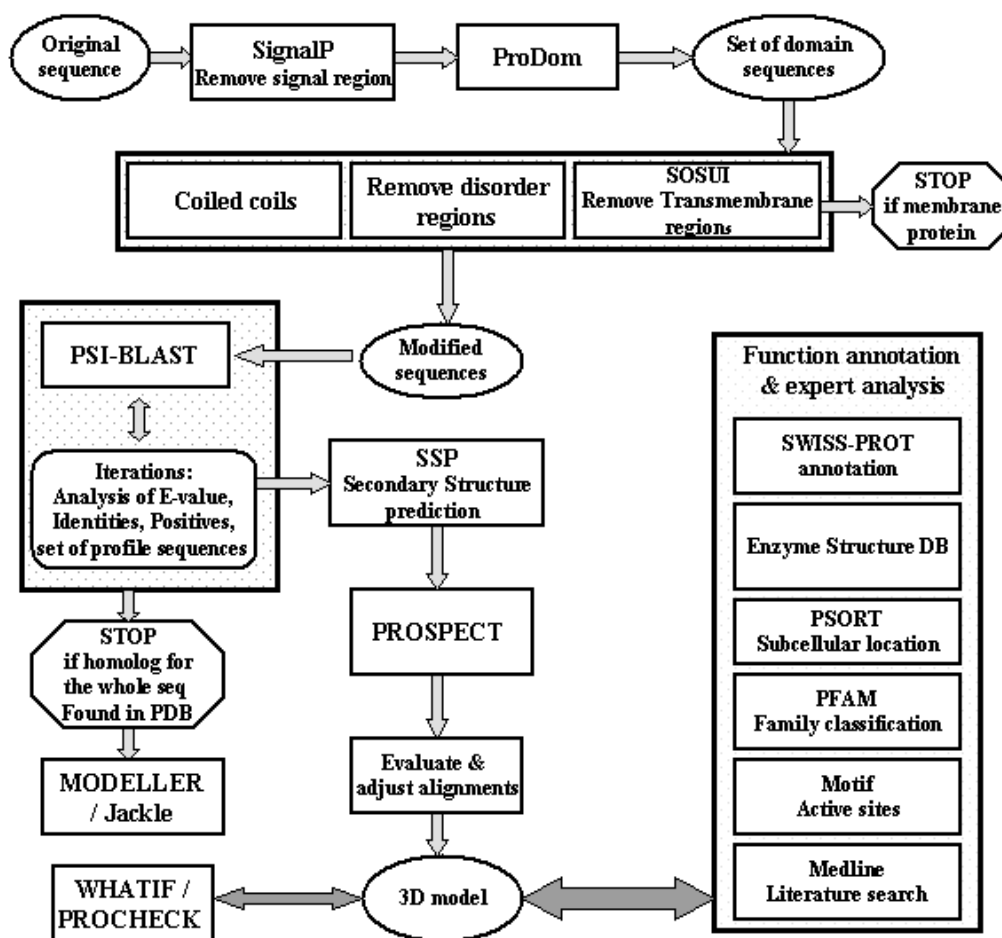


Figure 1. The prediction-process flowchart of the PROSPECT pipeline. Each rectangle represents an application of a computational tool; each oval represents a data set.

(3) **Secondary structure prediction using in-house tool SSP** (unpublished result). SSP uses a neural network technique to make secondary structure prediction, and its prediction accuracy is comparable to PSI-

PRED, which is close to 80% for predicting  $\alpha$ -helix,  $\beta$ -strand, and loop (Jones, 1999). The prediction result is used as an input to PROSPECT.

**(4) Homology search using PSI-BLAST** (Altschul et al., 1997). If a significant hit is found in PDB (Westbrook et al., 2000), threading may not be necessary. A significant hit in SWISSPROT (Bairoch and Apweiler, 1999) or some other databases can provide useful information such as the EC number of an enzyme and functional annotation. A pre-selected E-value threshold ( $10^{-4}$ ) is used as the default value for PSI-BLAST hit against PDB (release of Nov. 2002).

**(5) Prediction of membrane protein and its transmembrane regions using SOSUI** (Hirokawa et al., 1998). SOSUI's prediction accuracy of transmembrane regions is very high (greater than 90%). Membrane proteins have significantly different physiochemical properties than soluble ones. Since there are only a few templates available in PDB for membrane proteins and the energy function used in threading is derived from globular proteins, threading methods generally do not work for membrane proteins (Miyazawa and Jernigan, 1996). If a protein is predicted to be a transmembrane protein, the PROSPECT pipeline provides only the secondary structure predictions. However if a membrane protein has a soluble domain, the PROSPECT pipeline will predict the structure of this domain.

**(6) Protein threading using PROSPECT** (Xu and Xu, 2000). PROSPECT is a protein threading program with a number of unique features, compared to similar programs. These include a unique capability to rigorously deal with residue-residue contact potential, a key energy term for protein fold recognition (Xu and Xu, 2000). In addition, PROSPECT has a unique way of incorporating evolutionary information (position-dependent sequence profile) into all its energy terms, including the singleton energy term, the residue-residue contact term and the mutation energy term (Kim et al., 2002). These unique features have helped the program perform significantly better than many other threading programs.

For each fold prediction, PROSPECT provides a normalized threading score, calculated using a support vector machine (SVM) approach, based on the raw score of threading and other calculated features of the query sequence. PROSPECT also calculates a z-score that measures the reliability of the prediction (Kim et al., 2002), as shown in Table 1. The z-score is the threading score in standard deviation unit relative to the average of the threading raw score distribution of random sequences with the same amino acid composition and sequence length against the same structural templates. In practice, the average and the standard deviation are estimated by repeated threading between a template and a large number of randomly shuffled query sequences.

z-score interval	Probability to be correct	Confidence level	Similarity
< 6	<0.3	unlikely	fold/unrelated
6 - 8	0.35	low	superfamily/fold
8 - 10	0.63	medium	superfamily/fold
10 - 12	0.85	high	superfamily
12 - 20	0.96	very high	family/superfamily
> 20	>0.99	certain	family

Table 1. Interpretation of z-score. The first column represents the z-score range. The second column shows the probability of a sequence-template pair sharing the same fold within a certain z-score range. The third column shows a corresponding qualitative confidence level. The fourth column provides a possible relationship between the query and template protein in terms of the SCOP protein family classification, *family*, *superfamily*, and *fold* (Murzin et al., 1995). A query protein and its template are classified into a protein family if they have clear evolutionary relationship with significant sequence identity between them (often 25% or higher). Two proteins are considered to belong to the same superfamily if their structural and functional features suggest that they have a common evolutionary origin but not necessarily have a high sequence identity. They are considered to have the same fold if two proteins have same major secondary structures in the same arrangement and with the same topological connections but not necessarily have significant sequence similarity nor necessarily have a common evolutionary origin.

**(7) Construction of atomic structure using MODELLER and Jackle.** The PROSPECT pipeline employs both MODELLER (Sali and Blundell, 1993) and Jackle (Xiang and Honig, 2001) for detailed atomic model construction, based on provided sequence-structure alignments generated by either PROSPECT or PSI-BLAST. It allows the user to choose between the two programs for specific applications. The quality of the 3D model is mainly dependent on the sequence-structure alignment. If the alignment is correct, the RMSD for the 3D structure in the aligned region can be typically within 3 Å.

**(8) Quality assessment using WHATIF** (Vried, 1990). WHATIF provides a capability for evaluating the structural quality of a predicted structure. This includes the quality of side-chain packing and backbone conformations, the inside/outside occupancies of hydrophobic and hydrophilic residues, and stereochemical

quality of the predicted structure. Based on this assessment, the user can pick the best of the multiple structures derived from an alignment in the PROSPECT pipeline.

**(9) Information fusion using HitEvaluator.** HitEvaluator is an in-house program consisting of a set of rules for (a) cross-validating predictions of native-like folds and protein structures using information derived from different sources, (2) ranking and selecting the final fold and structure predictions, and (3) refining threading alignments using additional information. Currently the pipeline employs about 30 rules to help decision-making during its prediction process, and this number will be significantly increased in the next phase of expert system development. Additional rules for structure-based functional inference (see Section 3) are currently being added to the pipeline. The following gives one detailed rule in the PROSPECT pipeline:

*A structure template will be selected as a recognized structural fold if*

- 1) the E-value of the PSI-BLAST alignment between the template and the target protein is  $< 0.02$ , or*
- 2) the E-value is between 0.02 and 1.0 and its raw threading score or the z-score is ranked among the top 50 structural templates, or*
- 3) the template's EC number matches at least the first three digits of the EC number of the target protein and the raw threading score or the z-score of the template is ranked among the top 50 templates, or*
- 4) the ranks of its raw threading score and the z-score are both among the top 20 predicted folds.*

The following subsection provides a more detailed description of the rules employed by the PROSPECT pipeline.

## **2.2. Expert rules for improving structure predictions**

Currently, rules that have been developed or under development are classified into two categories: (a) rules for fold selection and (b) rules for alignment refinement.

### **Rules for fold selection based on multiple information sources and fold ranking**

**(1) Functional comparison.** The most helpful supplemental information for evaluating whether a proper fold has been predicted is often the function of the query protein. Many proteins with unknown structures have functional annotation, either from experiments or from sequence-based homology search. If a template

is among the top ranking hits (either by z-scores or by raw scores) and has the same or similar function to the query, this improves confidence that the template represents the native-like fold of the query protein. If both the query and template proteins are enzymes, their functional comparison is done by comparing their Enzyme Classification (EC) numbers. Otherwise, their functions are compared through matching functional keywords.

**(2) Consensus among different scores.** Experience shows that the “correct” fold templates often have top ranking simultaneously in the raw score, SVM score, and the z-score, even when the top hits have relatively low z-scores. Hence if a template ranks very high (top 10 among all structural templates) by all scoring schemes, it often indicates that the template probably represents a native-like fold of the query protein. Such a criterion often provides the best guess for the templates of the native-like fold for the query protein.

**(3) Fold Consensus among different templates.** It is often a good indication that a top-ranked template has the correct structural fold when it shares the same or similar fold with some other top-ranked templates. Note that the structural templates may not share any significant sequence similarity, but they can still represent the same fold. If multiple templates representing the same fold are among the top hits, the confidence level for the templates to represent the native-like fold of the query protein should be high.

**(4) Secondary structure comparison.** Consistency between the predicted secondary structures by SSP and the secondary structure assignments of the predicted structure by DSSP is often a good indication about the correctness of a fold template. A poor agreement (less than 50%) between them usually indicates an incorrect prediction. Based on this, many templates can be excluded from consideration for native-like folds.

**(5) Motif matching.** Some motifs identified through sequence-based analysis, such as coiled coils and WD40 repeats, have long segments in sequence, and they can be predicted accurately. These motifs can be used to verify whether a template represents the native-like fold of the query protein, by checking if both query and template have the same motif. For example, in the vitronectin structure prediction (Xu et al., 2001b), a sequence-based analysis found WD40 motifs. The template identified (1gen) also contains WD40 repeats. This increased the confidence of the prediction.

## **Rules for alignment refinement**

**(1) Using active sites to refine sequence-structure alignment.** When two proteins have a similar function, they often have the same type of active site, which should be aligned with each other in the

sequence-structure alignment. Aligning the active sites provides a highly useful constraint in the threading process, and can improve the accuracy of threading alignments. The active site information of a query sequence is predicted by using MOTIF (<http://www.motif.genome.ad.jp>), PROSITE (Hofmann et al., 1999), PRINTS (Attwood et al., 1999), or BLOCKS (Henikoff et al., 1999). The active site information for a structural template can be obtained from databases such as PDBsum (<http://www.biochem.ucl.ac.uk/bsm/pdbsum/>). Phylogenetic analysis or multiple sequence alignment can be used to make active-site prediction more accurate. If the predicted active site is conserved among the protein family, or at least conserved among the neighbors on a phylogenetic tree, the predicted active site is more likely to be real.

**(2) Compactness of aligned region on template.** Generally, a protein structure should be compact. Hence compactness of a predicted structure can be used to assess the *goodness* of a threading alignment. A good estimate for the compactness of a predicted structure or any of its structural domains is its normalized radius of gyration. If the value is above 3.0, the aligned portion is not compact enough and the alignment is probably not correct.

### **2.3. Using PROSPECT pipeline through Web interface**

Applications of the PROSPECT pipeline are typically triggered from the Web interface, although in case of batch processing, a command line interface (not publicly available) can also be used. The Web interface is used to submit a query sequence to the prediction server, to choose a set of prediction tools and their parameters, to start the execution, and to review the status of the prediction process and results. A user can run all the available tools using the default setting, where all that is needed is for the user is to load/paste a protein sequence and hit the “Submit” button. The pipeline can also be used in an advanced mode, where options can be selected and parameters modified before the job is submitted (as shown in Fig. 2). The pipeline assigns a unique Query ID to each job submitted to the server, which can be used to retrieve execution results after the web browser had been closed. All the data related to the particular Query ID is kept at the server for an extended period of time. Upon submission of the request, the interface provides continuous feedback to the user about the pipeline’s prediction status, by periodic update of the status web page. The status web page also allows the user to inspect the results of individual tools, during and after the execution of the pipeline request. The status page is also linked to the results of individual tools. All tools are applied to the target sequence or to its subsequences, and the status page reflects the status of analysis of each domain (see Fig. 3). It indicates if the domain is still being processed by a particular tool, or how successful the analysis was. This page also has links to pages with more detailed information about each

application's processing status. After all domains had been processed, the system considers the whole pipeline run successful if there were no failures of tools on any sequence segment.

## Computational Biology at ORNL

[Home](#) [About Us](#) [Analysis Tools](#) [Information Resources](#) [Projects & Research](#)  
Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

QueryID  ProcType

ORNL Protein Pipeline. Query: T0129 (casp-pipeline-global)

[Home](#) [Simple Status](#) [Detail Status](#) [Output](#) [Error](#) [Params](#)

### Target Sequence

ID:  Name:

Type:  Organism:

Input Sequence:

FASTA File:

all\_services

**Signalp** Signal peptide cleavage site identification [CBS-DTU](#)

Figure 2. The (partial) Web interface for submitting a job and configuring the pipeline.

# Computational Biology at ORNL

The screenshot shows a web interface for the ORNL Protein Pipeline. At the top, there are navigation tabs: Home, About Us, Analysis Tools, Information Resources, and Projects & Research. Below these is a dark bar with links: Channel, Generation, Grail, GrailEXP, Pipeline, Parser, and PROSPECT. The main heading reads "ORNL Protein Pipeline results for query T0129 (casp-pipeline-global)". Below this is another set of navigation tabs: Home, Simple Status, Detail Status, Output, Error, and Params. A status bar indicates "Pipeline Finished" with a progress indicator. A legend shows "Processing" (white square), "Success" (grey square), and "Failure" (black square). The page lists two sequence domains: "Sequence\_0 domain (whole sequence: 1 -- 182)" and "Sequence\_1 domain (80 -- 179)". For each domain, there are links to various analysis tools: ProdomBlast, Constructor, Prospect2Casp, Sosui, Coils, PsiBlast, Ssp, Prospect, HitEvaluator, Modeller, and Whatif. At the bottom, there are links for Site Map, Feedback, Life Sciences Division, ORNL, Disclaimer, and Webmaster.

Fig. 3. The Web page for status report of a submitted job with links to all related analysis and prediction results.

## 3. Interpreting Results of PROSPECT Pipeline

Though the PROSPECT pipeline generally works well on many types of proteins just by running the default options without human intervention, one can improve the performance of structure prediction and obtain more functional information by using additional bioinformatics tools and human judgement. Our approach to this generally considers three parts: (1) human evaluation of threading results for template selection, (2) manual refinement of structural model, and (3) functional inference based on predicted structures. More expert knowledge in these areas is being integrated into the pipeline, so that over time, less manual work will be required.

### 3.1. Enhanced manual interpretation of the template

When the confidence level of a computer prediction is low (with  $z$ -score less than 8), human evaluation is often helpful in selecting a good template based on not only the pipeline results but also the collected structural and functional information through literature and Internet search. Literature search about each query sequence plays an important role in accurately characterizing a protein's structure and function. A

query protein can often be found in a protein database, e.g., SWISSPROT, which contains the literature references about the protein. The abstracts of these references and related publications in PubMed (<http://www.ncbi.nlm.nih.gov/PubMed/>) can provide valuable structural and functional information about a query protein. One can also use the protein identification number and related keywords obtained from protein databases to do an Internet search for possible unpublished results. Key pieces of experimental data, such as the location of a disulfide bond, are extremely available and can be used to verify whether a template may possibly represent a native-like fold. If no template can be found to represent the native-like fold with good confidence and the z-score is low (less than 6), the query protein may be assigned as a possible new fold. The identification of possible new folds may be useful for identifying novel fold targets in the structural genomics (Burley et al., 1999).

### **3.2. Manual refinement of predicted structures**

Manual refinements of the automatically generated structures can often improve the accuracy of these structures, based on information that is currently difficult or non-trivial to apply by the PROSPECT pipeline. Sometimes, the refinement process needs to be carried out iteratively. The following procedure for such refinements is as follows:

**(1) Manual adjustment.** Most often, MODELLER is used to generate a 3D structure based on provided sequence-structure alignment. MODELLER works reasonably well when a query and its template have high sequence similarity, but it was not designed for cases where the query protein has no significant sequence similarity to its template. Hence, the quality of structural models generated from a direct run of MODELLER can be poor for such cases. One can improve the quality of a structural model by doing refinements using loop modeling, multiple templates, etc. A user should also visually check structural models. If the user finds some obvious errors in a model, e.g., overlap of residues in space, the alignment should be fine-tuned and the model rebuilt.

**(2) Domain docking.** When a query protein is partitioned into several domains for separate threading jobs, different domains generally have different native-like templates that cannot be easily connected. A user can apply the protein-docking program GRAMM (Vakser, 1996) to build a structural model for the whole query protein by docking the domains together (see the example of vitronectin in Xu et al., 2001b). GRAMM is particularly suitable for this type of application, since it allows inputs of low-resolution structures such as predicted domain structures.

### 3.3. Structure-based functional inference

For many hypothetical proteins, fold recognition may provide useful information about function. In many cases, the inference can suggest potential function or a type of function, (e.g., whether a protein binds DNA), and it may further suggest the function of each domain and important active sites. Here we provide a list of rules we have been using for functional inference based on predicted structures.

**(1) Suggesting possible functions based on a predicted fold.** Although proteins with the same fold may not have the same function, their functions are often evolutionarily related. One can suggest a possible function of a query sequence from its predicted fold, using the SCOP database (Murzin et al., 1995), which provides a hierarchical classification of proteins into families, superfamilies, and folds. When the z-score of our prediction is high, the query and the template are likely to be in the same family or superfamily, as shown in Table 1. When the z-score is not high, one can check the functions of all the superfamilies in the predicted fold, to identify the most likely function based on additional information collected about the query protein.

**(2) Suggesting function type based on consensus among top hits.** Sometimes, the user may not be able to identify a specific fold for a query protein reliably, since the prediction confidence level may be low or the query protein may represent a new fold. However, the top hits often may mostly contain one specific type of protein, e.g., DNA-binding proteins. It is possible that PROSPECT has captured in the list some common features among such a type. In such cases, the consensus among the functions of the top hits (e.g., signal transduction proteins, protein related to phosphate group, etc.) may suggest the function of the query protein. A user can easily check the functions of all the top-hit templates by using the Web-interface of the pipeline output, which lists the functions of each template.

**(3) Identifying function through motifs.** To further pin down the function of a query protein from an identified fold, one can apply functional motifs as additional constraints during the process of functional assignment. If the predicted structure contains a functional motif (conserved residues at a particular position in the 3D structure, not necessarily close to each other on the protein sequence) of a protein in a fold family, the query protein is likely to have the same function as the template containing the motif. Zhang *et al.* (1999) have shown that structure-based motifs help to identify whether two proteins in the same fold family share the same function. The authors have constructed a database of functional motifs for known structures (e.g., EF-hand motif for calcium binding), called SITE. Currently, SITE contains identified functional motifs from about 50% of the SCOP superfamilies. A comparison of motifs could be made between the proteins in

the predicted fold family and the predicted model of the query protein to determine if the functional motif is conserved. One can also search the predicted protein structure against PROCAT (Wallace et al., 1996), which is a database of 3-D enzyme active site templates.

**(4) Sequence-based function search.** A protein family classification based on Pfam (Bateman et al., 1999) can provide useful function annotation. A prediction of sub-cellular locations using PSORT (Nakai and Horton, 1999) and SubLoc (<http://www.bioinfo.tsinghua.edu.cn/SubLoc/>) may assist such a functional inference, by narrowing down the possible functions to a small list. For example, if a protein is predicted to be extracellular, it is not likely to be a DNA binding protein.

**(5) Confirming function from electrostatics.** The user can calculate continuum electrostatics of a predicted structure using DELPHI (Honig and Nicholls, 1996) and visualize the electrostatics profile on the surface of the protein using GRASP (Nicholls et al., 1991). Electrostatics profiles may confirm functional assignment or active site identification. For example, if a protein is suspected to be involved in DNA binding based on the results from PROSPECT, a strong electrostatic potential on a region of the protein surface may provide evidence of a DNA binding site.

#### **4. Computational Studies of Protein Structures in *Cyanobacteria***

This section presents one application of the combined computational/manual protocol described above for genome-scale protein structure and function prediction to predicted genes in three cyanobacterial genomes; *Synechococcus* sp. WH8102 and two strains of *Prochlorococcus* sp. (MED4 and MIT9313). The procedure provided both a global analysis of the structural folds in the three genomes and a detailed study of several predicted proteins that have been suggested to be essential for the function of carboxysomes, the common microcompartments that are presented in these photosynthetic microorganisms.

##### **4.1 Overview of three genomes**

The cyanobacterial community in the world open oceans is dominated by small unicellular forms of two genera *Synechococcus* and *Prochlorococcus*. It has been suggested that these microorganisms, in various marine environments, compete for a similar ecological niche such that the total biomass of the two genera is relatively constant (Chisholm et al., 1992). Together, these organisms are major primary producers in large oligotrophic central gyres of the world's oceans. Although the two genera are frequently present together, *Synechococcus* is widely distributed and dominant in surface water that is rich in nutrients whereas

*Prochlorococcus* is limited to 40°N-40°S latitudes and often found in oligotrophic waters (Partensky et al., 1999). Furthermore, though both ecotypes of *Prochlorococcus* are abundant in temperate and tropical oceans, absorbing blue light efficiently for their carbon fixation, *Prochlorococcus* sp. MIT9313 is adapted to lower light conditions (at deeper ocean) than *Prochlorococcus* sp. MED4. Because regeneration of organic carbon is a critical step in response to anthropogenic inputs of CO<sub>2</sub> into the atmosphere and thus highly relevant to global carbon recycling, a major focus of biological oceanography has been to study, predict, and manipulate the process of carbon fixation in the ocean. To support these scientific pursuits, the genomes of *Synechococcus* sp. WH8102 and two strains of *Prochlorococcus* sp. (MED4 and MIT9313) have been sequenced recently by the Department of Energy's Joint Genome Institute. The availability of these complete genomes enables researchers to study the global properties of protein folds of these genomes, using computational biology approaches. Such studies will provide one mechanism to better understand the structure and function of the proteins encoded by these genomes.

#### 4.2. Global analysis of protein structural folds in three genomes

Protein structure predictions were carried out for all the predicted genes in the three genomes using the PROSPECT pipeline. The gene predictions were extracted from ORNL's Genome Channel at <http://compbio.ornl.gov/channel>. Each genome took about one week for the pipeline to finish all the predictions. These results can be accessed through the Internet at:

*Synechococcus* sp. WH8102: <http://compbio.ornl.gov/PROSPECT/syn/>

*Prochlorococcus* sp. MIT9313: [http://compbio.ornl.gov/PROSPECT/pmar\\_mit/](http://compbio.ornl.gov/PROSPECT/pmar_mit/)

*Prochlorococcus* sp. MED4: [http://compbio.ornl.gov/PROSPECT/pmar\\_med/](http://compbio.ornl.gov/PROSPECT/pmar_med/)

Overall, the PROSPECT pipeline identified structural homologs in PDB with reasonable level of confidence (through either PSI-BLAST with E-value less than 10<sup>-4</sup> or PROSPECT with Z-score 6.0 or above) for 54.8%-59.8% of all the ORFs in each of the three genomes. Together with annotations of membrane proteins, about 80% of all the ORFs in each genome are characterized. Table 2 provides a summary of these prediction results.

Species	<i>Synechococcus</i> sp. WH8102	<i>Prochlorococcus</i> sp. MIT9313	<i>Prochlorococcus</i> sp. MED4
Total number of ORFs	2502	2251	1694
Membrane proteins	548 (21.9%)	560 (24.9%)	436 (25.7%)

PSI-BLAST hits	867 (34.7%)	867 (38.5%)	640 (37.8%)
PROSPECT ( $z = 20$ )	328 (13.1%)	137 (6.1%)	196 (11.6%)
PROSPECT ( $12 = z < 20$ )	81 (3.2%)	53 (2.4%)	47 (2.8%)
PROSPECT ( $10 = z < 12$ )	39 (1.6%)	28 (1.2%)	23 (1.4%)
PROSPECT ( $8 = z < 10$ )	55 (2.2%)	38 (1.7%)	25 (1.5%)
PROSPECT ( $6 = z < 8$ )	126 (5.0%)	111 (4.9%)	80 (4.7%)
Total number of structural homologs predicted	1496 (59.8%)	1234 (54.8%)	1011 (59.7%)

Table 2. A summary of predicted structural folds in three cyanobacterial genomes. Membrane proteins are predicted using the SOSUI program. For soluble proteins, PROSPECT was applied to a gene only when it does not have a PSI-BLAST hit. Each row represents the total numbers of structural homologs predicted in three genomes, in a particular range of z-scores of PROSPECT hits or with PSI-BLAST hits. For an interpretation of each z-score category, we refer the reader to Table 1.

For all genes with predicted structural folds in PDB as listed in Table 2, we have performed a distribution analysis of their predicted folds. We found that the distribution of the fold occurrence for the three genomes follows a power-law distribution as shown in Fig. 4, similar to what has been observed in 20 other genomes (Qian et al., 2001). More specifically, a few folds occur many times in the genome while many folds occur infrequently. Tables 3-5 list the ten most commonly occurring superfamilies and their SCOP identification numbers for each genome. Most structural folds in these lists are common to all three genomes, including (1) P-loop containing nucleotide triphosphate hydrolases, (2) NAD(P)-binding Rossmann-fold domains, (3) membrane ion channel-forming peptide antibiotics containing D-aminoacids, (4) S-adenosyl-L-methionine-dependent methyltransferases, (5) FAD/NAD(P)-binding domain, (6) PLP-dependent transferases, alpha/beta-hydrolases and (7) nucleotide-diphospho-sugar transferases. Two superfamilies (membrane all-alpha and UDP-Glycosyltransferase/glycogen phosphorylase) only show up in the top-ten list of *Synechococcus* sp. WH8102; two superfamilies (winged helix DNA-binding domain and MOP-like) only show up in the top-ten list of *Prochlorococcus* sp. MIT9313; and two superfamilies (nucleic acid-binding proteins and thioredoxin-like) only show up in the top-ten list of *Prochlorococcus* sp. MED4.

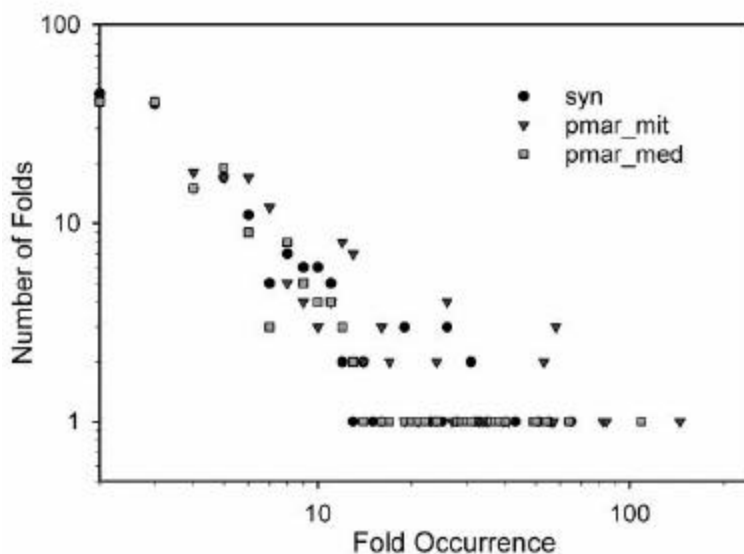


Figure 4. The distributions of fold occurrences in the three genomes. The horizontal axis shows the number of ORFs represented by the fold and the vertical axis shows the number of folds that represent a particular number of ORFs. For example, a point at (10, 5) represents a total of five structural folds in a particular genome, each of which represents ten ORFs. Both axes are in logarithmic scale. The plot for each genome is close to linear distribution in the logarithmic scale, indicating the distribution of the fold occurrence of the three genomes follows a power-law distribution.

Rank	Superfamily	SCOP	# of occurrences
1	P-loop containing nucleotide triphosphate hydrolases	c.37.1	137
2	NAD(P)-binding Rossmann-fold domains	c.2.1	55
3	Nucleotide-diphospho-sugar transferases	c.68.1	35
4	Membrane ion channel-forming peptide antibiotics containing D-aminoacids	j.1.1	33
5	S-adenosyl-L-methionine-dependent methyltransferases	c.66.1	32
6	FAD/NAD(P)-binding domain	c.3.1	31
7	Membrane all-alpha	f.2.1	28
8	PLP-dependent transferases	c.67.1	26
9	Alpha/beta-hydrolases	c.69.1	26
10	UDP-glycosyltransferase/glycogen phosphorylase	c.87.1	24

Table 3. Top ten most commonly occurring superfamilies in *Synechococcus* sp. WH8102.

Rank	Superfamily	SCOP	# of occurrences
1	P-loop containing nucleotide triphosphate hydrolases	c.37.1	145
2	PLP-dependent transferases	c.67.1	58
3	S-adenosyl-L-methionine-dependent methyltransferases	c.66.1	57
4	NAD(P)-binding Rossmann-fold domains	c.2.1	53
5	Alpha/beta-hydrolases	c.69.1	53
6	FAD/NAD(P)-binding domain	c.3.1	41
7	Membrane ion channel-forming peptide antibiotics containing D-aminoacids	j.1.1	39
8	Nucleotide-diphospho-sugar transferases	c.68.1	33
9	Winged helix DNA-binding domain	a.4.5	30
10	MOP-like	b.40.6	27

Table 4. Top ten most commonly occurring superfamilies in *Prochlorococcus* sp. MIT9313.

Rank	Superfamily	SCOP	# of occurrences
1	P-loop containing nucleotide triphosphate hydrolases	c.37.1	109
2	NAD(P)-binding Rossmann-fold domains	c.2.1	54
3	Membrane ion channel-forming peptide antibiotics containing D-aminoacids	j.1.1	50
4	S-adenosyl-L-methionine-dependent methyltransferases	c.66.1	38
5	Alpha/beta-hydrolases	c.69.1	36
6	FAD/NAD(P)-binding domain	c.3.1	30
7	PLP-dependent transferases	c.67.1	29
8	Nucleotide-diphospho-sugar transferases	c.68.1	24
9	Nucleic acid-binding proteins	b.40.4	22
10	Thioredoxin-like	c.47.1	21

Table 5. Top ten most commonly occurring superfamilies in *Prochlorococcus* sp. MED4.

#### 4.3. Computational analysis of predicted carboxysome proteins

The carboxysome is a polyhedral inclusion body that is found in a variety of microorganisms (Cannon et al., 2001). The components and the overall structure of the carboxysome is poorly understood although it is known that an enzyme named ribulose1,5, biphosphate carboxylase/oxygenase (RuBisCO) -- the major enzyme that convert carbon from an inorganic to organic form -- constitutes about 60% of the total carboxysomal proteins (Friedberg et al., 1993 and Cannon et al., 1983). Sequence homology search suggests that in the three organisms under study, the functional RuBisCo unit consists of eight large subunits (*rbcL*) and eight small subunits (*rbcS*). It is also known that the genomes of *Synechococcus* WH8102, *Prochlorococcus* sp MED4 and MIT9313 contain only one copy of the large subunit (*rbcL*) and one copy of the small subunit (*rbcS*), whose locations are adjacent to each other in the genome sequence. This close arrangement implies that these genes may be co-regulated or co-transcribed, as in many other cyanobacteria that have been experimentally verified. Besides RuBisCo, several carboxysomal shell proteins have been identified in *H. neapolitanus*. For example, *ccmK* (*csoS1* homolog) was identified through peptide sequencing (English et al., 1994). This gene appears to be duplicated twice in the *H. neapolitanus* genome (*csoS1A*, *1B* and *1C*). Two other genes encoding shell proteins *csoS2* and *csoS3* were identified using a battery of techniques (Baker et al., 1999 and 2000). These genes, along with ORFA and ORFB, are clustered with the two structural genes for RuBisCO (*cbbL* and *cbbS* in *H. neapolitanus*) on the genomic sequence. Recent sequence analysis has shown that these carboxysome shell genes are not only found in *Synechococcus* WH8102 and the two strains of *Prochlorococcus*, but their clustered relationship along with the RuBisCO genes (*rbcL* and *rbcS*) is also conserved in the genomic sequences (Cannon et al., 2001). The clustered arrangement and the conserved homology of these genes suggest that these (predicted) proteins are among universal components of carboxysomes.

In contrast to the well-studied function of RuBisCo, little is known about the functions of the shell proteins. In *Synechococcus PCC7942*, disruption of *ccmK* (homologous to *csoS1*) and *ccmL* (homologous to ORFA and ORFB) by insertion of chloramphenicol drug marker gene  $Cm^r$  correlates with a severe reduction in cell growth at the atmospheric CO<sub>2</sub> level (Price et al., 1993). These observations suggest a critical role of *csoS1* and ORFA/ORFB in promoting cell growth under the atmospheric CO<sub>2</sub> level. However, despite the intensive efforts to elucidate the structure and function of carboxysomes, many important questions concerning carboxysome structure and functions of shell proteins remain unanswered. Aiding to the effort of elucidating carboxysome structure and function in this study, we have applied the PROSPECT pipeline, in conjunction with other bioinformatics tools, to predict structures and functions for the proteins that may be critical for the function of the carboxysome in carbon fixation.

Sequence analysis of *Synechococcus* sp. WH8102 and two strains of *Prochlorococcus* sp. (MED4 and MIT9313) genomes reveals many similarities in the predicted protein sequences and the organization of carboxysome genes (Cannon et al., 2001). Figure 5 shows that these predicted genes are, in general, arranged in the order of a *ccmK1*, a *rbcL*, a *rbcS*, an operon of *csoS2*, *csoS3* and two other ORFs, and another copy of *ccmK* gene (Cannon et al., 2001). Many of these predicted genes are highly similar across the three genomes. Furthermore, the first copy and the second copy of *ccmK* genes (e.g., *ccmK1* and *ccmK2* in *Synechococcus* sp. WH8102) in all three organisms share a C-terminal core of 87 amino acids whose sequence identities range from 64-99%, whereas the *ccmK2* orthologs in *Synechococcus WH8102* and *Prochlorococcus MIT9313* contain an N-terminus of low complexity region consisting of many prolines, alanines and glycines. Consistent with the high levels of sequence similarities between these predicted proteins, the intergenic regions spacing among *csoS2*, *csoS3*, ORFA and ORFB are also very similar in all three organisms (see Figure 5). Due to their closeness in the genomic sequence, it is highly probable that these proteins belong to the same operon, as it has been reported in *Thiobacillus neapolitanus* (Shively et al., 1996). Additionally, Figure 5 shows that the intergenic regions between *ccmK1* and *rbcL*, *rbcS* and *csoS2* and ORFB and *ccmK2* (*csoS1* ortholog) genes are similar, whereas the intergenic regions between *rbcL* and *rbcS* range from 60-108 nucleotides in all three organisms. As various lengths of the intergenic region between *rbcL* and *rbcS* have been reported (Tabita et al., 1990), it is interesting to observe that other intergenic spacing between neighboring ORFs from *ccmK1* (*csoS1*) to ORFB is conserved in these organisms. These observations implicate that whereas transcription of *csoS2*, *csoS3*, ORFA and ORFB are tightly controlled in an operon, the regulation of gene expression of others genes such as *ccmK1* and *rbcL*, or *rbcS* and *csoS2* operon, is probably transcriptionally co-regulated.

Though several other genes required for carboxysomal biogenesis have also been identified by genetic analysis (Cannon et al., 2001), their roles are unclear and these will not be addressed. Structure prediction has been focused on the genes presented in Figure 5. Careful examination of the predictions from the PROSPECT pipeline for all the genes in Figure 5 (other than the two RuBisCo ORFs, i.e., *rbcL* and *rbcS*, which have close homologs of known structures) reveals that none of them has significant PSI-BLAST hit or high z-score hit by PROSPECT. Hence, further manual assessments of the prediction results were performed.

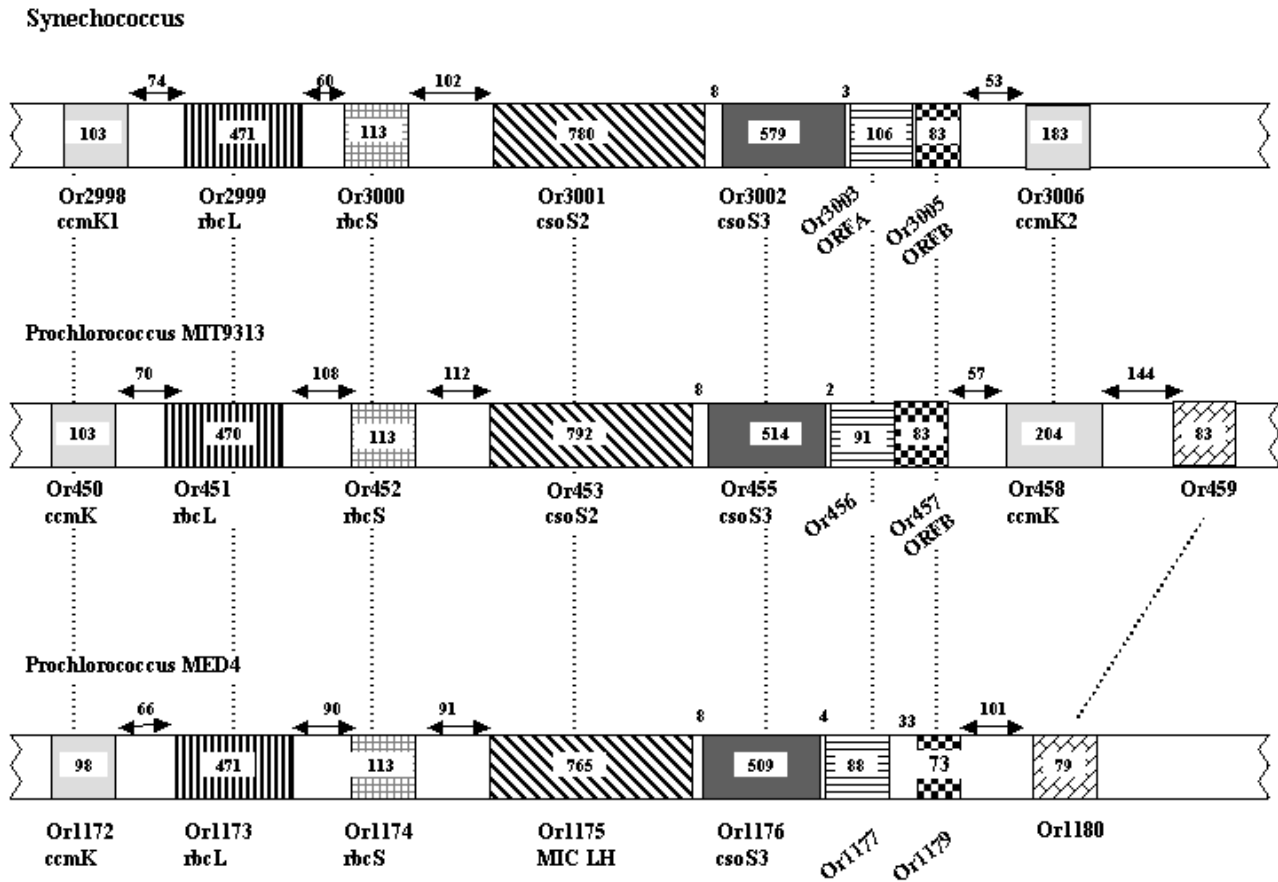


Fig. 5. Comparison between the putative carboxysome-related ORFs clustered in the genomic sequences in the three genomes. Each ORF is represented as a shaded rectangle with the ORF identification number and gene name (if available) marked under the rectangle and the number of amino acids of the gene marked in the rectangle. The numbers of nucleotides in the intergenic regions are marked between corresponding rectangles. The orthologous relationships between genes in different genomes is represented by dashed lines.

Sequence analysis, using PSORT (<http://psort.nibb.ac.jp/>) and SubLoc (<http://www.bioinfo.tsinghua.edu.cn/SubLoc/>), predict that all the genes presented in Figure 5 are localized in the cytoplasm. These results are in agreement with analysis using SignalP, which suggests that none of the protein sequences contain signal peptide (Nielsen et al., 1997). Based on the results from the PROSPECT pipeline and manual predictions, the final predictions for the structures and functions of these orthologs are shown in Table 6. All the selected templates consistently rank at the top across all three genomes in terms of its raw score, SVM score and z-score (see the prediction result websites provided in Section 4.2).

ORF	Template	Template function	Predicted ORF function
ccmK1	Iris	Ribosomal protein S6	Metallochaperone
csoS2-1	liir-A	Glycosyltransferase	Structural protein
csoS2-2	1fnf	Fibronectin	
csoS3	lqlt-A	Vanillyl-alcohol oxidase	Oxidase
ORFA	1kt9	Diadenosine tetraphosphate hydrolase	Phosphate hydrolase
ORFB	1kt9	Diadenosine tetraphosphate hydrolase	Phosphate hydrolase
ccmK2	Iris	Ribosomal protein S6	Metallochaperone
Or459 (MIT9313)	1dcp	Pterin-4a-carbinolamine dehydratase	Dehydratase

Table 6. Structure and function predictions for genes in Fig. 5. The table lists the ORF name, the PDB code (and chain name) of the template used, the function of the template, and the predicted function based on most related superfamily of the template fold. csoS2-1 and csoS2-2 are the N-terminal domain and C-terminal domain of the csoS2 ORF, respectively.

For the two copies of ccmK genes, i.e., or2998 and or3006 in *Synechococcus* sp. WH8102, or450 and or458 in *Prochlorococcus* sp. MIT9313, and or1172 in *Prochlorococcus* sp. MED4, it was predicted that their conserved C-termini adopt a ferredoxin-like fold, as shown in Fig. 6(a). Interestingly, the large subunit (rbcL) of RuBisCo also contains a ferredoxin-like fold at the N-terminus, although it has no significant sequence similarity to the ccmK protein. The structural similarity suggests that the ccmK gene and the rbcL gene may be evolutionarily related through gene fusion and duplication. The best template in terms of threading scores was Iris, the ribosomal protein S6. Given the context of the carboxysome, ccmK probably does not function as a ribosomal protein and functions of all superfamilies that adopt the ferredoxin-like fold were examined to find a possible function for ccmK1. The most believable and relevant function is metallochaperone, which may assist folding and stability of other carboxysome proteins. This is consistent with experimental data that shows that disruption of ccmK1 gene in *Synechococcus* sp PCC 7942 results in a high carbon requirement phenotype, and disruption of the upstream region of the gene cassette that contains ccmK inhibits carboxysome maturation (Price et al. 1993; Orus et al., 1995). Moreover, it was reported that the chaperon activity is required for the efficient folding of the large subunit of RuBisCo (rbcL), and the distance between ccmK and rbcL is very similar in the all three genomes, suggesting that these two genes are transcriptionally co-regulated. Altogether, it is reasonable to hypothesize that a role of ccmK1 is to facilitate the folding/assembly of rbcL and some other proteins in the carboxysome together with another identified chaperon Cpn60 (Kovacs et al., 2001). In addition, the low complexity region at the N-terminus of

ccmK2 is highly hydrophobic and abundant in prolines, alanins and glycines. Due to these features, the secondary structure prediction shows that the N-terminus of ccmK2 is basically in loop conformation. Accordingly, this domain may provide the hydrophobic patches needed for chaperone activity in general. The hydrophobic N-terminus can also contribute to the stability of the carboxysome through facilitating protein-protein interaction.

For the csoS2 orthologs, due to the difficulty in finding a good template for the whole sequence, the protein sequence was partitioned into two domains. The domain partitions are 1-277 / 278-780, 1-237 / 238-765 and 1-219 / 220-792, for *Synechococcus* sp. WH8102, *Prochlorococcus* sp. MIT9313, and *Prochlorococcus* sp. MED4, respectively. Whereas the N-terminal domain matches the template liir-A, a glycosyltransferase gtfb, the C-terminal domain matches 1fnf, a fibronectin fragment (see structural models in Figure 6(b,c)). This fibronectin fragment encompasses the 7th-10th type III repeat, whose presence is found in many proteins with a broad range of biological function, and an RGD motif which is an integrin-fibronectin interaction site (Leahy et al., 1994; Steward et al., 2002). Therefore, the C-terminus of csoS2 could play a structural role, or it may be important for promoting protein-protein interactions. In addition, from PSI-BLAST search, csoS2 are homologous to histones, which also play a role as structural proteins, even though their molecular structure is unknown. Interestingly, it has been reported that antibodies raised against CsoS2 labels the edges of the carboxysome of *Thiobacillus neapolitanus* (Baker et al., 1999). Combined with the experimental evidence, it may be expected that csoS2 is present in the carboxysome shell (Baker et al. 1999). Sequence homology and protein folding of these proteins further suggest that csoS2 may be engaged in a structural role in the forming of the backbone of the carboxysome.

For the csoS3 orthologs, the best template that we identified is 1qlt-A, which is vanillyl-alcohol oxidase (see structural models in Fig. 6(d)). Interestingly, the N-terminal domain of vanillyl-alcohol oxidase structure, like the ccmK genes, belongs to the ferredoxin-like fold (categorized by SCOP classification) whose superfamilies include many proteins that act as either enzymes or binding domains. Again, the large subunit (rbcL) of RuBisCo also contains a ferredoxin-like fold at the N-terminus, although it has no significant sequence similarity to the csoS3 protein. The structural similarity suggests that csoS3 and rbcL may be evolutionarily related through gene fusion and duplication. In *Thiobacillus neapolitanus*, csoS3 expression is less abundant when compared to the expression of RuBisCo genes and CsoS1, and the protein was shown to be associated with the periphery of the carboxysome (Baker et al., 2000). Due to the position of csoS3 in the cluster of genes that are important for carboxysome function and its lower abundance in a cell, it is possible that csoS3 could function as an enzyme, or participate in protein-protein interaction that is critical for the function of the carboxysome. In line with this reasoning and the results from protein structure

prediction, we hypothesize that *csoS3* may function as a reductase, which is one of the most recurrent enzymes among proteins that contain a Ferredoxin-like fold. As a reduced environment is required for the optimum activities of a few critical enzymes in carbon fixation process (such as sedoheptulose biphosphatase and fructose biphosphatase), the function of *csoS3* as a reductase could facilitate the process of converting carbon source from inorganic form to organic form through RuBisCo activity.

As we know from experimental results, disruption of *ccmL*, which is homologous to ORFA and ORFB, leads to a high-carbon-requirement phenotype in *Synechococcus* sp. PCC7942 (Price et al., 1993). ORFA, ORFB, and Or459 (MIT9313) may also work in the same biological pathway based on the predictions. ORFA and ORFB are homologous, and as expected they are predicted to adopt the same template, which is 1kt9 (diadenosine tetraphosphate hydrolase, see structural models in Fig. 6(e)). This superfamily contains enzymes that are associated with substrates containing two or more number of phosphate groups and we also know that transferring of a phosphate group is a common mechanism in the activation of intermediates of the calvin cycle that involves in carbon fixation process. Hence, ORFA and ORFB may be involved in a metabolic pathway that manages the dynamics of phosphate utilization. Moreover, sequence comparison shows that ORFA and ORFB are homologous to ethanolamine utilization protein *cchB*, suggesting these proteins are not unique to cyanobacteria. Based on the information, it is likely that ORFA and ORFB plays a role to process the product/intermediate that contains multiphosphate groups in the Calvin cycle either upstream or downstream of RuBisCo.

*or459* in *Prochlorococcus* sp. MIT9313 and *or1180* in *Prochlorococcus* sp. MED4 are predicted to adopt the template 1dcp (pterin-4a-carbinolamine dehydratase) with the best confidence among all the genes in Fig. 5 (other than *rbcL* and *rbcS*), with Zscores 9.9 and 10.5 respectively. The ortholog of this gene in *Synechococcus* sp. WH8102 (*or3010*, separated by 3 genes from *or3006*) is also not far from the genes in the genomic sequence shown in Fig. 5. Since the z-score of these predictions are high, it is very likely that this ortholog has a function similar to a dehydratase. We know that pterin-4a-carbinolamine dehydratase catalyzes a dehydration reaction and such a reaction does not exist in the Calvin cycle. Therefore, we predict that *or459* acts as an enzyme to assist a reaction of RuBisCo.

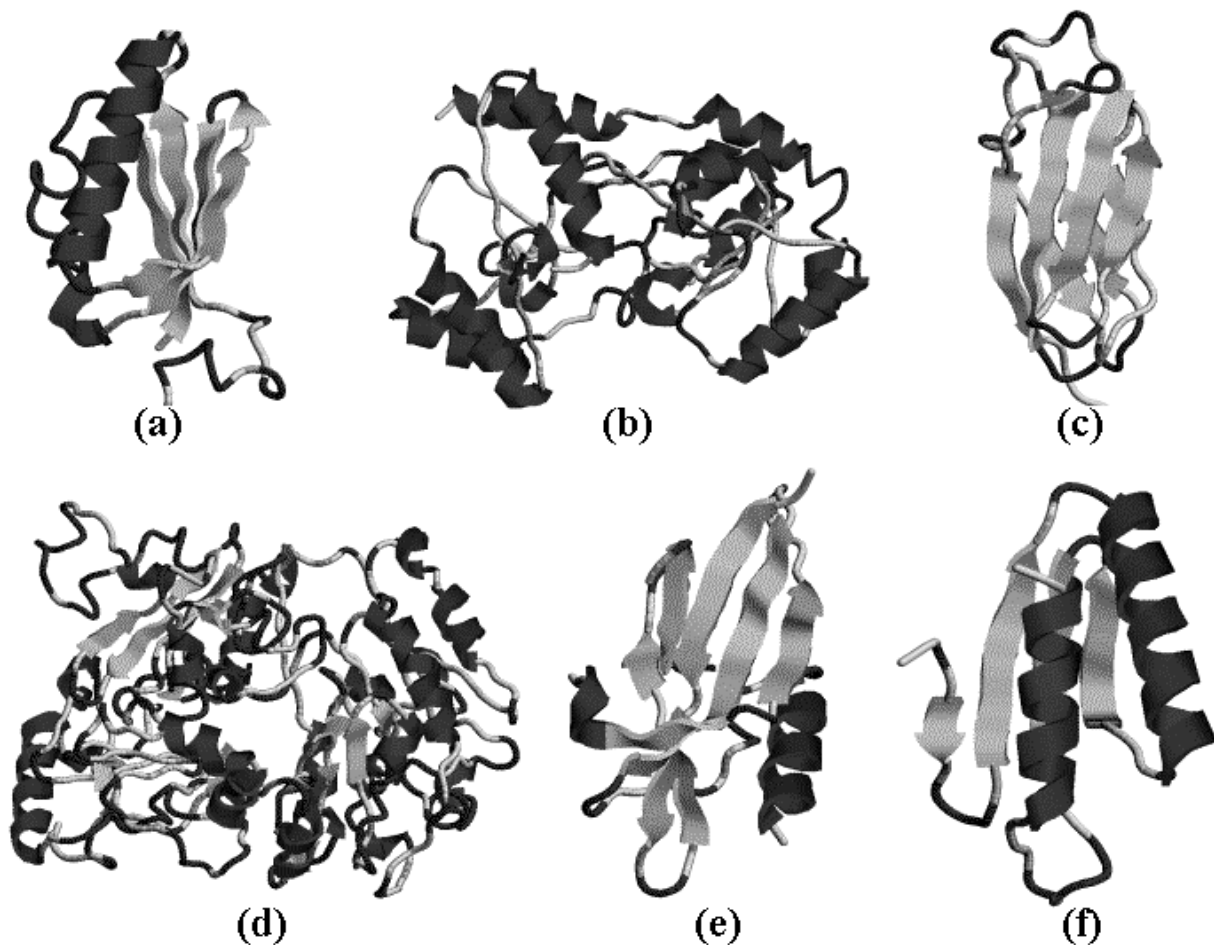


Figure 6: Predicted structures for (a) *ccmk1*, (b) *csoS2* N-terminal domain, (c) *csoS2* C-terminal domain (d) *csoS3*, (e) ORFA, and (f) Or459 (*Prochlorococcus* sp. MIT9313). (a-e) are 5 ORFs from *Synechococcus* sp. WH8102. (c) shows only one of the repetitive domains of the predicted structure.

## 5. Discussion

The PROSPECT pipeline, presented in this paper, provides a comprehensive toolkit for characterization of protein structures and functions. Its capability for assessing its prediction reliability, linking multiple tools in a seamless fashion, and doing the predictions in an efficient manner has made it possible to make structure/function predictions at genome scale. The automated prediction results provide a good rapid first-pass prediction for most of the genes in a microbial genome. The prediction results on three genomes suggest that ~80% of the genes in a microbial genome can be given some level of functional assignment, significantly higher than sequence-based genome annotation methods. The advantage of the prediction

pipeline over sequence-based approaches is not only in the higher number of genes that can be provided with functional assignments, but also due to the significant additional information that can be used to guide experiments, potentially speeding up the process of protein structure and function studies.

This pipeline can be used for a variety of purposes. One can do global analysis of a genome or to compare different genomes, as we have done here for three cyanobacterial genomes. Typically, it was found that by analyzing the top 100 predicted structural folds for each ORF, even in cases with no high z-scores, it was possible to discover a great amount of information, as was demonstrated for the case study of carboxysome. Experience suggests that over 90% of the predicted genes in a microbial genome could have some level of structural and functional assignment, when using the pipeline in conjunction with manual predictions by an expert predictor. A byproduct of the genome-scale structure prediction is that it could be used by structural genomics centers to help prioritize their target lists for experimental structure determination. This will allow them to focus on proteins with novel folds.

As we have demonstrated, a significant portion of our manual prediction procedure employed in this work could be automated, which should further reduce the amount of human involvement in structure-based genome annotation. Work will continue towards building a more comprehensive expert system that employs a large number of decision-making rules based on human expert predictors.

This early attempt at genome-scale structural/functional assignments of predicted genes demonstrates that it is possible to significantly expand the scope of structurally and functionally assignable genes, using threading-based method in conjunction with a rule-based system and on top of sequence-based methods like PSI-BLAST. Further research can push the application of such capability in two directions. (a) More detailed analysis and comparison of structural fold distributions across multiple genomes could provide a basis for new insights about the differences between different genomes and the niche different organisms exploit. (b) More detailed analysis of the reliably predicted structural folds (with high z-scores) could possibly reveal greater detail about functional mechanisms of individual proteins and provide a foundation for understanding the role of individual proteins in molecular complexes or protein interaction networks. Such advanced capabilities for mining the structural fold space in a genome should complement the existing, mainly sequence-based, methods to make the computational tools more powerful for addressing challenging biological questions in the post genome-sequencing era.

## **Acknowledgments**

We thank our ORNL colleagues Frank Larimer, Loren Hauser, Miriam Land, Sergei Passovets, Kyle Ellrott, Philip LoCascio, Li Wang, Inna Vokler, and Dr. Brian Palenik of UCSD for helpful discussions. This research was sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under Contract No. DE-AC05-000R22725 managed by UT-Battelle, LLC. This work was also funded in part by the US Department of Energy's Genomes to Life program ([www.doegenomestolife.org](http://www.doegenomestolife.org)) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling" ([www.genomes-to-life.org](http://www.genomes-to-life.org)). The authors also thank Dr. Al Geist of ORNL for providing us the full access to the XTORC Linux cluster for running the PROSPECT pipeline on the three cyanobacterial genomes.

## References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389-3402.
2. Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J., Wright, W. (1999) PRINTS prepares for the new millennium. *Nucleic Acids Research*. 27, 220- 225.
3. Bairoch A (1993) The ENZYME data bank. *Nucleic Acids Research*. 21, 3155-3156.
4. Baker, S. H., Lorbach, S. C., Rodriguez-Buey, M., et al. (1999). The correlation of the gene *csoS2* of the carboxysome operon with two polypeptides of the carboxysome in *thiobacillus neapolitanus*. *Arch Microbiol* 172(4): 233-9.
5. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. L. (2002). The Pfam protein families database. *Nucleic Acids Research*. 30, 276-280.
6. Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 253, 164-170.
7. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. (1999) Swaminathan S. Structural genomics: beyond the human genome project. *Nat Genet*. 23, 151-157.
8. Cannon G.C., Bradburne, C.E., Aldrich, H.C., Baker, S.H., Heinhorst, S. and Shively, J.M. (2001) Micocompartments in Prokaryotes: Carboxysomes and Related Polyhedra. *Appl. Env. Microbiol.* 67, 5351-5361.
9. CASP. (1995) Protein Structure Prediction Issue. *Proteins: Struct. Funct. Genet.* 23, 295-462.
10. CASP. (1997) Protein Structure Prediction Issue, *Proteins: Struct. Funct. Genet*, Suppl. 1. 29, 1-230.
11. CASP. (1999) Protein Structure Prediction Issue. *Proteins: Struct. Funct. Genet*. Suppl. 3. 37, 1-237.
12. CASP. (2001) Protein Structure Prediction Issue. *Proteins: Struct. Funct. Genet*. Suppl. 4. 35, 1-199.

13. CASP. (2002) CASP5 Proceedings.
14. Chance, M. R., Bresnick, A.R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A., Almo, S. C., Bonanno, J. B., Buglino, J. A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M., Wang, L. K. (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.* 11, 723-38.
15. Chisholm, S.W. (1992) Phytoplankton size. Primary productivity and biogeochemical cycles. P.G. Falkowski and A.D. Woodhead. New York, Plenum Press: 213-237.
16. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* 282, 699-705.
17. 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.
18. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-6.
19. Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-6.
20. Henikoff, J. G., Henikoff, S., Pietrokovski, S. (1999) New features of the blocks database servers. *Nucleic Acids Research.* 27, 226-228.
21. Higgins D., T. J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
22. Hirokawa, T., Boon-Cheing, S., and Mitaku, S. (1998) Classification and secondary structure prediction system for membrane proteins, *Bioinformatics.* 14, 378-379.
23. Hofmann, K., Bucher, P., Falquet, L., Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Research.* 27, 215-219.
24. Honig, B. Nicholls, A. (1995) Classical electrostatics in biology and chemistry. *Science.* 268, 1144-1149.
25. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature.* 358, 86-89.
26. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *Mol. Biol.* 5-202.
27. Karplus, K., Barrett, C., Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics.* 14, 846-856.
28. Kim, D., Xu, D., Guo, J., Ellrott, K., Xu, Y. (2002) PROSPECT II: protein structure prediction program for genome-scale application, submitted.

29. Kovacs, E., van der Vies, S. M., Glatz, A., Torok, Z., Varvasovszki, V., Horvath, I., Vigh, L. (2001) The chaperonins of *Synechocystis* PCC 6803 differ in heat inducibility and chaperone activity. *Biochem Biophys Res Commun.* 289, 908-15.
30. Lander, E. C., et al. (2001) Initial sequencing and analysis of the human genome, *Nature.* 409:860-921.
31. Leahy, D. J., Erickson, H. P., Aukhil, I., et al. (1994). Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. *Proteins* 19(1): 48-54.
32. Li, Z., Scheraga, H. A. (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA.* 84, 6611-6615.
33. Lupas, A., Van Dyke, M., Stock, J. (1991) Predicting Coiled Coils from Protein Sequences, *Science.* 252, 1162-1164.
34. Miyazawa, S. and Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256, 623-644.
35. Montelione, G. T. and Anderson, S. (1999) Structural genomics: keynote for a Human Proteome Project. *Nature Struct. Biol.* 6, 11-12.
36. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
37. Nakai, K. Horton, P. (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends in Biochemical Sciences.* 24, 34-35.
38. Nicholls, A., Sharp, K. A., Honig, B. (1991) Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Structure, Function and Genetics.* 11, 281-296.
39. Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10:1-6.
40. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. (2001) Gerstein M. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol.* 314, 1053-66.
41. Orus, M. I., Rodriguez, M. L., Martinez, F. and Marco, E. (1995). Biogenesis and Ultrastructure of Carboxysomes from Wild Type and Mutants of *Synechococcus* sp. Strain PCC 7942. *Plant Physiol* 107, 1159-1166.
42. Partensky, F., W. R. Hess, and D. Vaultot. (1999) Prochlorococcus, a marine photosynthetic prokaryote of global significance *Microbiology and Molecular Biology Reviews.* 63, 106-127.

43. Price, G. D., Howitt, S. M., Harrison, K. and Badger, M. R. (1993). Analysis of a genomic DNA region from the cyanobacterium *Synechococcus* sp. strain PCC7942 involved in carboxysome assembly and function. *J Bacteriol.* 175, 2871-9.
44. Sali, A. Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815.
45. Skolnick, J., Kolinski, A. (1991) Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol.* 221, 499-531.
46. Steward, A., Adhya, S. and Clarke, J. (2002). Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J Mol Biol.* 318, 935-40.
47. Tabita, F. R., Gibson, J. L., Falcone, D. L., Lee, B. G., Chen, J. H. (1990) Recent studies on the molecular biology and biochemistry of CO<sub>2</sub> fixation in phototrophic bacteria. *FEMS Microbiol Rev.* 7, 437-43.
48. Venter, V. C., et al. (2001) The sequence of the human genome, *Science.* 291, 1304-1351.
49. Vriend, G. (1990) HATIF: a molecular modelling and drug design program, *J. Mol. Graphics.* 8, 52-56.
50. Vakser, I. A. (1996) Low-resolution docking: Prediction of complexes for underdetermined structures. *Biopolymers.* 39, 455-464.
51. Wallace, A. C., Laskowski, R. A., Thornton, J. M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Science.* 5, 1001-1013.
52. Wetlaufer, D. B. (1978). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA,* 70, 697-701.
53. Westbrook, J., et al. (2000) The Protein Data Bank: unifying the archive, *Nucleic Acids Res.* 30, 245-248.
54. Xiang Z, Honig B, Extending the accuracy limit of side-chain prediction, *J. Mol. Biol.* 2001, 311:421-430
55. Xu, Y. and Xu, D. (2000) Protein Threading using PROSPECT: design and evaluation, *Protein: Structure, Function, Genetics.* 40, 343 - 354.
56. Xu, D. and Xu, Y. (2002) Computational Studies of Protein Structure and Function Using Threading Program PROSPECT. In *Protein Structure Prediction: Bioinformatic Approach*, edited by Igor Tsigelny. International University Line publishers (IUL), La Jolla, CA. Pages 5-41.
57. Xu, D., Crawford, O. H., LoCasio, P. F., and Xu, Y. (2001a) Application of PROSPECT in CASP4: Characterizing Protein Structures with New Folds. *Proteins: Structure, Function, and Genetics (CASP4 Special Issue).* 46, 140-148.

58. Xu, D., Baburaj, K., Peterson, C. B., and Xu, Y. (2001b) A Model for the Three Dimensional Structure of Vitronectin: Predictions for the Multi-Domain Protein from Threading and Docking. *Proteins: Structure, Function, Genetics*. 44, 312-320.
59. Xu, Y., Xu, D., Crawford, O. H., Einstein, J. R., Larimer, F., Uberbacher, E. C., Unseren, M. A., and Zhang, G. (1999) Protein threading by PROSPECT: a prediction experiment in CASP3. *Protein Engineering*. 12, 899-907.
60. Zhu, G., Spellman, P. T., Volpe, T., Brown, P. O., Botstein, D., Davis, T. N., and Futcher, B. (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* 406, 90-4.
61. Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J., Godzik, A. (1999) From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Science*. 8, 1104-1115.