*e*Family: Bridging Sequence and Structure

Robert D. Finn[1], **Andreas Prlić**[1], Ujjwal Das[2], **Phil McNeil**[2], Nicola Mulder[2], Sameer Velankar[2], Antonina Andreeva[3], Dave Howorth[3], Mark Dibley[4], Tim Hubbard[1], Rolf Apweiler[2], Kim Henrick[2], Alexey Murzin[3], Christine Orengo[4], **Alex Bateman**[1]

1. Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK
2. The European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SD, UK
3. MRC Centre for Protein Engineering, Hills Road, Cambridge, CB2 2QH, UK
4. Department of Biochemistry and Molecular Biology, University College London, London, WC1E 6BT, UK

Abstract

The *e*Family project is working at improving the navigation between the disparate resources of protein structure and protein sequence. Integral to this work has been the establishment of an amino acid residue by amino acid residue mapping between the two resources, which is held in the MSD database. Within the project, there are also four protein family databases, two based on protein structure (CATH and SCOP), two based on protein sequence (InterPro and Pfam). The mapping has allowed the information contained in all the databases to be exchanged and queried from a user's point of view. In this paper, we describe how the information contained in the different databases is being exchanged using both the DAS (distributed annotation system) and Web services. Examples of a DAS client using these resources and a Web based tool using the Web services are presented. We also outline how these Web services can be combined with other services to form biological analysis pipelines.

Background to the field of Molecular Biology

In this section we briefly introduce aspects of molecular biology that are used in this paper for the benefit of readers outside the field. Within each of the billions of cells in the human body resides the genetic information in DNA. Genes are regions of the DNA that are used to make messenger RNAs in a process called transcription. The messenger RNAs are then used to provide a template to produce proteins in a process called translation. DNA, RNA and proteins are all polymers of a small pool of monomers. Proteins (polymers) are composed of chains of the twenty possible amino acids residues (monomers) in differing combinations. The series of residues that make up proteins can be represented as a string such as HITVEMNC, known as a protein sequence. These liner protein sequences normally fold into complex three-dimensional (3D) shapes. The precise structure formed is dictated by the protein sequence. When folded in 3D, disparate sequence motifs come together to form functional sites that allow the proteins to perform biochemical reactions that are essential for life.

There has been an exponential growth in the number of determined protein sequences (currently over 500,000) and 3D protein structures (currently over 25,000) in the public databases, with no signs of abating. The relative numbers of known protein structures is less than the number of known protein sequences due to the difficulty in experimental determination of protein structures. Given the growth of the protein database, protein analyses may seem like an endless task. However, the majority of proteins appear to fall into a few thousand protein families (1). Therefore, if we can place new sequences and structures into existing families the analysis task becomes much simpler.

Overview

The *e*Family project brings together 5 of the world's leading molecular biology databases that are based in the UK: CATH (2), InterPro (3), MSD (4), Pfam (5) and SCOP (6). These databases are built upon protein sequence or structure. MSD is the European database that archives protein structures. Pfam and InterPro classify proteins into families based on sequence, whereas SCOP and CATH classify proteins based on their known structure.

Historically, the resources for archiving protein sequence and structure have been developed independently, leading to difficulties in navigating between the two. As the number of protein sequences and structures increases rapidly the need to integrate the two types of data becomes more pressing. The *e*Family project is working towards bridging these two resources, thereby allowing seamless navigation between protein structure and sequence to the end user. The *e*-Science area that the *e*Family project is most concerned with is data grids, rather than compute grids. Our objective is to produce a series of well defined Web services that allow access to the data. These Web services will allow the establishment of analyses that are currently not possible or would be very time consuming with the member database's existing Web interfaces. Furthermore, the access to these Web services must be simple and readily available to the end user, who may be a non-computer expert.

Connecting Resources

A priori - Before the *e*Family member databases can be connected in a computational sense, a common co-ordinate system must be agreed to allow data exchange. Thus, a core element in the *e*Family project is the production of the non-trivial residue by residue mapping between the sequence (UniProt) and structure databases by the MSD project. Currently, the MSD database has cross references to UniProt for 99% of protein structure entries and residue mappings for more than 97% of protein entries.

Data Dissemination Using the DAS Standard

The distributed annotation system (DAS) provides a communication protocol used to exchange biological annotations (7). DAS is based on the theory that annotation should not be provided by a single centralised database, but rather be spread over multiple sites. Data distribution, performed by DAS servers, is separated from visualisation, which is performed by DAS clients. Currently DAS usage is bias towards genomic DNA sequence annotation rather that protein sequence. Consequently, the current DAS specification does not readily allow protein structural or family data to be distributed, and therefore no suitable clients for structure and family data have been developed. We have proposed an extension to the DAS specification (http://www.sanger.ac.uk/xml/das/documentation/new_spec.html) and developed a prototype client to integrate sequence and structure data (Figure 1).

Data Dissemination Using Web Services

In addition to DAS, we are developing Web services that allow calculation of results/information that is beyond the scope of DAS. A prerequisite to developing Web services for exchange of data is a defined XML schema, referred to as the *e*Family schema. The *e*Family schema has to model protein family definitions from the four different protein family databases, structure and sequence alignments, database cross references and the residue by residue mapping between UniProt and MSD. Consequently the data model is very complex, so an API has been developed for use with the *e*Family schema. The API has been specifically designed so it can be readily integrated into BioPerl, with the proposal for its integration into the core distribution submitted recently. For example, the API allows a Pfam alignment wrapped up in the *e*Family XML to be returned as a BioPerl alignment object and *vice versa*. By integrating the connection methods and API into BioPerl, the *e*Family member databases will become more accessible to the community at large and the methods employed for handling the XML are generically written so they can be readily applied to other schemas. Furthermore, databases outside the

project can readily use the *e*Family schema to export their data.

A simple, but elegant, use of these Web services is a Web based tool for domain comparison that integrates results from SCOP, CATH and Pfam (Figure 2). Upon request, the SCOP and CATH family definition Web services are queried and the results for the structurally defined families are compared to those from Pfam (which are defined on sequence), a process which relies on the MSD mappings between structure and sequences.

Combining Services to Form Work Flows

The Web services that we are providing also allow access to some compute resources. The Pfam analysis tool, PfamScan, has been deployed for the calculation of domain matches using SOAPlab. SOAPlab, part of the myGRID project allows the deployment of command line tools as production quality Web services. Production quality Web services can then be integrated with other Web services to perform scientific analysis. We have found Taverna, again part of the myGRID project, an excellent platform for the integration of Web services from the *e*Family project and external Web services. Taverana is so appealing as it is an "off the self" package that end users can easily install and use.

Future Work

Currently, there are only a limited collection of DAS- and Web services. We intend to increase the number of services drastically over the next year to allow many different types of analyses to be performed. The member databases will also be using the Web services to enrich the information content of their sites. However, Web services are based on the client/server model, which could result in overload on the server. While we perceive that Web services will be the main model for disseminating the information, in some cases it may be appropriate for users to replicate the data locally. The replication may often only require replication at a certain granularity of the data, rather than the complete database. Currently, the MSD search database is replicated at 11 sites worldwide. Further work on replication and update mechanisms for all of the databases in the project is required.

References

1. Chothia C. One thousand families for the molecular biologist. Nature. 1992 357:543-544.
2. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res. 2003;31:452-455.
3. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res. 2003;31:315-318.
4. Golovin A, Oldfield TJ, Tate JG, Velankar S, Barton GJ, Boutselakis H, Dimitropoulos D, Fillon J, Hussain A, Ionides JM, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Pajon A, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan GJ, Tagari M, Tromm S, Vranken W, Henrick K. E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res. 2004;32:D211-216.
5. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res. 2004;32:D138-141.
6. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004;32:D226-229.
7. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. The distributed annotation system. BMC Bioinformatics. 2001;2:7. Epub 2001.

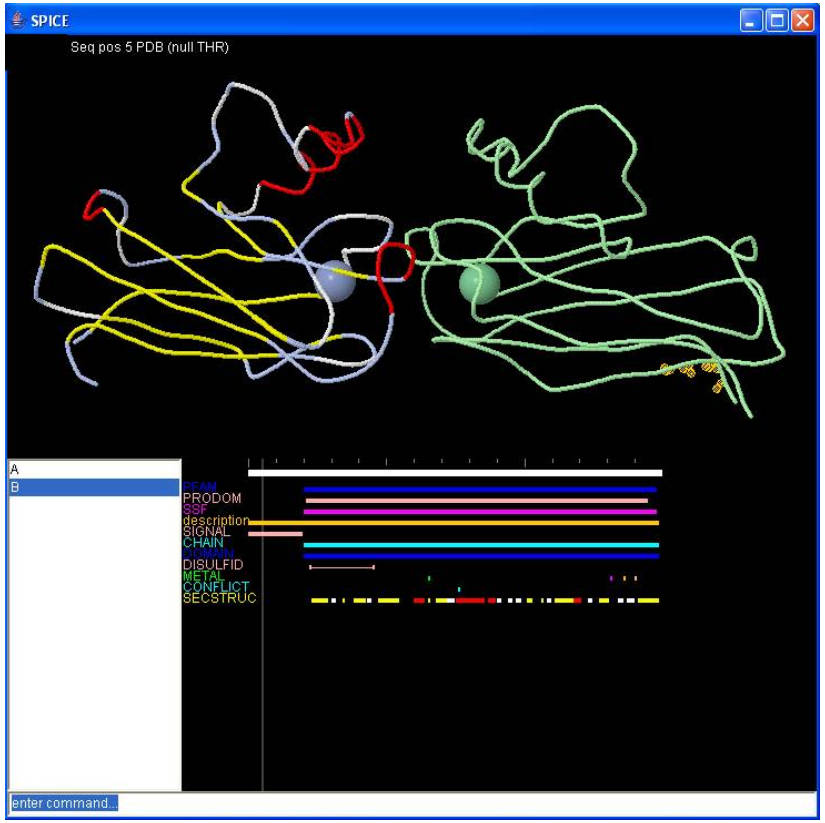

Figure 1 – Top is a graphical representation of the 3D structure of Azurin (PDB accession code 1A4A). The 3D structure information coming from a DAS source. **Bottom, left**, the annotated chain in the structure is highlighted. **Bottom, right** are different annotations about the active chain, served from different DAS services. The bottom annotation, secondary structure (SECSTRUC: coloured red, yellow and white), has been used to colour the 3D structure (**top, left**).

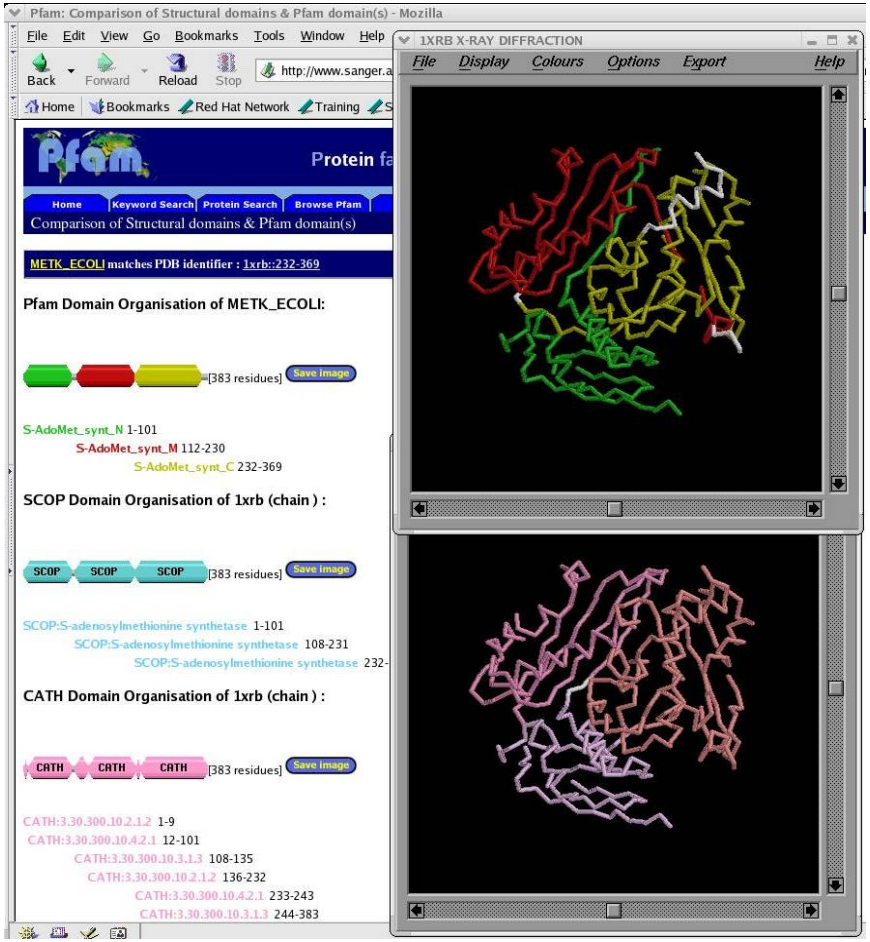

Figure 2 – Left, a web page that combines the results of web services provided by Pfam, SCOP and CATH to show how each database defines the protein domains in a single protein. The web page also allows the user to show the domain definition from each database projected onto the protein structure (**right**).