LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-209958

Protein Model Database

Krzysztof Fidelis, Alexei Adzhubej, Andriy Kryshtafovych, Pawel Daniluk

February 24, 2005

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Laboratory Directed Research and Development
ER

Protein Model Database

03-ERD-063

Krzysztof Fidelis (PI), Alexei Adzhubej, Andriy Kryshtafovych, Pawel Daniluk
Biology and Biotechnology Research Program
Lawrence Livermore National Laboratory
L-448, 7000 East Ave., P.O. Box 808, Livermore, CA 94551
Tel. (925) 423 4752, fax (925) 424 3130
fidelis@llnl.gov

Livermore, February 14, 2005

Abstract

The phenomenal success of the genome sequencing projects reveals the power of completeness in revolutionizing biological science. Currently it is possible to sequence entire organisms at a time, allowing for a systemic rather than fractional view of their organization and the various genome-encoded functions. There is an international plan to move towards a similar goal in the area of protein structure. This will not be achieved by experiment alone, but rather by a combination of efforts in crystallography, NMR spectroscopy, and computational modeling. Only a small fraction of structures are expected to be identified experimentally, the remainder to be modeled. Presently there is no organized infrastructure to critically evaluate and present these data to the biological community. The goal of the Protein Model Database project is to create such infrastructure, including (1) public database of theoretically derived protein structures; (2) reliable annotation of protein model quality, (3) novel structure analysis tools, and (4) access to the highest quality modeling techniques available.

This project was not completed.

First year accomplishments

The project was funded in May of 2003. This report covers the first year of funding. During this time we have essentially completed development of the new methods necessary to analyze, compare, and classify protein structures. We have also developed the relational database to store and query the data. Structure analysis techniques are described in **Section 1.**, the relational database system in **Section 2.**, and some organizational issues are covered in **Section 3.**

1. Novel techniques of structure analysis

We have developed three new analytical techniques necessary for an effective evaluation of models and for development of the estimators of model quality.

1.1. Basis for the implemented model analysis techniques

The most straightforward approach to the comparison of two protein structures is a rigid body superposition. When not carefully applied, however, this type of analysis is likely to produce misleading results, for example for multi-domain structures, where one domain is shifted relative to another. Also, a rigid body superposition will not produce satisfactory results for models characterized by a gradual deformation of one structure versus another. Instead we have developed a formalism based on local descriptors of protein structure, described in the following patent application and paper:

Fidelis, K., and Kryshtafovych, A. 2003. Local descriptors of protein structure. Patent application IL-10728, USA.

Hvidsten, T.R., Kryshtafovych, A., Komorowski, J., and Fidelis, K. 2003. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. Bioinformatics, 19 Suppl 2: 81-91.

The idea is to correctly identify similarities in the local topology by systematically comparing the residue-attached subsets of structure. Local subsets are designed to encompass the structure elements that are proximal with respect to the reference residue. An example of such a subset is shown in **Fig. 1.** This formalism constitutes the starting point for the structure analysis techniques described in sections **1.2.** – **1.4.**

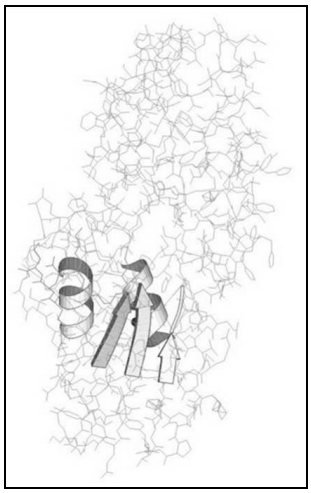

Fig. 1. Example of a residue-attached subset of structure as defined for residue 80 of cobalt chelatase (PDB code 1qgo, chain A). Residue 80 is represented by a black sphere partially obscured by the middle strand.

1.2. Model consensus algorithm

We expect that for many proteins, especially those with high impact on current research topics in biology, there will be more than one model available, provided either by separate research groups using comparable modeling techniques or by separate approaches altogether. It would then be highly desirable to perform a comparative analysis with the goal of obtaining a single structural representation that is better than any of the contributing results. Recent advances in structure prediction suggest that when multiple predictions are generated on any given target, the most abundant high-scoring models are closer to the native structure than the model with the highest score (e.g. Bonneau et al. 2002). This is the rationale behind the model consensus approach. Within the PMD system we expect to have obtained models from several research groups. These models in general would be associated with different reliability estimates. During the first year of this project we have developed a method to identify, in a set of models, all the similar local structures. The similarities are mapped onto the sequence of the modeled protein, and a consensus alignment generated. A single model is constructed based on the consensus local structure contributions, which are weighted by the estimated quality of the contributing models.

1.3. Detection of distant similarity

We have also moved to fill an important gap in the area of structure similarity detection, a methodology we need to address several of the model quality evaluation tasks we will address in FY05. The problem to reliably identify remote structural homology among proteins is well known and still not satisfactorily resolved (e.g. Lesk et al. 2001, Aloy et al., 2003). The issue arises in difficult to model cases, when for a modeled protein there is no straightforward relationship with an existing structure, i.e. no comparative modeling techniques can be used. The challenge is to recognize all the correct topological features in a model, even when the overall model quality is relatively poor. The challenge is also to recognize any evolutionary relationships between proteins for which we only have hypothetical structures and those for which full structural data are available. The commonly used rigid body superposition based techniques are not suitable for identifying remote homology among structures. Instead, we have used the descriptors of protein structure to research the similarity at the level of local structure. We have investigated the neighborhood of each amino acid in a protein for similarity to other protein models or structures. Geometry-based similarity relation for descriptors can be transitively extended using descriptors originating from other proteins (i.e. PDB structures). We have also developed a method of hierarchical clustering of protein structure descriptors using minimal cuts (Mirkin 1996, Sharan and Shamir 2000), and once the descriptors have been clustered used the obtained equivalence classes as a reference for the transitive extension of the relation of similarity. The results of this approach have been presented at the Pacific Symposium on Biocomputing:

> Daniluk, P., Kryshtafovych, A., Hvidsten, T. R., Komorowski, J. & Fidelis, K. 2004 Identifying structural similarity of proteins using local descriptors of protein structure. Pacific Symposium on Biocomputing, p.61, Hawaii, USA.

We plan to publish a paper describing the results of this work later this year. By evaluating a broader range of structure similarity, expressed at the level of elements of structure rather than entire domains, this method will allow us to more precisely assess the correctness of each model. It will also allow calibrating model quality estimators we plan to develop.

1.4 Automated classification of structure

Cataloguing protein structures according to their architecture and topology allows for better use of models through comparisons with existing structures and through identifying evolutionary relationships between proteins. In general, structure is much more conserved than sequence, implicating higher probability of detecting such evolutionary relationships when operating in the structure domain. From the point of view of this project, an automated classification of structure, especially at the level of local structure, is necessary for the subsequent evaluation of model quality. Specifically it will allow organizing the results according to structural families and, at the lower level, structural features. This in turn will be used in the development of model quality estimators that are specific to particular elements of structure. During the first year of the project we have developed a structure classification software package comprising modules which (a) identify local similarity between proteins, (b) assemble similar elements of structure along the protein

sequence, (c) identify similar folds and provide a measure of their relatedness, (d) allow visualization of the similarity and dissimilarity between structures. The results of this approach have been presented at the Pacific Symposium on Biocomputing:

Kryshtafovych, A., Hvidsten, T.R., Komorowski, J., Daniluk, P., Wilczynska, M., and Fidelis, K. 2004. Automated local structure based classification of protein folds. Pacific Symposium on Biocomputing, p. 98, Hawaii, USA.

2. Design and implementation of the database system

Knowledge of protein 3D structure is one of the key elements of the information spectrum ranging from genes to the molecular function they encode. It is however much more valuable if assessed as a part of a knowledge system imbedded in a relational database (RDB), with other elements of that system also accessible. An RDB system allows, when the data is loaded, to perform multiple complex searches needed for data analysis (e.g. assessment of model accuracy) without writing new software for each analysis task separately. The software is already there – it's the database management system, rendering savings on development tasks. We have constructed the technical foundations of the Protein Model Database in four steps: (1) Design of the database schema; (2) Design of the dictionary and dictionary extensions; (3) Parsing of the dictionary and loading it into the Dictionary Definition Language (DDL) database; and (4) Schema extraction from the DDL and creation of the model database. Since the initiation of the project we have completed steps 1-3 of the above plan. The data description language incorporates the macromolecular Crystallographic Information File (mmCIF) dictionary and is based on the Self Defining Text Archive (STAR) definitions. The dictionary contains definitions of data items in the STAR/mmCIF files and is used to store these data in the relational database modules. The PMD dictionary was developed to describe the many aspects of the protein structure modeling process and of the modeled structures. The main data categories include prediction and modeling techniques, model quality assessment, and structure classification assignments. Categories providing direct compatibility with protein sequence, metabolic pathway and protein structure (Protein Data Bank) databases have also been developed.

3. Protein Model Database staff and collaborations

We have recruited the PMD staff: Alexei Adzhubei, a leader in protein model database development and creator of the mmCIF dictionary now used by the Protein Data Bank (PDB) and Lukasz Szajkowski, a 2004 top graduate of the program in Biocomputing at the Poznan University of Technology, Poland. Alexei has been heavily involved in the development of the PMD prior to his arrival. During the first year some of the tasks were addressed by the Protein Structure Prediction Center staff, Andriy Kryshtafovych and Pawel Daniluk. Some of the local clustering work was performed as a no-cost collaboration with Prof. Komorowski's group at the Linnaeus Centre for Bioinformatics, Uppsala University, Sweden. We have also conducted talks regarding establishing a collaboration with the Andrej Sali group at the UC San Francisco and QB3. Dr. Sali is a key player in the area of large-scale protein modeling and his support for the PMD is invaluable.