EMBnet.journal
Bioinformatics in Action

Protein.spotlight, Issue 177 - e868

The art of biocuration

a special article to celebrate Swiss-Prot's 30th Anniversary

Vivienne Baillie Gerritsen, Marie-Claude Blatter

Museums have their curators. Art galleries too. Their job is to look after collections they are knowledgeable about and present them to an audience in a way that makes sense and is informative. Biocurators do the same. Ever since the advent of computers and advanced technology in the life sciences, the quantity of biological data has grown exponentially and been stored in databases. The simple piling up of data, however, is of little help not only to researchers but also to computers. To be useful, they need to be sorted some way or another. Such a step is easily performed by specialized software. But as for many things, without a human touch something lacks. Swiss-Prot is a protein sequence database that sprung into existence 30 years ago when protein sequences were still trickling in. In those days, every sequence could be nursed. Today, however, millions of protein sequences are produced on a monthly basis. How does Swiss-Prot cope? Thanks to its biocurators.

Photo: ©Vivienne Baillie

The life of a biocurator may sound simple. But it is not. Dozens of articles have been published to define this new species of life scientist. Biocurators are professional scientists who are trained to collect, sort, synthetize, organise and validate biological information which is then disseminated via databases. The nature of their job demands the patience and thoroughness of a librarian and they have been referred to – whether endearingly or not – as 'museum cataloguers of the internet age', 'those who prefer computers to pipettes', 'self-confessed bookworms', or 'monk copyists'. A somewhat narrow-minded and outdated view.

Biocuration is not only a science in itself but, like the technology that nourishes it, it is continuously evolving. Biocurators have to adapt fast as they face the unending production of huge amounts of data and have to deal with ever-changing biological knowledge. In Swiss-Prot, for instance, although the number of protein sequences does not exceed the half million mark, it provides data that have been checked and improved by its biocurators. This upgraded information is in turn (re)used for automatic annotation of new incoming sequences. As such, biocurators have become an essential and central piece of the life science puzzle.

About 70 biocurators currently work for the SwissProt database – which is now part of the UniProtKB knowledgebase – in Switzerland (SIB), England (EBI) and the US (PIR), most of whom (50) work at SIB. They are biologists, biochemists and chemists with a strong background in wet-lab research, and generally hold PhD degrees. Their job consists in reviewing experimental and predicted data for each and every protein with a sharp critical eye, as well as verifying in detail every protein sequence. In this way, they provide a complete overview of any information there exists on a given protein.

The information is extracted from various sources; most of it regards experimental data which is drawn from the literature. Biocurators reconcile any conflicting results and then compile them into a concise, comprehensive report – both in free text and structured format – with controlled vocabularies that can be read by a machine. The process of expert biocuration adds a wealth of knowledge to UniProtKB/Swiss-Prot records, and includes information related to a protein's function, structure and subcellular location but also a wide range of sequence features such as active sites or

post-translational modifications, besides the protein's interactions with other proteins.

Huge amounts of data means tons of articles. Does a biocurator read every single publication there is on a given protein? The answer is no. A crucial step in expert curation is knowing how to identify a representative subset of publications that will provide a complete overview of the information that is available at the time. Any information added manually to Swiss-Prot is linked to its source – in this way, users can trace each piece of information to its origin. For maximum efficiency, Swiss-Prot biocurators generally deal with groups of related proteins – such as proteins that belong to the same family or are found in different species – as the background knowledge already exists within the database. Swiss-Prot biocurators also collaborate and leverage the work of complementary curated resources to ease consistency and data exchange, thereby ensuring that biocuration efforts are not duplicated.

But this is all very theoretical. How exactly do Swiss-Prot biocurators deal with the knowledge that is pouring in? The recent characterization of the Notum protein in humans and the common fruit fly, Drosophila melanogaster, provides an excellent example. One Swiss-Prot biocurator waded through over 100 articles that were linked to the term "Notum" – which also happens to be the name given to the dorsal portion of an insect's thoracic segment. According to well-established criteria, two papers were sufficient to extract the information needed to acquire a comprehensive overview of the protein.

Notum was initially characterized in D.melanogaster as a protein that had a vital role in developmental morphogenesis by inhibiting the Wnt signalling pathway – a pathway that passes signals into a cell via cell surface receptors. For over 20 years, scientists thought the pathway was inhibited by Notum via the hydrolysis of specific proteoglycans known as glypicans. However, the two selected articles contradicted these results: inhibition occurs via serine depalmitoleylation of the Wnt proteins themselves. Thus the role of Notum as an inhibitor of the pathway is confirmed, but the mechanism is different.

Thanks to this new experimental data, an update of the function – and the protein's name – was echoed across 10 different species in UniProtKB/SwissProt. But the biocuration did not stop there... New gene ontology (GO) terms were created, a new enzyme commission (EC) number was issued, the positions of the enzyme's active sites were annotated from its 3D structure in the Protein Data Bank (PDB), and the modification of the Wnt protein by depalmitoleylation and its consequences have been annotated using controlled vocabulary defined by the RESID database of Protein Modifications. And, so as not to lose track of any information even if it has become redundant, the former function of the Notum protein is described in detail under a 'CAUTION' section.

This example illustrates how essential expert biocuration is, and how the correct identification of only a few targeted publications can give rise to information that is not only vital but has a chain reaction effect. Thanks to manual biocuration, new biological functions continue to be unveiled within what are thought to be well characterized protein families. Correct and up-to-date information is spread across other databases in addition to being fed back into the automatic annotation and function prediction systems loop. Existing close and mutually beneficial collaborations between different resources are also heightened, demonstrating yet again how essential expert biocuration is in maintaining biological knowledge.

Cross-references to UniProt

Palmitoleoyl-protein carboxylesterase NOTUM, *Homo sapiens* (Human): Q6P988
Palmitoleoyl-protein carboxylesterase NOTUM, *Drosophila melanogaster* (Fruit fly): Q9VUX3

proteinspotlight
> ONE MONTH, ONE PROTEIN <
Protein Spotlight (ISSN 1424-4721) is a monthly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles descrive a specific protein or family of proteins on an informal tone.
http://web.expasy.org/spotlight
Swiss Institute of Bioinformatics