Cheminformatics at the EBI: an update

Following EMBL-EBI's acquisition of the Galapagos NV large-scale drug discovery databases last year, Christoph Steinbeck and John Overington have been working hard to integrate the new data into the EBI's infrastructure.

Accompanying the transfer of the data with his own move from BioFocus DPI, a Galapagos NV subsidiary, to the position of ChEMBL team leader at the EBI, John has overseen the databases since their early days. While he and his team continue to maintain the new resource, Christoph's team have been closely involved in the integration of the chemical part of the data into the EBI's infrastructure.

Before coming to the EBI in January 2008, Christoph and his group had been tackling questions of metabolism using a cheminformatics approach. Their focus was on automating the process of metabolite structure elucidation, and also how best to encode, store, disseminate and analyse chemical data, developing a range of cheminformatics methods and algorithms.

The culmination of this was the development of the Chemistry Development Kit (CDK), an open source Java library for structural cheminformatics. As Christoph describes it, the CDK was "…a real success story. Our reputation as pioneers in open access databases and open source toolkits in cheminformatics was probably one of the reasons why we were recruited to the EBI." More than just a database, the CDK was also designed to be used for data analysis, for example building structure-activity predication methods.

On moving to the EBI, Christoph's team used the CDK to provide structure searching capabilities for the ChEBI database. The next step is to incorporating the new data. "A combination of the CDK and the new chemogenomics data will allow us to create open structure-activity models and to assist efforts in wet lab screening in areas such as library design," enthuses Christoph. "My team also created an open source chemical search engine for the new drug activity data, which will be the first open source chemistry search engine for the widely used Oracle Database system, based on the CDK."

The thread that runs throughout the work of Christoph and his team has been their focus on openness, recognising that cheminformatics and drug discovery has traditionally been seriously hampered by its closed-source and closed-access nature. As Christoph explains, "Historically, developments in chemistry were closed and protected, due to the potential impacts of discovering the next big drug. With the development of the open source movement from the late 1990s onwards, this has changed, and the sharing of software and information is beginning to be perceived as something beneficial rather than dangerous.

> "The sharing of software and information is beginning to be perceived as something beneficial rather than dangerous"

"Industry has also realised the value of collaborative pre-competitive development, which was traditionally performed in-house. We have seen a big shift by industry towards the use of open source cheminformatics software, such as the CDK, and open access data, such as the new chemogenomic data."

The ChEBI database also benefits from the EBI's Industry Programme; an open forum for life-science companies, providing training and information about EMBL-EBI resources, and a platform for discussion of pre-competitive and collaborative activities. "For us the collaborations within the EBI's industry programme is a great opportunity to put our methods and toolkits to use in a real-world scenario."

With the Chemical Biology retreat held earlier this year, bringing together experts from throughout the EMBL organisation for the first time, chemical biology certainly seems to becoming an increasing area of EMBL's activities. "The retreat gave us the opportunity to learn more about each other's work, assess the stage we're at and gave us the scope for a longer-term plan," says Christoph. "There are definitely opportunities for us all to work together and it will be exciting to watch these collaborations develop. On a personal level, I am particularly interested in seeing how our computationally developed algorithms can be verified in the lab."

Looking to the future, Christoph sees a big new challenge ahead for the EBI. "We are looking at expanding our activities in the area of metabolomics – the third large '–omics' area. We have just had an international workshop involving researchers from the main European and Canadian metabolomics initiatives. The participants unanimously expressed the need for central metabolomics resources of international scope, supported by minimal information guidelines for depositing data. We aim to work towards meeting these needs by creating, hosting and maintaining metabolomics resources to close the gap between this new field and the more advanced areas of genomics and proteomics."