Institution: University of Manchester

Unit of Assessment: UoA5

Title of case study: PRINTS and InterPro – online resources that facilitate discovery of pharmaceutical and commercially relevant information in proteomic and genomic data-sets

1. Summary of the impact

Automation of genomic data analysis has become essential. High-throughput sequencing technologies are producing data faster than can be managed and interpreted, meaning that much biomedical information remains unused.

Research led by Attwood introduced a unique method for protein sequence characterisation and a derived database of diagnostic protein signatures (PRINTS). This led directly to the development of a new database (InterPro), now routinely used to annotate the world's largest protein sequence archive (UniProt), and complete genomes and metagenomes. The databases and their search tools have been exploited in the private sector (including SMEs and multi-national pharmaceutical and agrichemical companies), generating workflows that have yielded candidate drug targets and provided insights into disease mechanisms.

2. Underpinning research

These impacts are explicitly based on research that was funded and undertaken at the University of Manchester (UoM) from 1999 to date. Key researchers were:

Professor Teresa Attwood (1999 to date)

Post-Doctoral Research Associates: **Dr Alex Mitchell** (2000-2011; EBI InterPro Content Manager, 2011 to date), **Dr Jane Mabey** (2001-2006), **Dr Mike Croning** (1999-2002), **Dr Phil Scordis** (2000-2001)

Research Assistants: **Mr Paul Bradley** (2002-2005), **Mr Ala Uddin** (2001-2003), **Mr Julian Selley** (1999-2002)

PhD students: **Mr Neil Maudling** (2000-2005), **Ms George Moulton** (2000-2005), **Ms Anna Gaulton** (2000-2004), **Mr Will Wright** (1999-2001)

The underpinning research aims to develop protein family databases, and tools for family analysis and annotation. At one end of the spectrum is high-throughput genome annotation and at the other is fine-tuned functional characterisation of pharmaceutically relevant proteins. As such, many of the team's research projects have been funded (~£359k) by major pharmaceutical companies and SMEs (AstraZeneca, Pfizer, Cambridge Drug Discovery, Roche Discovery, GlaxoSmithKline), highlighting the value of the work in drug-discovery programmes.

The key steps at the UoM were:

1. The motivation was to create a pipeline for automatic annotation of the mass of raw genomic data entering the European Bioinformatics Institute's (EBI's) TrEMBL database. InterPro was first released in October 1999, with seed data from PROSITE, Pfam and PRINTS [1]. InterPro's power comes from integrating complementary analysis methods, allowing it to provide more robust functional diagnoses (*i.e.*, the diagnostic sum is more powerful than its component parts).

2. Attwood and her team showed that their uniquely selective fingerprint method for sequence analysis could uncover new, distantly related, members of protein families (*e.g.*, a range of new tubulins in protozoal parasites was discovered [2]).

3. Attwood's team then exploited the selective nature of the method to establish an innovative hierarchical approach to protein family fingerprinting, for the first time allowing extremely fine-grained functional analyses. They showcased the approach using the pharmaceutically important GPCRs, creating a compendium of >200 family- and subfamily-specific GPCR fingerprints [3].

The approaches have been extended to numerous families. Such work directly augmented PRINTS [4,5], doubled its size, and provided >1,000 new family- and subfamily-specific signatures to InterPro. These strands of research remain fruitful and continue to feed into the impact: PRINTS is actively maintained and continues to supply InterPro with unique hierarchical fingerprints [6].

3. References to the research

The research was published in leading life science journals (*Trends in Pharmacological Sciences, Current Biology*) and in the top journal in the field of bioinformatics for database publications (*Nucleic Acids Research*).

1. Apweiler, R., **Attwood, T.K.**, Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., **Croning, M.D.**, Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*. 29 (1). p. 37-40. DOI: 10.1093/nar/29.1.37

2. **Vaughan, S.**, **Attwood, T.K.,** Navarro, M., Scott, V., McKean, P.,**Gull, K.** (2000) New tubulins in protozoal parasites. *Current Biology.* 10 (7). p. R258-259. DOI: 10.1016/S0960-9822(00)00414-0

3. **Attwood, T.K.** (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol.Sci*. 22 (4). p. 162-165. DOI:10.1016/S0165-6147(00)01658-8

4. **Attwood, T.K., Croning, M.D.,** Flower, D.R., Lewis, A.P., **Mabey, J.E., Scordis, P., Selley, J., Wright, W.** (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 28 (1). p. 225-227. DOI: 10.1093/nar/28.1.225

5. **Attwood, T.K.**, **Bradley, P.**, Flower, D.R., **Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A.**, Paine, K., Taylor, P., **Uddin, A.,** Zygouri, C. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31 (1). p. 400-402. DOI: 10.1093/nar/gkg030

6. Hunter, S., Apweiler, R., **Attwood, T.K**., Bairoch, A., Bateman, A., Binns, D, Bork, P., Das, U., Daugherty, L., Duquenne L., Finn, R., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., **Mitchell, A.** *et al*. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*. 37. p. D211-215. DOI: 10.1093/nar/gkn785

4. Details of the impact

Context

Before research at UoM, protein sequence annotation was largely done manually, and the gold-standard repository, Swiss-Prot, had become a bottleneck, preventing early access to genomic data. To tackle this problem, TrEMBL was established at the EBI as an unannotated Swiss-Prot supplement. Attwood and her research team established InterPro, the world's first integrated protein family repository, following on from PRINTS. Its creation made possible the implementation of robust, high-throughput annotation workflows for TrEMBL and also for complete genomes. In turn, this led to the development of new drug-discovery pipelines and global usage of InterPro.

Pathways to impact

PRINTS has been maintained at UoM for 14 years. Attwood re-established the umbrella website, DbBrowser (the access point for PRINTS and its analysis/annotation tools) in March 1999, and set up an FTP site for their anonymous download.

InterPro has grown in size and complexity, now having a dozen partner databases focusing on gene- and domain-families, protein folds, architectures and superfamilies. Hosted at the EBI, with continued input of new diagnostic signatures from its partners (including PRINTS), InterPro has become the pre-eminent instrument for analysis and functional annotation of uncharacterised genomic data. Since 2008, InterPro has featured in ~30 educational national and international workshops, and is covered in two EBI online e-learning courses: *InterPro: Quick tour* and *Introduction to Protein Classification at the EBI*. The audience for these workshops includes clinicians, researchers, technologists and industrialists. Various PRINTS- and InterPro-specific modules are also available as part of GOBLET, a global, community-based organisation providing a centralised public facility for sharing educational materials. The creation of GOBLET was led by

Attwood, who is the current Chair.

Reach and significance of the impact

Establishing one of the world's largest resources for protein family annotation:
InterPro was quickly adopted as part of the TrEMBL (now part of UniProt) annotation pipeline, and continues to contribute to the rule-based automatic annotation system for protein sequences in UniProtKB/TrEMBL [A]. Sequences in UniProt include fingerprint annotations, with hyperlinks back to PRINTS, making PRINTS one of UoM's most heavily used resources, *e.g.*, since 2008, traffic to PRINTS and its services has grown from ~2.5 million hits/annum to >3.4 million. In 2010, InterPro accounted for 12% of all Web visits to the EBI: it now draws ~45 million hits/annum from the UK/Europe, USA, Canada, Japan, *etc.* (typically, ~3 million of these from industry), and supports ~2 million searches/month.

New sequencing technologies will assure both a major ongoing impact on InterPro usage statistics, and a continued role for this pivotal database in genome- and proteome-annotation programmes [*e.g.*, B-D]. In 2011, InterPro contributed to the annotation of ~13 million proteins, making it one of the world's largest, most successful publicly available protein family annotation resources, on a par with the NCBI's Conserved Domain Database, and resulting in its recognition as one of ELIXIR's core data resources [D].

The advantages of using protein fingerprints:
Fingerprints uniquely provide hierarchical 'superfamily to subfamily' diagnoses, so users of PRINTS, InterPro and UniProt benefit from more fine-grained functional insights than are given by InterPro's 'catch-all' methods: *e.g.,* PRINTS classifies GPCRs and ion channels (the most cited and downloaded 3D structures, owing to their pharmaceutical relevance) into ~400 families and subfamilies. Such functional 'fine-tuning' improves the specificity of Gene Ontology (GO) mapping within InterPro, mappings that are now cross-referenced >66 million times in UniProt and provide functional terms for >11 million proteins. GO annotation provided by InterPro is the largest source of automatic GO annotation for proteins from all organisms [E].

Providing valuable information to pharmaceutical and commercial organisations:
Around 10% of PRINTS >1,200 and InterPro's >3,100 cumulative citations are by authors from commercial organisations (including pharmaceutical companies like GSK, Merck, Novartis, Roche, Pfizer, AstraZeneca, Intervet and agricultural, agrifood and agrichemical companies like BASF). Approximately 1.3 million webpages have been served to visitors from commercial domains from Jan-June 2013 [F].

InterPro developments have fuelled tangible outcomes for pharmaceutical industries, including:

1. Identifying specific genes and signalling pathways that may contribute to motor neuron degeneration in amyotrophic lateral sclerosis [G] (authors affiliated to Genentech).
2. Identifying gene families that play a critical role in chronic skin infections caused by *Trichophyton rubrum* (*e.g*., athlete's foot) [H] (author affiliated to Proctor & Gamble).
3. Providing evidence that combined inhibition of simultaneously active receptor tyrosine kinases can lead to an added anti-cancer effect [I] (author affiliated to Novartis).

One purpose of the *American Recovery & Reinvestment Act of 2009* was to increase economic efficiency by spurring technological advances in science and health. A project was developed to sequence individual human genomes for $1,000, and hence translate sequencing into a viable clinical tool, underpinning the worldwide goal of 'personalised' medicine. Efforts to achieve this goal are now generating data on an unimaginable scale, most of it useless without annotation. InterPro harmonises world-wide sequence-annotation projects and therefore provides an efficient platform for transforming raw data into pharmaceutically and clinically useful information; it thereby helps to realise both the global financial investment in, and the future clinical outcomes of, these transformative new sequencing technologies.

5. Sources to corroborate the impact

A. UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*., 41(Database issue), D43-47

B. Lamesch, P., *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, 40, D1202-D1210

C. Schneider, M., *et al.* (2009) The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. *J. Proteomics,* 72(3), 567-573. PMID: 19084081

D. ELIXIR: Overview, Progress & Futures. http://ec.europa.eu/research/biotechnology/eu-us-task-force/pdf/20th-meeting/presentation_of_elixir_project_en.pdf

E. Burge, S., *et al*. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database*, Article ID bar068. DOI:10.1093/database/bar068

F. Letter from EMBL-EBI, *corroborating the important contributions of InterPro and PRINTS.*

G. Phatnani, H.P., *et al.* (2013) Intricate interplay between astrocytes and motor neurons in ALS. *Proc. Natl. Acad. Sci. USA*. 110(8), E756-765. DOI: 10.1073/pnas.1222361110

H. Martinez, D.A., *et al*. (2012) Comparative genome analysis of *Trichophyton rubrum* and related dermatophytes reveals candidate genes involved in infection. *MBio*. 3(5), e00259-12. DOI: 10.1128/mBio.00259-12

I. Harbinski, F., *et al*. (2012) Rescue screens with secreted proteins reveal compensatory potential of receptor tyrosine kinases in driving cancer growth. *Cancer Discov.*, 2(10), 948-959. DOI: 10.1158/2159-8290.CD-12-0237