 Nucleic Acids Research, 1999, Vol. 27, No. 1 **251–253**

The PRESAGE database for structural genomics

Steven E. Brenner*, Derren Barken and Michael Levitt

Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305–5126, USA

Received September 4, 1998; Revised and Accepted October 26, 1998

ABSTRACT

The PRESAGE database is a collaborative resource for structural genomics. It provides a database of proteins to which researchers add annotations indicating current experimental status, structural predictions and suggestions. The database is intended to enhance communication among structural genomics researchers and aid dissemination of their results. The PRESAGE database may be accessed at http://presage.stanford.edu/

BACKGROUND

Structural genomics projects attempt to provide an experimental structure or a good theoretical model for every protein in all completed genomes (1–5). This new field of research is thriving because of the powerful new computational approaches for reliable homology modeling, improved methods for experimental structure determination, and increased awareness of protein structure and its potential to elucidate function. For functionally characterized proteins, structure can aid understanding of mechanism and provide insights into mutations. Genomic sequencing is also revealing huge numbers of proteins whose function is unknown; in these cases, the power of structure to reveal distant evolutionary relationships provides a tool for identifying biochemical roles. More globally, understanding of the complete repertoire of protein structures in different organisms provides fascinating insights into molecular and organismal evolution.

Computational and experimental approaches each have critical roles in structural genomics projects. Experimental research provides essential information about a relatively small number of individual proteins, while computational approaches can expand that knowledge and apply it to the potentially large families of related proteins. In practice, computational approaches are first used to assign protein structures to genomic proteins whenever possible. The remaining proteins are clustered into families, and representatives from these families are selected for experimental characterization. The newly solved structures are compared with other proteins of known structure in classifications such as SCOP (6), CATH (7) or FSSP (8), to yield information about their evolution and thence about function.

Already, the field has made impressive gains. Experiments aimed at solving the structure of a member of a new family are set to double this year. Computational analyses have also flourished, with at least a dozen groups making structural assignments of one or more complete genomes.

Unfortunately, information in the field is highly fractured. Though the Protein Data Bank (PDB) (9) remains a reliable repository of solved structures, there has not been any coordination in the selection of new structures. This has led multiple groups to inadvertently begin studies on the same protein, even though there are more than enough important families to go around. Similarly, computational studies have often been performed in isolation, with researchers unaware of their colleagues' efforts or the details of their work. Worse, lack of consistent organization and repositories for these data have made these results virtually inaccessible to biologists outside the field.

The PRESAGE (Protein Resource Entailing Structural Annotation of Genomic Entities) database is intended to improve communication among structural genomics researchers, by providing a repository of capsule information about progress in the field. Further, as structural information approaches sequence data in its pervasiveness, PRESAGE will aid in the distribution of this knowledge to the biology research community.

DATABASE MODEL

The core of PRESAGE is a database of protein sequences (derived from SWISS-PROT plus TrEMBL; 10) with structural genomics annotations. Unlike curated databases such as SWISS-PROT, the authors of the database do not create and edit these annotations. Instead, any active structural genomics researcher may submit information. Original contributors retain full credit (or blame) for their annotations. To help ensure proper attribution, entries have links with information about the contributor, as well as optional links to relevant literature references and associated Web sites—which may themselves be databases.

The database will also provide annotated summary data and analyses. However, its main goal is to allow structural and computational biologists to contribute to structural genomics projects and to disseminate that information.

ANNOTATIONS

The fundamental unit of information in the PRESAGE database is an annotation, which is attached to a single protein sequence entry. At present, PRESAGE has two main classes of annotations, (i) experimental and (ii) prediction, with several subsidiary and additional varieties. In the near future we will also be adding annotations of family membership.

Every annotation records the name of the annotator, the date on which it was entered, and allows contributors to specify which region of the protein they are annotating. Annotations have details

*To whom correspondence should be addressed. Tel: +1 650 725 0754; Fax: +1 650 723 8464; Email: brenner@hyper.stanford.edu

specific to their class, and also permit free-text comments, listings of relevant papers with MEDLINE references, and links to other Web sites associated with the annotation.

Experimental

An experimental annotation indicates that a protein has been selected for structure determination and tracks the progress towards the solved structure. Like the NCBI/HUGO Human Genome Sequencing Index (http://www.ncbi.nlm.nih.gov/HUGO/) that records sequencing efforts, this information will help coordinate research on structure. Principally it is intended to prevent inadvertent overlapping studies, but it also provides a view onto what structural data will become available in the future.

Experimental annotators record the stages their experiments have reached and specific details associated with those stages (e.g., the method of structure determination or the organism used for cloning).

Prediction

Computational biologists can register predicted structures for proteins, at three levels of detail. The simplest is 'assignment', which associates a region of the sequence with a known structure, and asserts that the two proteins will share a common fold. An 'alignment' prediction augments this information by indicating how the database sequence maps onto the solved structure, and a 'model' further provides predicted three-dimensional coordinates for the protein sequence. Comparison between groups' annotations could reveal strengths and weaknesses of different methods and provide additional background to users of the data.

Annotation records for all classes of predictions include the matched PDB entry and details about the regions whose folds are believed to be common. Storing alignment and model data is the responsibility of the annotators, who keep the data on their own Web sites. This introduces potential data integrity problems, as the information on an annotator's site may not match that which was available when the annotation was made. In order to alleviate this difficulty, the PRESAGE database obtains an MD5 message digest of the alignment and/or model at the time of annotation. The MD5 message digest (defined in RFC 1321; available from http://info.internet.isi.edu/in-notes/rfc/files/rfc1321.txt) is a cryptographic technique which condenses the content of these files into a short string that acts as a signature for the data, and it is virtually impossible to generate a different data-file which will produce the same MD5 signature. Therefore, users of PRESAGE can verify that the alignments or models stored on annotators' Web sites are the same as those posted when an annotation was made: it is only necessary to compare the MD5 signature archived at PRESAGE with one generated from the current files.

Annotators are also requested to provide an indication of the method used to make the prediction, as well as measures of confidence in the reliability of the prediction. There are presently no constraints on the reliability measures that may be entered. However, as more structures are solved, users will be able to see when methods have exaggerated significance of matches.

Recommendation and request

To help guide selection of proteins for experimental characterization, researchers in structural genomics may recommend proteins for study. Reasons for suggesting a particular protein family might include pervasiveness throughout many different species but lack of known function or structure. It is also possible to request that a particular structure be solved, in the hope that the offer will appeal to a crystallographer or NMR spectroscopist. It is usually assumed that the requestors will provide appropriate materials (e.g., purified proteins), and these are indicated as part of the annotation.

FACILITIES

PRESAGE contains several methods of retrieving entries, including searches by various identifiers [including those used by SWISS-PROT and TrEMBL (10), GenBank (11) and the EMBL Data Library (12), TIGR (13), PDB (9) and SGD (14)], or by keywords in the SWISS-PROT description and comments about the proteins. Lookups for proteins annotated by a particular contributor are also available, as are searches by keywords in annotations or by types of annotation.

For researchers interested in tracking structural knowledge of a particular protein, the 'awareness' function may be especially valuable. This allows a user to register interest in a protein, and he or she will receive Email notification when annotations are made to that protein.

In the future, we plan to implement searches by family. These can either search for proteins within a family explicitly defined by an annotator, or the family may be defined on-the-fly using a sequence comparison algorithm with a desired threshold.

We will also be producing summary analyses of the data in PRESAGE. For example, many groups have made structural predictions for all the proteins in *Mycoplasma genitalium*, so these may be presented as a condensed report. To make the results comparable, the reports would 'normalize' all of the predictions to a single representative of the fold (e.g., using the scop database), and would indicate degrees of annotator-specified reliability.

AVAILABILITY

The database is publicly available at http://presage.stanford.edu/ . Contributors and individuals wishing to use the awareness function may register on-line, through links from that page. An interface for linking to PRESAGE will also be documented at the site.

CONCLUSION

Structural genomics will be a collaborative effort involving the talents of many experimental and computational researchers. Because of the expertise and time required to perform structure determination, the number of experimental scientists involved will be considerable. Continued advances in the methods and interpretation of sequence comparison mean that it will be valuable to draw upon a variety of predictions, whose effectiveness and reliability may vary greatly. With PRESAGE, all can contribute. We hope that the database will help link researchers in the decentralized field of structural genomics and make their results readily available.

ACKNOWLEDGEMENTS

The authors thank Joel Sussman and Otto Ritter for useful discussions at the beginning of the project, and Patrice Koehl for critically reading the manuscript. S.E.B. is an Alfred P. Sloan and US Department of Energy Postdoctoral Fellow in Computational Molecular Biology. This work is funded by DOE grant DE-FG-03-95ER62135.

REFERENCES

1 Gaasterland,T. (1998) *Nature Biotechnol*., **16**, 625–627.
2 Gaasterland,T. (1998) *Trends Genet*., **14**, 135.
3 Kim,S.H. (1998) *Nature Struct. Biol*., **5**, 643–645.
4 Pennisi,E. (1998) *Science*, **279**, 978–979.
5 Shapiro,L. and Lima,C.D. (1998) *Structure*, **6**, 265–267.
6 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.
7 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
8 Holm,L. and Sander,C. (1998) *Nucleic Acids Res*., **26**, 316–319.
9 Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*. Data Commission of the International Union of Crystallography, Cambridge, pp. 107–132.
10 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res*., **26**, 38–42.
11 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res*., **26**, 1–7.
12 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res*., **26**, 8–15.
13 White,O. and Kerlavage,A.R. (1996) *Methods Enzymol*., **266**, 27–40.
14 Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) *Nucleic Acids Res*., **26**, 73–79.