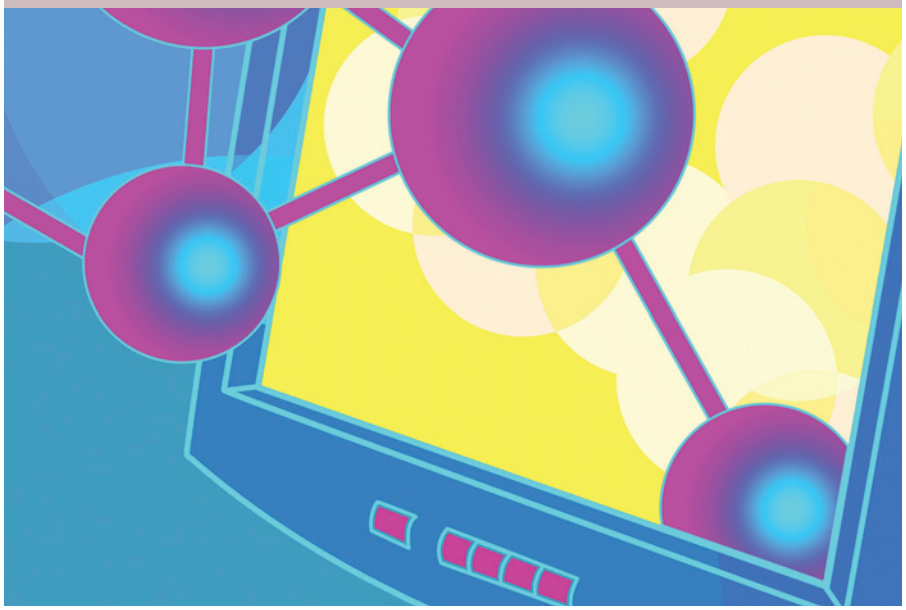
NEWS FEATURE

Dealing with a data dilemma

The pilot programme Molecular Libraries Initiative, part of the NIH Roadmap, gave academic researchers access to high-throughput screening technology and expertise. Now, as this initiative enters its next phase, a key question is how to make the most of the data generated. David Bradley investigates.



Launched in 2004, the National Institutes of Health (NIH) Molecular Libraries Initiative (MLI) set out to provide academic researchers with the tools and infrastructure to identify chemical probes for exploring biological pathways. This is in addition to potential starting points for drug discovery, particularly for neglected diseases. Such efforts are now bearing fruit. For example, in March this year a collaboration involving the Molecular Libraries Screening Center Network (MLSCN) — the MLI component that performs high-throughput screening of a large compound library for assays developed by academia — reported the identification of a novel class of lead compounds for schistosomiasis (*Nature Med.* 14, 407–412; 2008).

At the heart of the MLI is PubChem, a public database of chemical structures and their biological activities, which contains compound information from the scientific literature as well as the screening and probe data generated by the MLSCN. At present, PubChem includes information on more than 19 million chemical structures and data sets from more than 1,000 assays. Its establishment was a reflection of the key role that cheminformatics has to play in making the most of the vast amount of data generated by the MLSCN.

"Cheminformatics integrates advanced statistical, data mining and computational science approaches on the one hand, and chemical, biological, pharmacological and toxicological knowledge on the other," says Professor Alexander Tropsha, Director of the Exploratory Center for Cheminformatics Research (ECCR) at North Carolina University, USA. This ECCR is one of a group of six such centres that was given funding through the NIH Roadmap to research the development of cheminformatics methodologies and tools in support of the MLI. "Because of this unique position, the discipline actually embodies the spirit of the NIH Roadmap," Tropsha says.

Martin Griffies, a consultant specializing in informatics, agrees on the importance of strong cheminformatics capabilities. "All small-molecule drug discovery companies rely on cheminformatics as a vital component of their processes," he says. "Researchers need to be able to create, store, search and retrieve information about chemical entities, and make calculations upon their molecular properties."

Soon, the pilot MLI will become the Molecular Libraries Program (MLP). Ajay, Program Director, Cheminformatics, National Human Genome Research Institute, NIH, explains that the MLP will have almost identical goals to the MLI. The MLP will build on the pilot so that both NIH and academic screening centres will operate efficiently through three different types of centres: Comprehensive Centers, Specialized Screening Centers and Specialized Chemistry Centers.

However, some researchers are worried that the further development of cheminformatics aspects is not getting the attention and specific funding that is needed. "While the MLP provides significant funding for chemistry (for example, for building the screening library) and biology (for example, for assay development and screening), there is a significant gap in data processing and analysis, which is only in part covered by the PubChem development team," says Tudor Oprea, Co-Principal Investigator and Director of the Informatics Core, New Mexico Molecular Libraries Screening Center, part of the MLSCN, and Professor at the University of New Mexico School of Medicine, Albuquerque, USA. "That gap is about to grow wider, as more chemicals and more screens are deposited to PubChem every week."

"There is a huge amount of chemical genomics data already available from the MLP as well as from other contributors to PubChem," agrees Tropsha. "Consequently, there is a growing need for robust data analysis and knowledge discovery in expanding databases, which makes it hard to understand why the NIH would not continue to allocate targeted funding for cheminformatics research."

"The major problem comes from the misguided perception that off-the-shelf cheminformatics software can be used to address the needs of the MLP," says Oprea. Simply providing a place to deposit data does not allow it to be used optimally. "PubChem does not curate the data as deposited by screening centres," Tropsha says, "so understanding the value of the PubChem data with rather low signal-to-noise ratio requires very thorough and laborious cheminformatics approaches."

Tony Willliams of ChemSpider.com in Wake Forest, North Carolina, also notes the quality issues that are associated with freely available data. "PubChem was not meant to be a house for catalogue depositors of chemicals, but was built to house the small-molecule collections being used for screening as part of the screening initiative," he explains. "But, people dumped various quality data in there… the database is polluted with structure errors and information errors."

"Perhaps the key lesson [from the MLI] is that the data analysis in terms of uncovering trends is not an easy and straightforward

process," Tropsha adds. "The data sets deposited in PubChem are highly imbalanced, with the ratio of active to inactive compounds on average of 1:1,000 or even worse, which is typical of high-throughput screening campaigns." He points out that the false-positive rates in primary screens are also high, and many actives are not confirmed even in secondary and/or confirmatory assays.

"Although a certain component of production cheminformatics can be used to support the MLP, there is a research component — followed by development and production — that is currently not available," says Oprea. He highlights three major needs. First, chemistry-cognizant software for processing and analyses of vast amounts of data. Second, tools that consider the relationships between assays, phenotypic screens, and thus go beyond target or biochemical assays, allowing for certain degrees of fuzziness as the assay results are unreliable. And third, incorporating pathway, structural, genomics and other non-MLP information in an integrated and comprehensive manner.

Indeed, the ECCRs have identified broad areas within computational and theoretical chemistry to develop collaborative arrangements for follow-on research, as well as acting to publicize cheminformatics needs and producing publicly available tools. "One downside to cheminformatics research in the past was that a lot was commercial or proprietary and so expensive or unavailable," says Rajarshi Guha, a visiting professor at Indiana University, USA, and a data mining and cheminformatics expert. Increasingly though, "these tools can be accessed under liberal licences and used in a variety of ways," he says.

By demonstrating the possibilities in which cheminformatics and bioinformatics converge through distributed computing and Web 2.0 technologies, the ECCRs have helped to energize the academic cheminformatics community. Such efforts could drive forward the development of the field, suggests Chris Leonard, Director of Translational Research and Technology at Memory Pharmaceuticals, New Jersey, USA, who served as an external reviewer on the panel that established the MLI in 2004. "The enabling technologies for the next wave of leveraging cheminformatics to address information handling and broader prediction of off-target pharmacology or toxicology in the drug discovery process will be the blending of the semantic web (Web 3.0) and chemical structure-aware analytical tools," he says.

Although the MLP does not have a dedicated cheminformatics component currently, a workshop co-organized with the NIH on the role of informatics within the MLP, led by Irwin Kuntz at the University of California at San Francisco, USA, and Peter Wipf at the University of Pittsburgh, USA, explored the chemical and biological needs for cheminformatics as part of the NIH Roadmap. The workshop raised the need for high quality data and data storage, with secondary concerns including the need for efficient data-analysis tools and a high level of transparency in the ultimate interface so that a broad spectrum of users can access the data. "We have just initiated a process to take action on recommendations from the workshop," Ajay says, "there are of course no guarantees that the process will finally result in a specific cheminformatics (data analysis) type initiative."

Whatever the outcome, it is clear that the cheminformatics challenges need to be addressed in some way if the opportunities created by the MLP are to be fully exploited. "There are no current tools, either in industry or academia that could address all major complex challenges created by the MLP," says Tropsha. "I am sure the major methodological discoveries and important novel applications in cheminformatics are ahead of us."