3.1 Publishable summary

Functional annotation of proteins is an important step in understanding biological systems, diseases and pathogenesis; protein families and domains are invaluable pointers that help biologists to find distantly related proteins and predict their functions. A daunting array of resources, each with different strengths and weaknesses, is now available to search genomes and proteomes for "protein signatures" - diagnostic entities that are used to recognise a particular domain or protein family. In order for end users to utilise these resources effectively in their research, they have been amalgamated together into a single resource (a database called "InterPro"). The continued generation of these signatures, their annotation and integration into InterPro and the provision of software and databases to serve them to the public is therefore of great importance to the life science community.

The IMPACT (IMproving Protein Annotation through Coordination and Technology) project involves a Consortium of 9 experienced partners, some of whom have been collaborating for almost 12 years in the field of protein family and domain prediction (these include the InterPro, PRINTS, PROSITE, SUPERFAMILY, SMART, Pfam, CATH-Gene3D and ProDom databases). IMPACT aims to harness existing technologies and use them to dramatically improve these information resources. Such improvements will be achieved through:

- Closer coordination of research activities of the Consortium partners, minimising the amount of duplicated effort in protein signature generation and annotation.
- Establishing or expanding upon standardised formats or protocols to encourage faster, more transparent data exchange between partners.
- Creating closer links to external user groups through training and outreach activities and consequently enhancing data access routes for them, such as web interfaces, web services and use of DAS ("Distributed Annotation System").
- Improving the InterProScan signature scanning software so that distributed compute resources are used more effectively within the consortium.
- Adding new features to the InterPro database, such as biological pathway information and functional labels, including Gene Ontology (GO) terms.

The progress of the IMPACT project towards these objectives may be followed by visiting the project website (http://www.ebi.ac.uk/impact/) and individual consortium partners' websites.

During the first reporting period (1st January 2008 to 31st December 2008), the main focus of the project was to establish procedures to facilitate the smooth running of the Consortium. In the second reporting period (1st January 2009 to 31st December 2009), the plans from the first period were implemented, and preparations were made for the execution of developments in the third period. The final period (1st January 2010 to 30th June 2011) sees the culmination of the project's work, outlined below.

During the third and final period, the generation, quality checking and annotation of the diagnostic signatures that make up the core of InterPro was once again a major undertaking. At the start of the IMPACT project 3 years ago, just over 5 million proteins (5,148,042 – InterPro release 16.1) from the UniProtKB protein knowledgebase matched at least one diagnostic signature; that is, InterPro was able to add some form of annotation to those proteins (either conserved domains or membership of functional families). In 2011, this total has virtualy tripled to almost 13 million proteins (12,897,947 – InterPro release 32.0). This has been achieved by the continued generation of signatures that cover novel functional space by IMPACT partners. This is challenging, as the majority of well-known, large families and domains were already described many years prior to the start of the IMPACT project, and so developments have been targeted out of necessity towards

niche families and rarer domains that describe smaller numbers of proteins. The fact that IMPACT has resulted in a 5.2% increase in protein coverage within 3 years (proportion of proteins thta can be attributed with some form of annotation via InterPro) is therefore particularly impressive. InterPro has continued utility in the functional annotation of new individual genome projects, and is being increasingly used to characterise community functions from environmental samples (also known as metagenomics). It is expected that metagenomic projects will be a new source of novel biology that will need to be described by InterPro's signatures in the future, beyond the end of IMPACT.

When signatures are incorporated into the central InterPro database, there are two phases to their integration:

i) A quality-checking phase, where the accuracy and sensitivity of the signatures are measured, and any signatures that are making spurious matches are fed back to the source signature database(s).
ii) An annotation phase, where a description of what the signature is representing is created, together with an appropriate name. The signature is also labelled with Gene Ontology (GO) terms, which utilise a controlled vocabulary to categorise the family, site or domain according to its function, localisation in the cell and the processes in which it is involved.

An outcome of the IMPACT project is that 10,067 InterPro entries are now associated with at least one GO term. There has therefore been an increase of over 5,000 GO associations since the start of the project. Additionally, data curators have also begun adding an "un-mappable" flag for those entries to which it is not possible to associate a term; approximately 3,300 entries currently have this status. Information allowing users to infer in which biological pathways a particular family or domain is implicated has also been added to InterPro, and it is expected this will have utility in the contextual interpretation of whole genome annotation (that is, is a pathway present in a genome or not?). All of these developments increase the biological value of InterPro for researchers studying genome and protein sequences.

The consortium has attained consensus regarding the terminology that should be used when describing signatures and the biological entities they represent. This includes formalising the definition of the entities represented in InterPro (Family, Domain, Binding Site, Conserved Site, Active Site, Post-Translational Modification), and working with nomenclature committees to ensure naming conventions are adhered to, leading to increased coherence for end users. Data are also now exchanged between consortium members using the IMPACT XML-based data exchange formats, which are used to represent the annotation of signatures and the proteins they match. This has speeded-up the process of sending, loading and checking data between InterPro's partners. Users can also access this format via the InterProScan software, where it is optionally used as input into a programmatic module for automatically annotating proteins with additional features, such as structurally-solved active sites and domain architectures. It is highly likely that we will nominate the standard format for inclusion in the Proteomics Standards Initiative (PSI) after the end of the IMPACT project.

Two major routes by which external users access the data in InterPro (the signature scanning software InterProScan, and the public web interface to the database) have been completely re-vamped during the project and a third (Distributed Annotation Service, DAS) has had new components added.

In the project's second year, a modular version of the InterProScan software was produced, which was able to run in a more flexible and robust manner than earlier versions. During 2010, this prototype version has evolved into a fully-functional piece of software, giving users the ability to search all 11 InterPro signature databases (note that databases and algorithms outside the scope of IMPACT are also included) against their sequences. Additional features have been added, including a new mechanism for looking up pre-calculated results for publicly available proteins; an option to

retrieve the newly included pathway association data from InterPro; and the possibility for users to relate predicted signature matches back to original genomic sequences, rather than just the encoded protein sequence translations. The new InterProScan software ("i5") is now available for download by external users from a Google Code repository.

The Distributed Annotation System (DAS) allows federated annotation of sequences by remote groups. This technology allows a client to download features from numerous databases, such as the IMPACT partners. We have extended this technology for use with multiple sequence alignments, which are a common currency of protein family and domain information. A DAS alignment viewer using a DAS widget Javascript library developed for this project (http://pfamsrv.sanger.ac.uk/docs/) allows partners to display their own alignments or those from other partners within their websites. Using this technology, PROSITE, Pfam and CATH-Gene3D serve out DAS alignments. In particular, CATH-Gene3D provide a service for structural alignments that will enable access to this valuable resource to the IMPACT partners. DAS has recently moved to a 1.6 specification, and the partners have upgraded to this new standard.

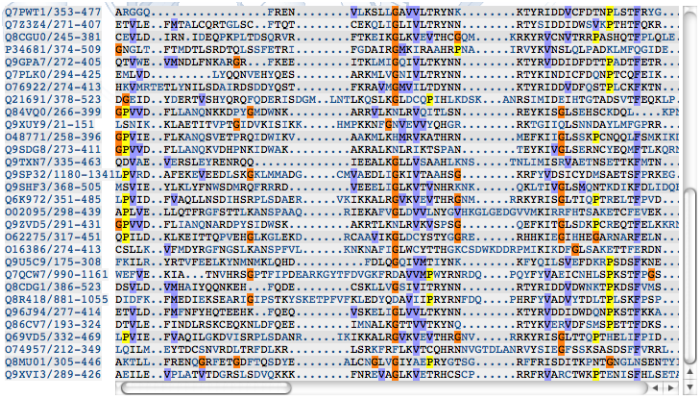

Figure 1. An example multiple sequence alignment served by DAS and displayed within a web page as a widget.

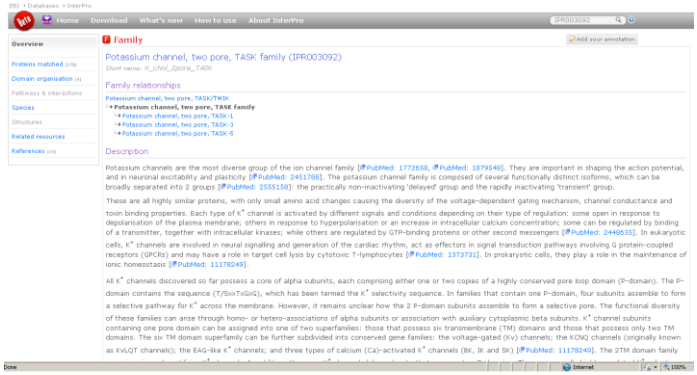

Figure 2: Screenshot of the new InterPro web interface. The design is much cleaner and easier to navigate than the previous site.

The InterPro web interface has been redesigned using a user-centred (UX) approach, leading to a resource that better reflects users' needs, and is hopefully easier for novice users to navigate and understand. All of the new data that have been included in InterPro as a consequence of IMPACT are now available via this interface (*e.g.*, new signatures, proteins, GO terms and pathway associations), and IMPACT-developed web services have also been integrated (*e.g.*, FASTA sequence download API). The new website has been available to users since December 2010 and thus far, has evoked a very positive response.

During the project, we aimed to deliver over 20 million web accesses to the InterPro website. We have surpassed this mark, delivering 45 million page impressions in the last 18 months alone. In addition, we have generated a similar number of web hits to the member databases collectively. However, we acknowledge that web hits are a poor measure of resource access, and are dependent on the measurement procedure used: *e.g.*, whether, and how, to eliminate hits from search engine robots (such as Google and Bing) indexing the sites. We prefer to look at the number of independent sites that have accessed the websites, as this is more stable, and better represents the extent of the influence of InterPro and the member databases. Of course, this metric is also susceptible to interpretation (*e.g.*, whether we count IP addresses, or institutions); nevertheless, it is

clear that tens of thousands of independent sites access InterPro and the member databases every month. There are large seasonal fluctuations, but these are similar year on year.

The successes of the IMPACT project have been disseminated to the scientific community in multiple ways. The IMPACT consortium has been represented at numerous international meetings and conferences, including the annual ISMB conference in Boston (July 2010) where a 50 minute "Technology Track" presentation was given about InterPro; the ICT 2010 meeting in Brussels, where IMPACT exhibited, and the 8th e-Infrastructure concertation meeting at CERN. In addition, twelve papers (so far) describing work arising from IMPACT have been published, or accepted for publication, in peer-reviewed journals over the course of the project, along with several ancillary publications, such as book chapters.

The second and third IMPACT training workshops were held in May and November 2010, providing consortium members and members of the InterPro and UniProt teams opportunities to exchange ideas, to understand how the different member databases create and maintain signatures, and, crucially, to interact with, and train, external users of InterPro. Both courses were deemed a success and it is likely that future post-IMPACT courses will follow a similar structure. The user-training events also provided a good opportunity for the consortium to get feedback about the developments that have occurred during IMPACT, and again these directly influenced the web re-design process. Having regular training and outreach events has allowed direct, informative contact with user groups, allowing us to develop better end-products.

Further information about the achievements of the project are available from the IMPACT project website, or are available, on request, from the project coordinator, Sarah Hunter (hunter@ebi.ac.uk).