A Web-Accessible Protein Structure Prediction Pipeline

Michael S. Lee
Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Ft. Detrick, MD;US Army Research Laboratory, Computational and Information Sciences Directorate, Aberdeen Proving Ground, MD
michael.lee@amedd.army.mil

Rajkumar Bondugula, Valmik Desai, Nela Zavaljevski, In-Chul Yeh, Anders Wallqvist, and Jaques Reifman
US Army Medical Research and Materiel Command (MRMC), Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, Ft. Detrick, MD
{raj, valmik, nelaz, icy, awallqvist, jreifman}@bioanalysis.org

Abstract

Proteins are the molecular basis of nearly all structural, catalytic, sensory, and regulatory functions in living organisms. The biological function of a protein is inextricably linked to its three-dimensional (3D) atomic structure. Traditional structure determination methods, such as X-ray and nuclear magnetic resonance techniques, are time-consuming, expensive, and infeasible for the millions of proteins that have been sequenced so far from various organisms. Alternatively, computational structure prediction methods provide a faster and more cost-effective, albeit approximate, alternative to experimental structure determination. We present a high-throughput protein structure prediction pipeline (dubbed "PSPP"), which given input protein sequences infers their 3D atomic structures. The pipeline was designed to be used with high performance computing clusters and to scale with the number of processors. The pipeline encompasses a core Perl module, a parallel job manager, and a Web browser graphical user interface accessible at our Website (www.bhsai.org). The software is currently installed at the Department of Defense (DoD) Maui High Performance Computing Center, and it is available for download along with its associated databases from our site. Currently, DoD scientists are using the pipeline in basic science and drug and vaccine development projects.

1. Introduction

Proteins play a major role in many structural, catalytic, sensory, and regulatory functions in living organisms. The structure of a protein is essential in understanding its function at the molecular level. Characterizing sequence-structure and structure-function relationships have been the goals of molecular biology for more than three decades. Traditional structure determination methods, such as X-ray crystallography and nuclear magnetic resonance techniques, are time-consuming, expensive, and infeasible for the millions of proteins that have been sequenced from various organisms. Predicting the structure of a protein computationally is an attractive alternative to experimental methods. The predicted structures are useful for classification, function annotation, binding interface description, binding partner prediction, and structure-based design of drugs and vaccines.

Computational protein structure prediction is a complex multi-step process that requires many tools and expertise to achieve the final goal of protein structure prediction. Many Web servers with applications exist that predict the structure of a given protein sequence. However, depending on publicly available servers is not practical for many reasons: they are shared resources with limited access, data confidentiality cannot be assured, and there is no assurance that the Web servers will be maintained in the future with appropriate upgrades to the latest available software and databases.

We introduce a software pipeline called Protein Structure Prediction Pipeline (PSPP) to automatically perform the multi-step protein structure prediction process. The PSPP consists of more than 20 individual

978-0-7695-3946-1/10 $26.00 © 2010 IEEE
DOI 10.1109/HPCMP-UGC.2009.29

software programs and databases that predict various aspects of protein structures of individual proteins. The PSPP can perform three types of structure prediction: comparative (homology) modeling, fold recognition, and *ab initio* fragment assembly. The pipeline encompasses a core Perl module that integrates the different software components, a parallel job manager, and a Web browser graphical user interface (GUI). The parallel job manager distributes the computational work over multiple processors. Users can submit structure prediction jobs via a secure Web browser GUI, which is based on the User Interface Toolkit (UIT)[1]. The results can be viewed either in the GUI or can be downloaded for viewing in text editors, spreadsheet applications, or Web browsers.

2. Methods

PSPP contains many freely available modules from several laboratories and some in-house components for predicting protein structure and other properties. Figure 1 depicts the workflow of the pipeline. The structural properties predicted by the PSPP include: protein secondary structure (short regions of regular structure), calculated by PSIPRED[2], SSPro[3], and MUPRED[4] programs; solvent accessibility (the extent to which solvent can access the surface of the protein) computed by ACCPro[3] and MUPRED[5] programs; transmembrane region (describes whether the protein spans the cell membrane, interior or exterior of the cell) inferred by the TMHMM[6] program and disordered regions (lack of well-formed structures) computed by Dispro[7] program. For almost all of these programs, the input is a protein position-specific substitution matrix (PSSM), in addition to the protein sequence itself. We calculate a single PSSM for each of the input proteins using PSI-BLAST[8] and the NR database. Once the structural properties are predicted using the input sequence, the proteins are divided into domains for tertiary structure prediction. The domain boundaries are determined using the FIEFDom[9] program. Optionally, Bayesian statistics can be applied in addition to the FIEFDom program to guarantee that domains do not exceed a length of 250 amino acids.

If the input domain sequence shares a high sequence similarity with other proteins with known structures, then homology modeling is used for tertiary structure prediction. The proteins related to the input protein are identified using the sequence alignment program, PSI-BLAST. If the input domain sequence is remotely related to other proteins and the relationship is not readily identifiable through programs like PSI-BLAST, then threading is used to establish the relationship, as seen in the flowchart in Figure 1. We use the PROSPECTII[10] component of the PSPP to identify remotely related proteins of the input proteins. Once the related proteins with known structures are identified, their crystal structures are used as templates by the NEST[11] program in the PSPP to build three-dimensional (3D) atomic structural models of the input protein. Unlike most other homology prediction servers, the predicted structures are scored and ranked using two different scoring schemes. The first scoring program is an in-house implementation of the DFIRE-AA[12], an all-atom statistical potential derived from analysis of the inter-atomic distances between pairs of atoms types in a large set of known protein structures. The second scoring module involves minimizing the structures using CHARMM/PARAM22[13] followed by scoring with the PARAM22 force field plus the GBMV2[14,15] implicit solvent potential.

If the input domain sequence is either not related to any other protein with known structures or the relationship cannot be established either by sequence alignment or by the PROSPECTII threading procedure, then a computationally intensive *ab initio* procedure must be used for structure prediction. PSPP incorporates Rosetta[16], a popular *ab initio* structure prediction program that employs fragment assembly. The program builds several backbone-only models from protein fragments of 3 and 9 amino acids in length. Then, the SCWRL[17] program is used to build side chains onto each of the backbone-only models output from Rosetta. These full-atom models are then ranked by DFIRE-AA score. The top-scoring structures are further ranked by the GBMV2 score, using the procedure described previously. Finally, the top-scoring models from this round are structurally compared against the SCOP[18] using the combinatorial extension (CE)[19] program. The goal of this last step is to annotate the best Rosetta models in terms of known protein fold types. While some proteins truly have unique folds never before observed experimentally, others may be distant homologues to a known fold that could not have been inferred by sequence alone.

We have made the pipeline available to Department of Defense (DoD) scientists by deploying the software onto the Maui DSRC/Jaws computing cluster. The pipeline can be accessed through a Web-based GUI (http://www.bhsai.org) implemented as a Web application using a variety of state-of-the-art software and libraries, including Java, J2EE, JavaServer Faces (JSF), ICEfaces, asynchronous JavaScript (AJAX), and XML. The Web application consists of server-side Java codes that use JSF and AJAX-based application programming interface (API) from ICEfaces. The Web application is deployed on an Apache Tomcat sever and uses hypertext transfer protocol over a secure socket layer connection for encrypting all of the data flowing to and from the user's

Web browser. The GUI uses UIT, a DoD-sponsored API, to allow authorized personnel to communicate and access High Performance Computing Modernization Program computational resources by verifying their credentials via SecurID-based Kerberos authentication tools. In addition to credential validation, the GUI makes it easy for the user to specify job-specific parameters, submit jobs, check the status of jobs, and analyze the results. The results of the predictions, both annotations and structural models, can be downloaded. The annotations are available in tab-delimited and HTML formats. In addition to the downloadable files, the results are also presented as a table in the GUI that can be sorted by various criteria including model rank and model energy. Finally, results can be searched by keyword. For example, the annotated results of a multiple sequence run can be searched to find the proteins that were predicted to be of a particular fold-type. Figure 2 depicts a typical screenshot of the GUI presenting the results of a protein structure prediction run.

3. Results

In this section, we first briefly describe how the results of a run are depicted in the GUI. Second, we discuss the scaling tendencies of the pipeline with the number of available processors. Next, we discuss structure prediction of egg white protein lysozyme[20] (Protein Data Bank ID: 2vb1) using our pipeline and comment on template selection based on energy. Finally, we briefly outline the current applications of the pipeline undertaken by various DoD life science laboratories.

GUI: In the results screen of the Web server GUI, as seen in Figure 2, all of the jobs submitted by the user are available through a drop-down list on the top. When a particular job is selected, the results are loaded. From the input sequence tree on the left side, sequence property results for each protein can be obtained by mouse-clicking on the name. Selecting the "+" button lists each domain for that protein. Selecting a specific domain leads to the tabbed windows shown in the right region. The results of either comparative modeling or fold recognition can be viewed by selecting the respective tabs. With each modeling category, the results can be sorted based on any column displayed in the results table. The homology modeling component displays identity rank, % sequence identity to the Protein Data Bank (PDB) template, number of aligned residues, number of identical residues, number of positive matches, number of gaps in the alignment, and title of the PDB entry. The fold recognition component shows confidence rank, SCOP template ID, % identity to template, SCOP fold family ID, and SCOP fold description. Finally, the *ab initio* component (not shown) lists Z-score rank, SCOP template ID, model score, SCOP fold family ID, and SCOP fold description.

Scaling: Different components of the pipeline have different computational costs associated with them: homology modeling typically takes roughly 1 processor-hour/structure, whereas the *ab initio* folding can take up to 100 processor- hours for a single 150 amino-acid-residue protein. The computationally-intensive *ab initio* component benefits greatly from the use of multiple processors. Figure 3 illustrates the computational speedup for generating and scoring Rosetta models as a function of the number of processors using an 8-processor job as the reference. The plot shows that Rosetta scales linearly with the number of processors up to 64 processors. In addition to slave processors, one processor is dedicated by the parallel Rosetta and scoring module job managers to monitor the job status and distribute tasks to slave processors.

The homology modeling and fold recognition modules also retain good scaling performance up to 64 processors (results not shown) with the caveat that the fold database must be copied to the local file system of the computing node. The problem is that the PROSPECTII fold recognition program performs thousands of file open/close operations per domain query, which degrades the performance of the shared file system if multiple domains are processed simultaneously. Since use of the disk on the computing node of the cluster is not always feasible, another solution will be pursued in the future. Specifically, we will modify a new open-source variant of the PROSPECTII program, which will load the fold database in large blocks rather than one fold template file at-a-time.

Analysis: One of the novel features of the PSPP is the evaluation of homology and fold recognition structural models by physical and statistical scoring functions. As an example, we predict the structure of hen egg white lysozyme. Traditionally, sequence identity of the query to the template is used to determine the suitability of a comparative model as seen in Figure 4a. However, this metric is not always useful. As seen in Figure 4b, the DFIRE-AA scoring function provides a better scoring criteria with respect to actual accuracy as measured by root mean squared deviation (RMSD) of the alpha-carbon trace of the comparative model to the known structure. Suppose, for example, that the only available homologous templates were the 97 structures of human lysozyme protein. All of these templates have roughly the same sequence identity (58 to 61%) with respect to hen egg white lysozyme. However, within the human template subset, the DFIRE-AA score correlates with model accuracy and thus provides a reliable measure for selecting a near-optimal model. In this hypothetical case, the template with the lowest DFIRE energy is PDB structure 1GAZ, which results in a predicted model that is 0.79 Å RMSD from the native structure of the hen egg white protein, as seen in Figure 5. While the best model one could have selected produces a 0.73 Å RMSD, other choices of human lysozyme templates could have resulted in predicted structures as far away as 2.67 Å RMSD from the native structure.

Applications: Researchers at the US Army Medical Research Institute of Infectious Diseases are using the PSPP to determine structures of proteins encoded by viral and bacterial genomes. These structures will be used to screen for small molecule inhibitors that can be developed into drugs. For example, in a structural genomic application of the PSPP, the variola (smallpox) genome consisting of 197 protein-coding genes was successfully annotated using homology modeling in less than six hours using 64 processors. High-quality structure predictions suitable for drug design were generated for about 10% of these proteins. Furthermore, in collaboration with the Walter Reed Army Institute of Research, the Biotechnology HPC Software Applications Institute is using the PSPP to characterize the structure of single-protein malaria vaccine candidates.

4. Conclusion

We have developed a Perl-based pipeline for protein structure prediction that integrates freely downloadable software components from various academic and government research laboratories. The pipeline is designed for deployment on high-performance computing clusters. It includes all-atom scoring, structural annotations of *ab initio* models and includes a GUI that facilitates user authentication, parameter specification, job submission, job monitoring, and access to results. The system has been applied to support multiple biodefense-related projects sponsored by the DoD Defense Threat Reduction Agency.

Acknowledgments

This work was sponsored by the US Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes Initiative.

Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This abstract has been approved for public release with unlimited distribution.

References

1. Monceaux, W., K. Rappold, P. Duett, S. Swillie, and R.S. Maier, "UIT Development: How to Access HPC Resources Using the UIT Web Service." *HPCMP Users Group Conference*, Denver, CO, 2006.

2. Jones, D.T., "Protein secondary structure prediction based on position-specific scoring matrices." *J Mol Biol,* vol. 292, pp. 195–202, Sep. 17, 1999.

3. Cheng, J., A.Z. Randall, M.J. Sweredoski, and P. Baldi, "SCRATCH: a protein structure and structural feature prediction server." *Nucleic Acids Res,* vol. 33, pp. W72-6, Jul. 1, 2005.

4. Bondugula, R. and D. Xu, "MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction." *Proteins,* vol. 66, pp. 664–70, Feb. 15, 2007.

5. Bondugula, R. and D. Xu, "Combining Sequence and Structural Profiles for Protein Solvent Accessibility Prediction." in *Computational Systems Bioinformatics, CSB2008 Conference Proceedings*, Stanford, CA, USA, pp. 195–202, 2008.

6. Viklund, H. and A. Elofsson, "Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information." *Protein Sci,* vol. 13, pp. 1908–17, Jul. 2004.

7. Cheng, J., M. Sweredoski, and P. Baldi, "Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data." *Data Mining and Knowledge Discovery,* vol. 11, pp. 213–222, 2005.

8. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res,* vol. 25, pp. 3389–402, Sep. 1, 1997.

9. Bondugula, R., M.S. Lee, and A. Wallqvist, "FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator." *Nucleic Acids Res,* vol. 37, pp. 452–62, Feb. 2009.

10. Kim, D., D. Xu, J.T. Guo, K. Ellrott, and Y. Xu, "PROSPECT II: protein structure prediction program for genome-scale applications." *Protein Eng,* vol. 16, pp. 641–50, Sep. 2003.

11. Petrey, D., Z. Xiang, C.L. Tang, L. Xie, M. Gimpelev, T. Mitros, C.S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I.Y. Koh, E. Alexov, and B. Honig, "Using multiple structure alignments, fast model building, and

energetic analysis in fold recognition and homology modeling." *Proteins,* vol. 53, Suppl 6, pp. 430–5, 2003.

12. Zhou, H. and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction." *Protein Sci,* vol. 11, pp. 2714–26, Nov. 2002.

13. Brooks, B.R., R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminatham, and M. Karplus, "CHARMm: A program for macromolecular energy, minimization, and dynamics calculations." *J. Comp. Chem.,* vol. 4, p. 187, 1983.

14. Lee, M.S., M. Feig, F.R. Salsbury, Jr., and C.L. Brooks, 3rd, "New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations." *J Comput Chem,* vol. 24, pp. 1348–56, Aug. 2003.

15. Lee, M.S. and M.A. Olson, "Assessment of Detection and Refinement Strategies for de novo Protein Structures Using Force Field and Statistical Potentials." *J Chemical Theory and Computation,* vol. 3, pp. 312–324, 2007.

16. Simons, K.T., C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." *J Mol Biol,* vol. 268, pp. 209–25, Apr. 25, 1997.

17. Bower, M.J., F.E. Cohen, and R.L. Dunbrack, Jr., "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool." *J Mol Biol,* vol. 267, pp. 1268–82, Apr. 18, 1997.

18. Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *J Mol Biol,* vol. 247, pp. 536–40, Apr. 7, 1995.

19. Shindyalov, I.N. and P.E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." *Protein Eng,* vol. 11, pp. 739–47, Sep. 1998.

20. Soundararajan, M., F.S. Willard, A.J. Kimple, A.P. Turnbull, L.J. Ball, G.A. Schoch, C. Gileadi, O.Y. Fedorov, E.F. Dowler, V.A. Higman, S.Q. Hutsell, M. Sundstrom, D.A. Doyle, and D.P. Siderovski, "Structural diversity in the RGS domain and its interaction with heterotrimeric G protein alpha-subunits." *Proc Natl Acad Sci U S A,* vol. 105, pp. 6457–62, Apr. 29, 2008.

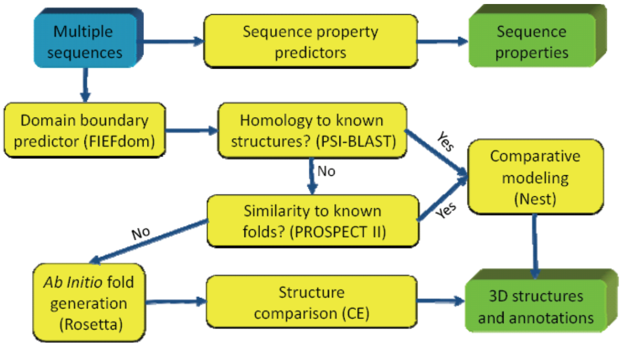
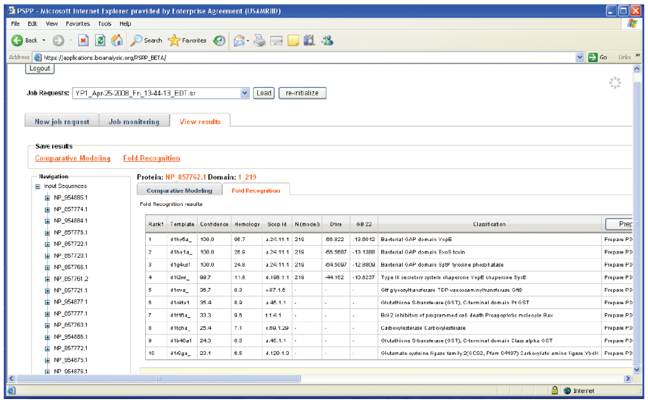

Multiple sequences
Sequence property predictors
Sequence properties
Domain boundary predictor (FIEFdom)
Homology to known structures? (PSI-BLAST)
Yes
No
Comparative modeling (Nest)
Similarity to known folds? (PROSPECT II)
Yes
No
Ab Initio fold generation (Rosetta)
Structure comparison (CE)
3D structures and annotations

Figure 1. Workflow for the protein structure prediction pipeline. First, 1-dimensional properties of the full protein sequence are predicted. Then, the sequence is divided into domains and routed through three possible tertiary prediction schemes: comparative modeling (PSI-BLAST), fold recognition (PROSPECT II), and *ab initio* fragment assembly (Rosetta). Box color legend: blue–*input sequences,* yellow–*pipeline processes,* green–*output*.



https://applications.bioanalysis.org/PSPP_BETA/

Job Requests: YP1_Apr-25-2008_Fri_13-44-13_EDT.sr

Save results

Comparative Modeling Fold Recognition

Navigation

Input Sequences
- NP_954885.1
- NP_857774.1
- NP_954884.1
- NP_857776.1
- NP_857722.1
- NP_857720.1
- NP_857768.1
- NP_857761.2
- NP_857721.1
- NP_954877.1
- NP_857777.1
- NP_857763.1
- NP_954886.1
- NP_857772.1
- NP_954875.1
- NP_954876.1

Protein: NP_857762.1 Domain: 1_219

Fold Recognition results

Rank1	Template	Confidence	Homology	Scop Id	N(model)	Dfire	GB 22	Classification	Prep
1	d1hy5a_	100.0	96.7	a.24.11.1	219	-68.822	-13.8012	Bacterial GAP domain YopE	Prepare PD
2	d1he1a_	100.0	26.9	a.24.11.1	219	-66.5687	-13.1388	Bacterial GAP domain ExoS toxin	Prepare PD
3	d1g4us1	100.0	24.8	a.24.11.1	219	-64.5097	-12.8809	Bacterial GAP domain SptP tyrosine phosphatase	Prepare PD
4	d1l2wi_	99.7	11.8	d.198.1.1	219	-44.102	-10.8237	Type III secretory system chaperone YopE chaperone SycE	Prepare PD
5	d1rrva_	36.7	9.3	c.87.1.5	-	-	-	Gtf glycosyltransferase TDP-vancosaminyltransferase GtfD	Prepare PD
6	d1okta1	35.4	8.9	a.45.1.1	-	-	-	Glutathione S-transferase (GST), C-terminal domain Pf GST	Prepare PD
7	d1f16a_	33.3	9.6	f.1.4.1	-	-	-	Bcl-2 inhibitors of programmed cell death Proapoptotic molecule Bax	Prepare PD
8	d1tqha_	25.4	7.1	c.69.1.29	-	-	-	Carboxylesterase Carboxylesterase	Prepare PD
9	d1b48a1	24.3	6.5	a.45.1.1	-	-	-	Glutathione S-transferase (GST), C-terminal domain Class alpha GST	Prepare PD
10	d1r8ga_	23.1	6.5	d.128.1.3	-	-	-	Glutamate-cysteine ligase family 2(GCS2, Pfam 04107) Carboxylate-amine ligase YbdK	Prepare PD

Figure 2. A typical results screen of the Protein Structure Prediction Pipeline Web-based Graphical User Interface

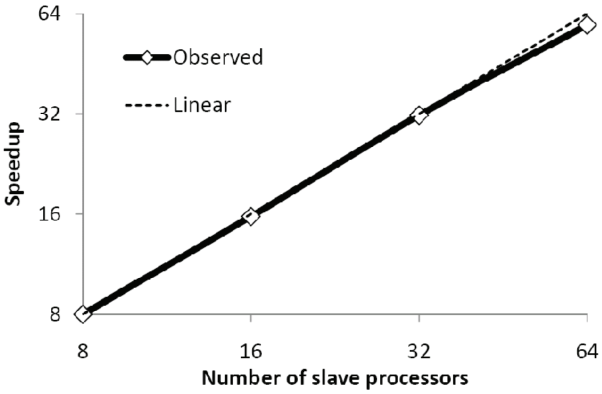


Figure 3. The speedup of Rosetta model generation and post-process scoring is directly proportional to the number of slave processors available to the pipeline. In addition to the slave processors, the pipeline employs one master processor to assign and monitor jobs. We have chosen to define the 8-processor run as the baseline; therefore, it is assigned an ideal 8-times speedup.

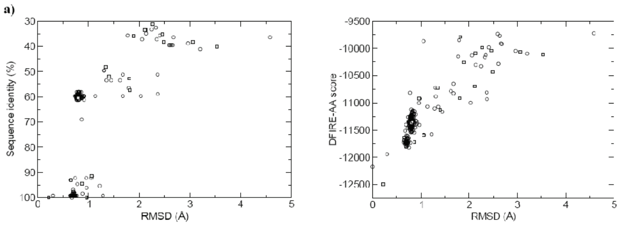


Figure 4. Evaluation of template-based models of hen egg white lysozyme[20] (Protein Data Bank ID: 2vb1) generated by comparative modeling (open circles) and fold recognition (grey squares) using two different discrimination tools: a) sequence similarity of template to query and b) all-atom statistical potential score (DFIRE-AA). RMSD refers to the root-mean squared deviation of the alpha-carbon trace between the model and the native structure.

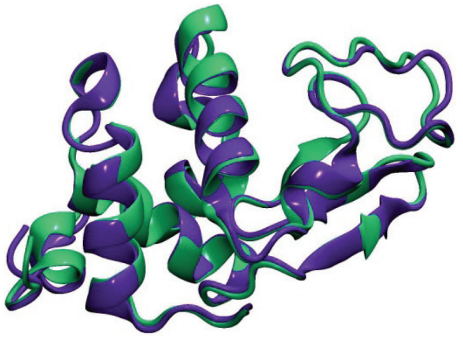


Figure 5. The lowest DFIRE-AA scoring homology model of hen egg white lysozyme built from a human lysozyme template. Predicted model (violet) is 0.79 Å RMSD from the native structure of hen lysozyme (green).