Bridging Sequence and Structure

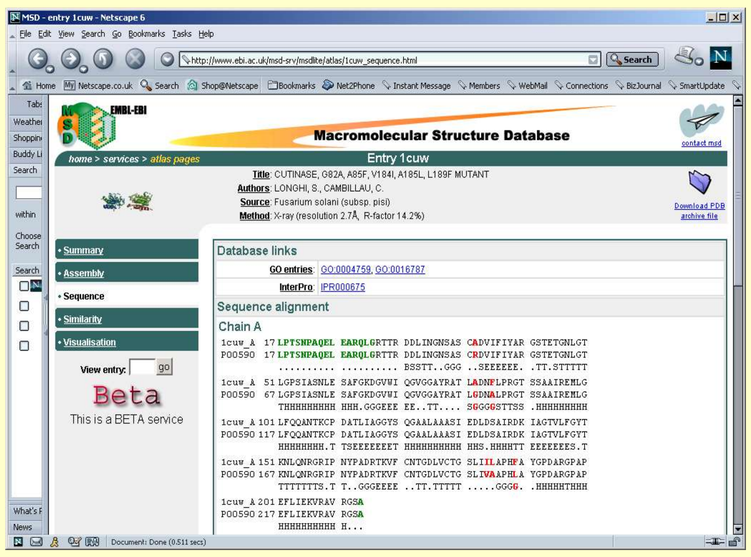
Robert D. Finn[1], **Andreas Prlić**[1], Ujjwal Das[2], **Phil McNeil**[2], Nicola Mulder[2], Sameer Velankar[2], Antonina Andreeva[3], Dave Howorth[3], Mark Dibley[4], Tim Hubbard[1], Rolf Apweiler[2], Kim Henrick[2], Alexey Murzin[3], Christine Orengo[4], **Alex Bateman**[1]

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

Introduction

The eFamily project brings together 5 of the world's leading molecular biology databases that are based in the UK: CATH, InterPro, MSD, Pfam and SCOP. These databases are built upon protein sequence or 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 eFamily project is working towards bridging these two resources, thereby allowing seamless navigation between protein structure and sequence to the end user, who is more often than not a non-computer expert.

Mapping Between Protein Structure and Sequence



Before the eFamily 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 eFamily 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.

The eFamily Website

As the eFamily project strived to maintain federated databases, the eFamily website, http://www.efamily.org.uk/, does not act as a data server. However, the eFamily website acts an information resource for the eFamily project. The site contains a rudamentry UDDI for both the webservices and DAS servers. People can access software downloads and documentation for the DAS and eFamily XMLs.

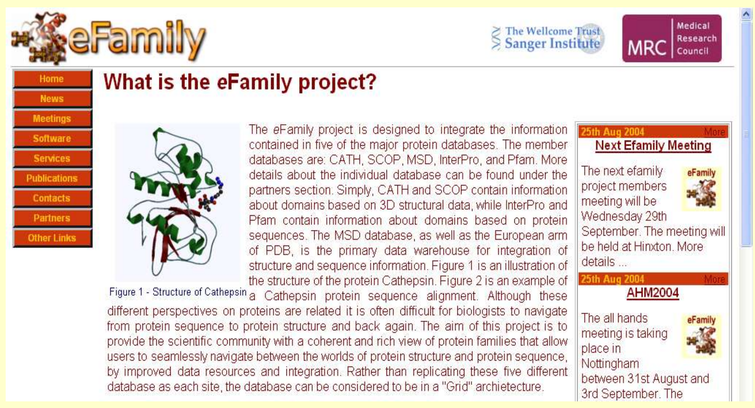
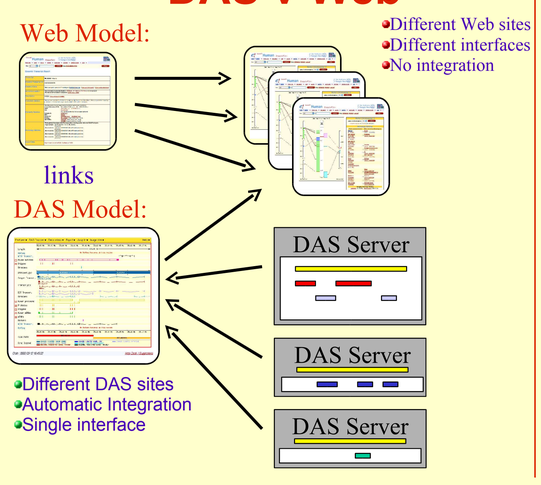


Figure 1 - Structure of Cathepsin

Data Exchange Mechanisms 1: DAS – The Distributed Annotation System

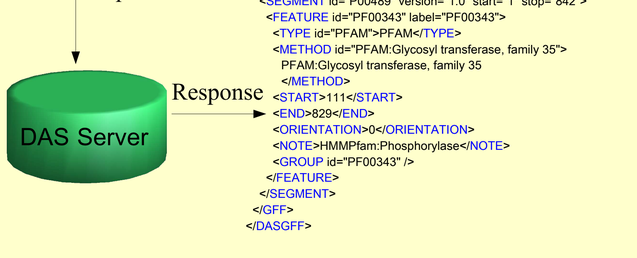
DAS provides a system that allows the mapping of a set of features or attributes to a stable identifier. DAS has been widely used to annotate biological data. Features/attributes are retrieved from a seft of disparate resource (DAS servers). The different feature information is then displayed using a DAS client. In the DAS system the data served is relatively simple (features on seqeunce), so it is the client, rather than the server that has to be sophisticated. This is in contrast to the web generally where the server data complex and the client simple.

DAS v Web



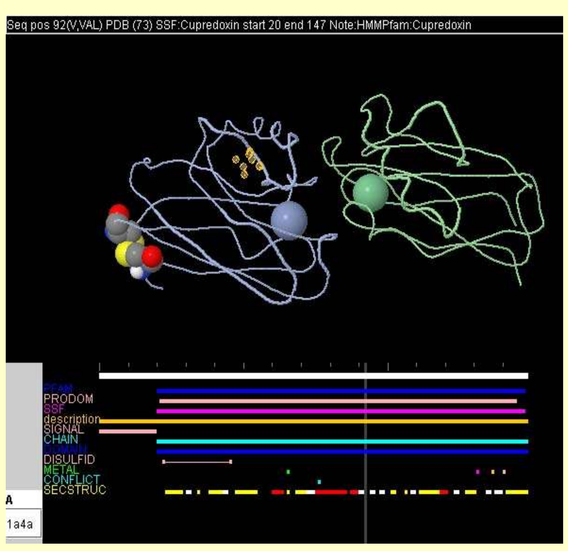
DAS Request/Response

http://das.sanger.ac.uk/das/interpro/features?segment=P00489



In the past, DAS usage has been biased towards genomic DNA sequence annotation rather protein sequence. Consequently, there was no way of encapsulating protein alignments or protein structure. As part of this project we have extended the DAS specification to allow sequence and structure aligments.

We have also developed a prototype client to integrate sequence and structure data. The DAS client, called SPICE, is shown left. The features are represented as bars underneath the sequence (white bar). The 3D structure of the white bar is in the top panel and coloured blue. The DAS feature highlighted in the structure is a disulphide bridge



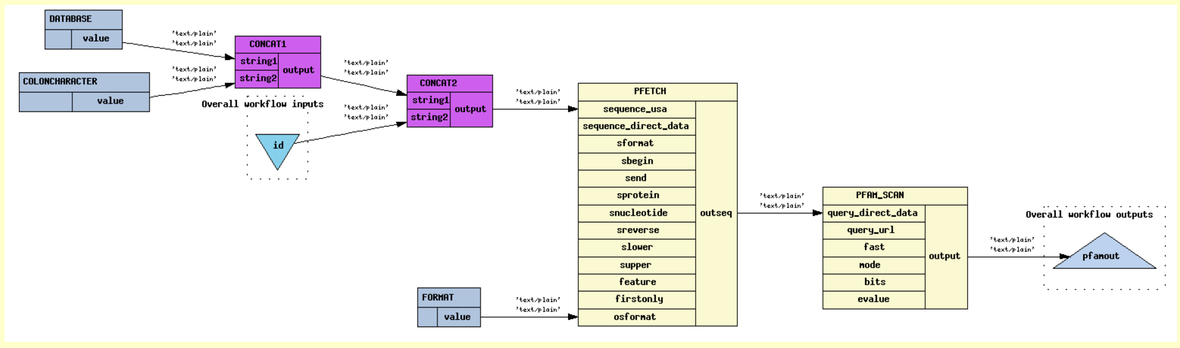
Data Exchange Mechanisms 2: Webservices

After investigation, eFamily has decided that one of the most appropriate methods for making data and compute resources available is via standard Webservices. Our Webservices have been produced using a variety of methods: Java and Apache Axis; Perl SOAP::Lite module or the third party software SOAPLab (http://industry.ebi.ac.uk). Avialable services are listed on the eFamily website

Exchange Format & API – As the eFamily members are going to be exchanging similar sorts of data, which will be used by third parties, we have developed an XML schema that models the mapping between sequence and structure, domain information and sequence and structure alignments. More information about the schema can be found at http://www.efamily.org.uk/xml/efamily/documentation/efamily.shtml. As the data model is complex we have written an API to perform the I/O between the data (often stored in BioPerl objects) and the XML. We plan to summit this API to BioPerl in the near future making the XML and API accessible to the wider scientific community.

Performing eScience

Using eFamily Web Services – Web Services can be accessed directly by software or multiple Web Services can be linked together to form workflows. Taverna, part of the myGRID project, provides the framework for graphically integrating Webservices (below).

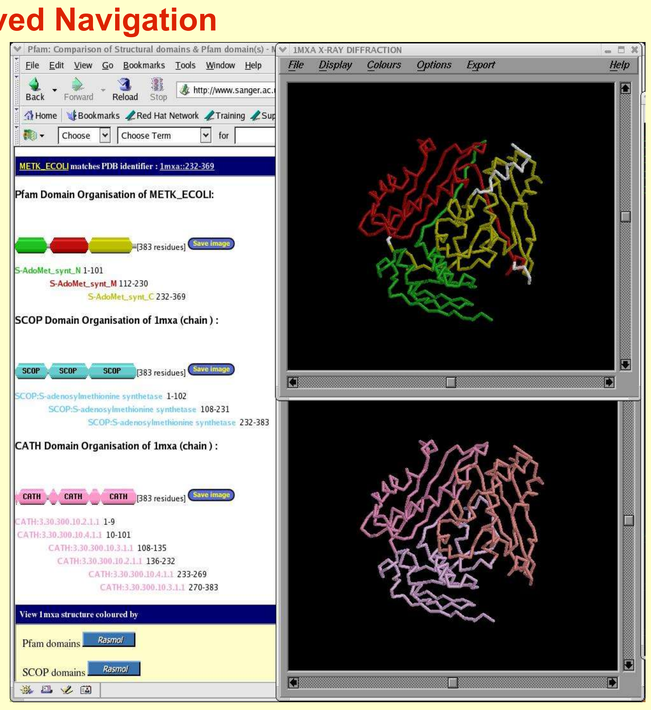


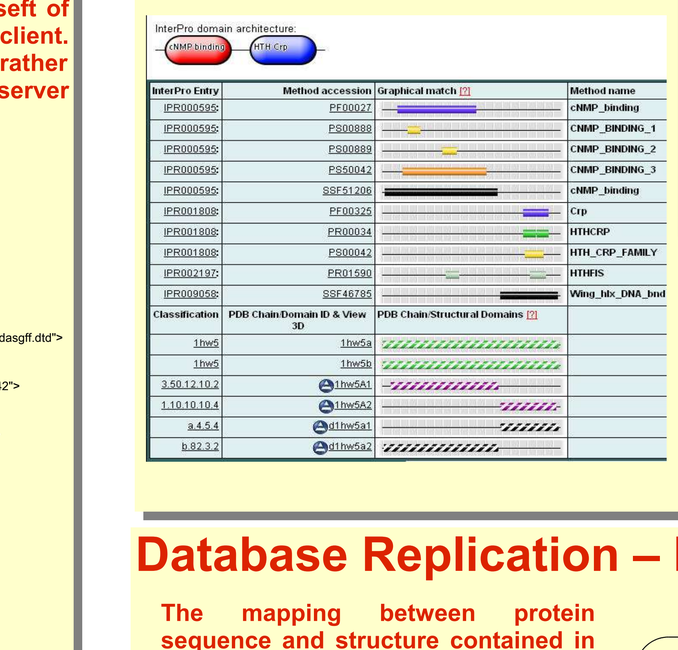
A workflow that integrates the EBI SRS services and pfam_scan service to annotate new sequences. The workflow takes a sequence ID, retrieves the sequence from the EBI, then calculates any matches to Pfam, returning a formatted list of results.

The eScience Behind the Scences

Web Domain Comparison & Improved Navigation

Although the Webservices we provide are programmatically accessible, we also use them as convenient visualisations on the Web. In Pfam, domains with a determined structure can be compared to the structurally defined domains from SCOP and CATH. The three different domain definitions are then displayed so that they can be compared (near-right). The domains can also be compared in 3D (far-right). The Pfam and SCOP domains are very similar, but the CATH domains have identified the strand swapping that occurs in the structure.





InterPro Entry	Method accession	Graphical match [?]	Method name
IPR000595	PF00027		cNMP_binding
IPR000595	PS00888		CNMP_BINDING_1
IPR000595	PS00889		CNMP_BINDING_2
IPR000595	PS50042		CNMP_BINDING_3
IPR000595	SSF51206		cNMP_binding
IPR001808	PF00325		Crp
IPR001808	PR00034		HTHCRP
IPR001808	PS00042		HTH_CRP_FAMILY
IPR002197	PR01590		HTHFIS
IPR009058	SSF46785		Wing_htx_DNA_bnd
Classification	**PDB Chain/Domain ID & View 3D**	**PDB Chain/Structural Domains [?]**	
1hw5	1hw5a		
1hw5	1hw5b		
3.50.12.10.2	1hw5A1		
1.10.10.10.4	1hw5A2		
a.4.5.4	d1hw5a1		
b.82.3.2	d1hw5a2		

Similarly, domains can be compared in InterPro (below). In this view, information from all the eFamily member databases and other InterPro consortium members can be compared. The sequence domains (top, solid boxes) can now be compared to the structural defined domains (bottom, striped boxes).

Database Replication – MSD replication Service

The mapping between protein sequence and structure contained in the MSD database is so heavily used be each member database, it has necessitated the limited replication of MSD at different sites.

Right is a schematic representation of how MSD replication and incremental updates are produced and exported using the replication service.

Currently, this mechanism has been used to replicate MSD at 11 different sites, with only two of these site eFamily project members.