J Comput Aided Mol Des (2009) 23:195–198
DOI 10.1007/s10822-009-9260-9

WARR'S PIECE

ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI)

Wendy A. Warr

Received: 17 December 2008 / Accepted: 13 January 2009 / Published online: 5 February 2009

John Overington holds a first degree in chemistry, and studied for his PhD on comparative protein modeling, drug design, and sequence-structure relationships with Sir Tom Blundell FRS, at Birkbeck College, London. After a postdoctoral fellowship with the Imperial Cancer Research Fund, London, he joined Pfizer Central Research in the UK and eventually became manager of the Molecular Informatics, Structure and Design Department, where he was responsible for cheminformatics, structural biology, target analysis, and molecular modeling. He joined Inpharmatica in 2000; Inpharmatica was acquired by Galapagos NV in December 2006. There he was Senior Director, Discovery Informatics of the Galapagos subsidiary BioFocus DPI. In October 2008 he was appointed Team Leader, Chemogenomics at the European Molecular Biology Laboratory (EMBL), working at the European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK. His ChEMBL group's research focuses on mapping the interactions and functional effects of small molecules binding to their macromolecular targets. The group studies the interactions of pharmacologically active molecules and their receptors. In particular it builds and maintains a series of large-scale drug discovery databases, known collectively as ChEMBL.

W. A. Warr (✉)
Wendy Warr & Associates, 6, Berwick Court, Holmes Chapel, Cheshire CW4 7HZ, UK
e-mail: wendy@warr.com

Interview

WAW: I wanted to talk to you mainly about ChEMBL [1], the new chemogenomics data resource at EMBL-EBI, but let's start with a few questions about you yourself. You were at Pfizer for nearly 9 years. What did you achieve there and what led you to move on to Inpharmatica?

JPO: I worked for Pfizer initially in computational chemistry. After a research background in protein modeling my interests were in integration of biological and chemical data and in structure-based approaches to drug discovery. While I was there, the group changed its role and scope to eventually become Molecular Informatics Structure and Design. In 2000 I left large pharma for the booming biotech sector, and joined Inpharmatica. I had known some of the senior management of Inpharmatica, Mark Swindells (CSO) and Malcolm Weir (CEO), for some time, and liked the vision of an informatics-based drug discovery company. Inpharmatica had lots of high performance computing capabilities, expertise in database implementation, etc., and it offered me the chance to build large chemistry-oriented databases, and start research on Knowledge Discovery from Data (KDD).

WAW: What went wrong at Inpharmatica?

JPO: Well, I don't think anything went specifically wrong at Inpharmatica; but fundamentally the market changed quite dramatically from 2000 onwards. After our first big round of fund-raising, it became increasingly harder to raise venture capital, and the value of technology platform companies in particular suffered. Things have progressively become tougher: the cold that pharma caught in 2001–2002 has progressed now to a heavy flu! But we did build up a biology lab, and followed this with our own medicinal chemistry facilities. We discovered, and filed patents on many new targets, and potential therapeutics. For

example, we discovered around 200 novel secreted proteins in collaboration with Serono in Geneva (now Merck-Serono). We complemented this external collaborative work with in-house programs targeting various small-molecule drug targets. By this time we had built a large-scale SAR database that we called StARlite, outsourcing the data entry and performing extensive manual curation and automated indexing in-house. Essentially what we were attempting to do with this work was trying to "codify" the rules for target identification and validation, and lead optimization. We learnt a lot from our experience of outsourcing work, and the process of outsourcing is still gathering pace within the broader healthcare sector. The cost savings are compelling for certain types of work, and we continue to use outsourcing in our work at the EMBL-EBI.

I think it is difficult to build a successful commercial business model for scientific database content, other than for a very large-scale company. At Inpharmatica I would like to think that we were innovators but we were small, and consequently relatively expensive. Another issue for small informatics companies is in the increasing use of Request For Proposal (RFP) processes for informatics investments, which tend to take a lot of investment to respond to, and also apply significant downward pressure on pricing, making it difficult to retain the very best, most experienced staff.

In general there are also long lead times for significant informatics expenditure, and we found that quite often the scientists we were dealing with at our customers would either move on to a new role, or experience pressure to reduce costs, and so this became quite frustrating. In 2006 we performed a strategic review of the Inpharmatica business, and this eventually led to the sale of the company to Galapagos NV, which has a service division, BioFocus DPI. Inpharmatica became part of BioFocus DPI.

WAW: So after 6 years at Inpharmatica you spent nearly 2 years with BioFocus DPI. One assumes that it did not all work out as well as everyone had hoped.

JPO: Galapagos had nine different research sites and we were a group of just 8–10 people based in central London, so we were a relatively small part of a large contract research organization. With so many sites, there was a clear need for consolidation, and in this process there was not enough of a compelling commercial case to relocate the staff (and the very extensive computer equipment required to do our business) to the main UK site of BioFocus DPI. So options for what to do with the Discovery Informatics business were explored. Eventually this process led to the acquisition of the databases by the EMBL-EBI, funded by a Wellcome Trust strategic award. I have now transferred to the EMBL-EBI to lead the Chemogenomics group there. I think it is quite an interesting case of a reversed spin-out from academia, and it has also ensured that the data are available to all.

WAW: And the rest, as they say, is history. In July 2008, BioFocus DPI announced the transfer of its predictive drug discovery databases to EMBL-EBI, and EMBL-EBI will make these databases publicly available online to drug discovery researchers worldwide. The Wellcome Trust made this possible by awarding £4.7 million to EMBL-EBI, but how long will the money last? Does the venture have a long term future?

JPO: Yes. The grant is for seven people for five years, and includes future data updates, improved curation, and integration with other genomics resources. Based on the interest we have had in the data so far, from large pharma, SMEs and the academic sector, there is a clear, long-term need for this type of medicinal chemistry SAR data.

WAW: Inpharmatica developed databases called StARlite, CandiStore and DrugStore, and EMBL-EBI is to make these resources publicly available. Tell me a bit more about these three databases.

JPO: StARlite is a medicinal chemistry database, abstracted from the primary literature, of known compounds and their pharmacological effects: an SAR "knowledge base". It now contains around 500,000 compounds, covering about 5,000 targets. When orthologs and non-human genes are removed as targets, we currently cover about 1,700 distinct human proteins. StARlite also contains more than 2 million experimental bioactivities, and of course it is possible to complement these experimental values with easier to obtain calculated descriptors. StARlite is updated with new data monthly, and the curation is an ongoing process, indexing both new data and also fixing historical assays, targets and end-points. The growth of StARlite, in terms of numbers of distinct compounds is around 10% per annum. At the moment, I think StARlite is incredibly complementary to other large public-domain chemistry resources such as PubChem and ChemBank.

DrugStore is a database of around 1,500 known drugs (both small molecule and protein-based) and covers their indications and molecular targets. We published an analysis of the number of targets within DrugStore recently [2]. In a collaboration with the Special Program for Research and Training in Tropical Diseases (TDR) group of the World Health Organization (WHO), we have also contributed data to a database of targets against neglected tropical diseases [3].

CandiStore is a database of about 12,000 clinical development stage compounds. We are quite selective with what we attempt to cover in CandiStore: essentially the compound structure, synonyms, target, highest development stage, etc. We simply would never have the resource to develop a system that could rival the scope of coverage of commercial intelligence systems, and we think we can advance science with just this simple view. One of the key applications of CandiStore is in the area of drug repurposing; given the

slowing of overall discovery output there is a lot of interest in drug repurposing right now.

Together, these databases allow us to track the progress of a compound (or protein therapeutic) from lead optimization, through clinical development and then on to commercial launch. All the databases will be available as full downloads, and web services, and also *via* a user friendly front end. As you can imagine, there is a lot of planning for integration with other EMBL-EBI resources, such as UniProt, Ensembl, ChEBI, Intact, etc. In addition, these data will also be freely available for incorporation in commercial offerings, if there is a demand.

WAW: You said something about "when CandiStore goes live".

JPO: Yes, it's a work in progress, we have focused on getting the data capture and annotation process developed, and also populated CandiStore for particular gene families and drug classes of interest to us, for example monoclonal antibodies. However, first of all we need to recruit the staff. The group is just me in a portacabin at the moment.

WAW: Chris Steinbeck spoke about ChEMBL at the German Chemical Society's cheminformatics meeting in October. Where does Chris fit in?

JPO: Chris heads the Cheminformatics group at EMBL-EBI. His group curates the Chemical Entities of Biological Interest (ChEBI) database, which is already up and running [4]. Chris and I will work together on an integrated chemical infrastructure at the EMBL-EBI. One of the new things for me is picking up from Chris the culture of Open Source, and Open Access, a key part of the Chemistry Development Kit (CDK) project [5, 6] which Chris leads.

WAW: The press release also mentions Strudle for binding site "drugability", and Kinase SARfari and GPCR SARfari. Chris Steinbeck didn't mention these. Where do they fit in?

JPO: Strudle is a database of protein structure binding sites. In cheminformatics at the moment "ligand efficiency" is big news and binding sites also need to be "efficient". We have built some structure analysis tools on the basis of some very simple physicochemical and geometric arguments to identify interesting, or "drugable" binding sites.

The SARfari systems we have developed so far are integration portals built around particular gene families of broad interest to the pharmaceutical industry (for example kinases and GPCRs). We have integrated sequence, alignment, structural and screening data with an easy to use and develop web front end to encourage data exploration: hence the (somewhat corny) name SARfari. A key ability of the SARfari systems is that they allow the loading of proprietary data, allowing a single system to query across public and private data. As you can imagine, the SARfari data content is derived from the core databases StARlite, DrugStore and CandiStore, and it automatically benefits from improvement in data content and curation of those.

WAW: How will ChEMBL relate to other EMBL-EBI resources?

JPO: We have been regularly asked for integration with microarray data and novel genome sequence data, so a big advantage of now being at the EMBL-EBI is that there are very well established data systems in all of these areas, so it is relatively straightforward to map our SAR data onto these existing datasets. One of the exciting possibilities of our proximity is that we can potentially map completely novel genome sequences with chemical data as they are assembled and annotated.

WAW: Yes, Chris said that CDK would be made usable with ChEMBL. He also said that an open source Oracle cartridge would be developed for chemistry searching. When? What will be the impact?

JPO: The project to develop an open source chemistry data cartridge at the EMBL-EBI is actively underway at the moment. I would not venture to commit to a delivery date on Chris's behalf, but suffice it to say, within ChEMBL we would be very keen users.

WAW: What other plans are there for ChEMBL?

JPO: Another advantage of EMBL-EBI is its history of data and web-service provision and delivery over the web. It is hard to develop delivery systems that are cost effective and scale well. In contrast to me personally, EMBL-EBI has a huge amount of experience in this area, which is essential as part of the successful delivery of the grant. This frees me up to think more about further data offerings to the community, and also to apply data mining to the data we already have, to attempt to improve the overall drug discovery process.

WAW: To conclude on a personal note, I am intrigued by the photos of you with mushrooms.

JPO: I have always been passionate about nature. I used to live on the coast, and so my interests then were primarily

marine, but now living in the middle of the countryside, I go walking in the woods with the dogs and have become increasingly fascinated by the ecology and folklore of

fungi. As you can imagine, the curator in me spends too much time sorting out my mushroom photos!

WAW: Well, thank you for letting us print one, and in particular, thank you for taking the time to tell me all about your ambitious plans for ChEMBL. I wish you every success.

References

1. ChEMBL http://www.ebi.ac.uk/chembl/. Accessed November 2008
2. Overington JP, Al-Lazikani B, Hopkins AL (2006)) Nat Rev Drug Discov 5:993. doi:10.1038/nrd2199
3. Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK et al (2008) Nat Rev Drug Discov 7:900. doi:10.1038/nrd2684
4. ChEBI http://www.ebi.ac.uk/chebi/. Accessed November 2008
5. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) J Chem Inf Comput Sci 43:493. doi:10.1021/ci025584y
6. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Curr Pharm Des 12:2111. doi:10.2174/138161206777585274