Integration, Analysis and Collaboration. An Update on Workflow and Pipelining in Cheminformatics

The workflow paradigm is a generic mechanism to integrate different data resources, software applications and algorithms, Web services and shared expertise. Such technologies allow a form of integration and data analysis that is not limited by the restrictive tables of a conventional database system. They enable scientists to construct their own research data processing networks (sometimes called "protocols") for scientific analytics and decision making by connecting various information resources and software applications together in an intuitive manner, without any programming. Therefore, software from a variety of vendors can be assembled into something that is the ideal workflow for the end user. These are, purportedly, easy-to-use systems for controlling the flow and analysis of data. In practice, certainly in chemical analyses, they are not generally used by novices: best use of the system can be made by allowing a computational chemist to set up the steps in a protocol then "publish" it for use by other scientists on the Web.

Until recently, in the cheminformatics field, only two solutions were in common use for capturing and executing such multi-step procedures, processing entire data sets in real time through "pipelines" or "workflows". The two technologies, both of them commercial solutions, are from InforSense,[1] which uses a workflow paradigm in its InforSense platform, and SciTegic[2] (now part of Accelrys) which uses data pipelining in Pipeline Pilot. New entrants to the market now open up many more options.

Pipeline Pilot

In Pipeline Pilot, users can graphically compose protocols, using hundreds of different configurable components for operations such as data retrieval, manipulation, computational filtering, and display. There are three options for the interface: the Professional Client, the Lite Client, and the Web Port Client. SciTegic offers collections of components covering chemistry, ADME/Tox, chemically-intelligent text mining, decision trees, gene expression, materials, modeling, R statistics, reporting, imaging, sequence analysis, text analytics, and the software packages Catalyst, and CHARMm.

The Integration Collection provides mechanisms to link external applications and databases into a Pipeline Pilot data processing protocol. After a tool is integrated and added to the library as a new component, end-users can employ it as they would any other Pipeline Pilot component, regardless of where the external application resides or how the integration works behind the scenes. For controlling Pipeline Pilot protocol execution from proprietary or third-party applications, SciTegic provides three different software development kits (SDKs): a Java SDK, a .NET SDK, and a JavaScript SDK.

Spotfire[3] and SciTegic have coupled Spotfire DecisionSite's interactive visual analytics with Pipeline Pilot's data processing protocols. Researchers can embed Pipeline Pilot computations in DecisionSite (without any scripting or programming) and deploy these throughout the enterprise. DecisionSite users can run analyses in Pipeline Pilot without leaving the DecisionSite environment. Pipeline Pilot is supported on Linux and Windows and is used by over 200 pharmaceutical, biotechnology, and chemicals companies. Applications have been reported in the cheminformatics literature.[4,5] SciTegic has recently announced a free academic version of Pipeline Pilot.

InforSense

The InforSense platform[1] provides a scalable, analytical middleware to build, deploy and embed analytical applications into Web portals, Customer Relationship Management (CRM) systems etc., using a Service Oriented Architecture (SOA) methodology. It is built on four sets of

components: analytical workflow, analytical portal, in-database execution, and grid. Analytical workflows allow users to compose and deploy different analytical methods and data transforms without coding.

(An analytical workflow is similar to a Pipeline Pilot protocol, and a group of InforSense modules is akin to a Pipeline Pilot component collection.) InforSense has over 700 different modules that can be visually composed. The InforSense analytical portal allows workflows to be packaged as reusable, interactive Web applications. Applications built from analytical workflows can be executed in the InforSense platform or the InforSense Grid. In-database execution technology enables workflows to be embedded within a database. This has the advantages of scalability, security and data fidelity, assuming all the data are in, say, Oracle, as analytics can take place without data extraction.

There are analytical modules for data mining, and statistics and domain specific modules, GenSense, ChemSense, BioSense and ClinicalSense, for genomics, cheminformatics, bioinformatics, and clinical research, respectively. ChemSense allows combination of "best of breed" tools from multiple cheminformatics vendors, including ChemAxon,[6] Daylight,[7] MDL (now Symyx),[8] Molecular Networks,[9] and Tripos,[10] and can integrate Oracle chemical cartridges. An SDK allows users to implement custom applications. The InforSense Spotfire connector provides interaction between Spotfire DecisionSite for visual analytics and the InforSense platform.

Pharmaceutical, biotechnology, consumer goods, healthcare, financial services, manufacturing and communications companies are using InforSense. AstraZeneca and GlaxoSmithKline are reportedly using the in-database technology.[11] Windows, Mac OS, Linux and Solaris are supported (subject to the software from third parties being able to operate on those platforms).

KNIME

An interesting new entrant to the market is KNIME[12] (the "K" is silent), a modular data exploration platform that enables the user to create data flows visually, execute analysis steps, and later investigate the results through interactive views on data and models. It was developed by the group of Michael Berthold at the University of Konstanz, Germany. It already incorporates several hundred nodes as part of the standard release (for "nodes" you may care to read "modules" or "components") for data processing, modeling, analysis and mining, as well as various interactive views, such as scatter plots and parallel coordinates. KNIME supports the integration of different databases such as SQL, Oracle, and DB2. Thus, data from different database sources can be joined, manipulated, partitioned, and transformed in KNIME and finally stored again in databases.

KNIME is based on the Eclipse[13] open source platform and is extensible through its modular Application Programming Interface (API): custom nodes and types can be integrated. KNIME is licensed under the Aladdin free public license and, in essence, is available free of charge for use in both non-profit and for-profit organizations. A research group in a pharmaceutical company, for example, can download and use KNIME freely for internal data analysis but if an organization wants to make money from distributing KNIME, KNIME will want some benefit from that. For-profit partners such as Tripos[10] and Schrödinger[14] are implementing the technology and also provide KNIME support options. Commercial support is seen as a vital factor in general adoption of open source tools in cheminformatics. The KNIME developers also provide support through a KNIME community that contains useful information and discussion forums.

The Schrödinger KNIME Extensions (currently in a beta version) have more than 100 nodes covering a range of computational tools for workflows focused on both ligand- and structure-based applications. Ligand-based tools include nodes for property and fingerprint generation, similarity searching, diversity analysis, clustering, conformation generation, common pharmacophore perception, database searching, and shape-based screening. For structure-based work there are nodes for docking, homology model building, structural refinement, and binding free energy estimation as well at many tools for general molecular modeling and

structural manipulation. Schrödinger provides support for both the Schrödinger KNIME Extensions and for the KNIME platform itself.

The Tripos Chemistry Extensions for KNIME package is an initial offering that provides researchers with functionality to manipulate, analyze, and visualize chemical data. Benchware 3D Explorer, Benchware DataMiner, Concord, Confort, DBTranslate and AUSPYX can be licensed for access through the extensions package. In future, Tripos will be releasing further nodes that allow access to all Tripos science through the KNIME platform, and will continue to deliver its scientific capabilities through KNIME as well as through its traditional interfaces. The company will concentrate on the usability of the nodes rather than just on wrapping of functionality Commercial support for KNIME is available through Tripos. (Note that some Tripos tools are also accessible as components from SciTegic's Pipeline Pilot.)

More recent node additions to KNIME are the cheminformatics capability provided by ChemAxon[6] (implemented by Infocom, an IT business company of the Teijin Group), and the THINK modeling suite (Treweren Consultants).[15] Symyx/MDL recently demonstrated prototypes of enumeration and structure searching nodes. Steinbeck's team is helping the University of Konstanz to write CDK nodes. The Nodes4KNIME[16] project is a new, open source initiative to develop independent nodes for use with KNIME.

Other open source solutions

At this point it should be admitted that this article has a cheminformatics bias. Scientists in a variety of disciplines (e.g., biology, ecology and astronomy) have made extensive use of Kepler[17], but Kepler has had little impact on cheminformatics. Furthermore, open source software is commonly used in bioinformatics but users of cheminformatics systems have tended to take a more "traditional" stance. Some pharmaceutical industry users are now beginning to think that it is important to lower the barriers between bio- and cheminformatics applications; initiatives such as the "druggable genome" do require closer integration[5]. The extensibility and flexibility of the more open solutions may provide something here that the more monolithic systems cannot.

The SOMA2 open source modeling environment[18,19] has been developed at the Finnish IT Center for Science (CSC). A workflow program, Grape, and XML descriptions of scientific programs allow researchers to link molecular modeling software into workflows. SOMA2 collects the data calculated in the workflow and stores the information in Chemical Markup Language (CML)[20-26] format. An extranet interface is used for user authentication, building the program interfaces and workflows, and sorting, filtering, and visualizing the results. The SOMA2 interface also uses third-party applications. For example, molecular structures are visualized using ChemAxon's Marvin software package.

Another open source package is Taverna.[27,28] which aims to provide language and software tools to facilitate use of workflow and distributed computing technology within the e-science community. It allows a bioscientist with a limited computing background and limited technical resources and support to construct complex analyses over public and private data and computational resources, all from a standard PC, UNIX box or Apple computer. Taverna is now funded as part of the Open Middleware Infrastructure Institute UK (OMII-UK) which has a mandate to ensure the existence of a supported and sustainable foundation upon which other projects can build. Development is hosted at the EBI and Manchester University in a model that is closer to that of a software house than an academic group, so, unlike many open source projects, Taverna has some guarantee of continued existence.

One commercial user has integrated the Schrödinger tools into Taverna, running a system on the North-West grid in the UK, and was able to do this relatively quickly because of the open nature of the platform. Integrating a cheminformatics tool is no different from integrating a bioinformatics tool. Taverna does not supply tools: it provides links for tools in an easily extensible environment

(as do some other open source workflow projects). Taverna also supports internal or external sharing and reuse of workflows, for example the sharing of portals in the myExperiment project.

The CDK-Taverna[29] solution combines three open-source projects: Taverna as a workflow container, Christoph Steinbeck's Chemistry Development Kit (CDK)[30,31] as a basic chemo- and bioinformatics library of more than 100 components, and Bioclipse as an Eclipse-based result viewer. Potential workflows provided by the CDK-Taverna solution address data filtering, migration and transformation, information retrieval, QSAR/QSPR or pharmacophore related tasks, data analysis (statistics, clustering, computational intelligence), analytical and spectroscopical support, and molecular modeling. The CDK-Taverna solution aims to provide a "best of both worlds": to be as flexible and extensible as software libraries such as CDK, and as user-friendly as professional, industrial IT systems.

Rajarshi Guha at Indiana University is working on a flexible and generalizable approach[32] to the deployment of predictive models, based on a Web service infrastructure using R. The infrastructure allows users to access the functionality of these models using a variety of approaches ranging from Web pages to workflow tools. This approach is lower level, and requires programming, but is more general than an approach such as KNIME. Guha feels that one of the disadvantages of KNIME is lack of support for Web service nodes. Such nodes would allow very easy integration of much functionality, especially in bioinformatics.

Differences between pipelining and workflow

Data pipelining is a specific form of workflow. In InforSense's workflow methodology, task 1 is completed then the data are handed off to task 2 which is completed before the data are handed on to task 3 and so on. In pipelining, task 1 is completed on compound 1 and the data are passed to task 2. Task 1 can then start on the next compound. All the data are never passed all at one time and there might only be a few records in memory at any one time. The process can scale without impact on memory and efficiency is gained if a downstream operation can be commenced on some records while an upstream operation is still working on others. With a workflow method, it is easier to "cache" the data (save them for security reasons or for future use) at the end of each task.

Like InforSense, KNIME also adopts a table-by-table approach, rather than a row-by-row approach (where a row represents one compound and its data, in a table of compounds), so KNIME too needs to store the result of each node in a file in most cases, whereas Pipeline Pilot behaves like a real pipeline. It is not necessary, however, to write to disk in KNIME: the default is to write to memory making the node execution faster. In terms of reproducing experiments and wrapping up workflows with data, caching is a plus from a legal standpoint and could have appeal to chemistry management. From the technical viewpoint, there are pros and cons for both methods: if you are doing clustering and generating a pairwise distance matrix then you need all the rows; if you are calling an external program then you probably do not want to call it multiple times.

The developers of KNIME say that table-by-table processing offers substantial benefits such as multiple iterations over the same data, which is rather important for many data mining algorithms; the ability always to view intermediate results on the connections between nodes even after the workflow has been executed; and the ability to restart the workflow at any intermediate node if the user, for example, changes some settings in the middle. The penalty is the need to store the data somewhere. KNIME tries to be smart about this and only stores the differences between consecutive nodes, but it ultimately stores the data on disk. SciTegic has pointed out that in data pipelining, a cache (of all the data) can be added to the components as a "finish here and resume" component.

This is not the place to delve deeper into computer science issues. The tabular data structure approach in KNIME was singled out for mention because of the current commercial interest in

KNIME. The reader who wishes to go more deeply into workflow technologies should also look into the Collection-Oriented Modeling and Design (COMAD) paradigm in Kepler, and into Taverna's rich iteration semantics which allow some complex operations to be expressed very simply.

Commercial considerations

As an over-generalization, one could say that SciTegic is supplying "vertical solutions", specializing in cheminformatics, while InforSense is spread across a group of diverse "horizontal" fields. InforSense does report many users in pharmaceutical companies, and ChemSense has the advantage of being vendor-agnostic, but Pipeline Pilot has made more impact with computational chemistry departments, partly, perhaps, because of the established reputation of Accelrys with those groups of users.

InforSense's marketing pitches use terms such as "business intelligence", "embedding intelligence throughout the enterprise" and "pervasive, process driven and predictive analytics". The company is aiming not just at pharmaceutical research, but also at medical research centers, sales and marketing groups, and operational risk compliance programs. It has had large injections of vendor capital of late, allowing it to diversify in a big way. Its in-database technology is probably of more importance in some markets than in others. Pipeline Pilot does not have in-database technology, but SciTegic has pointed out that, in practice, in cheminformatics, researchers tend to have data in all sorts of files.

It is probably fair to say that Pipeline Pilot is the market leader in the cheminformatics niche (there are only two hits for InforSense in SciFinder but more than 20 for Pipeline Pilot). SciTegic does have customers who use the infrastructure and reporting for completely different tasks (for example, one company has built a CRM from it), and some customers in the imaging and bioinformatics areas are not in the field of cheminformatics. SciTegic and its customers, have, however, focused their scientific publications on cheminformatics. Recently, SciTegic like InforSense, has started to talk of business intelligence, at least in terms in terms of "Scientific Business Intelligence". Accelrys appears to have trademarked this seemingly generic term.

Some people have an impression that commercial solutions have focused on IT infrastructure, perhaps a good theme on which to focus commercially. InforSense emphasizes database technology. Accelrys' own algorithms are delivered through Pipeline Pilot with Discovery Studio or Accord for Excel as the graphical user interface. Pipeline Pilot acts as the middle tier of Accelrys' architecture, sitting on top of databases and computing servers with molecular modeling code. It is certainly true that many users consider the two commercial solutions expensive. The different licensing model of KNIME is thus attracting interest. Equally important is its reported user-friendliness.

Not only does KNIME allow "best of breed" rapid deployment of workflows but as the pharmaceutical industry moves to an environment of shared risk, outsourcing and collaboration then it needs to be using tools, models, and frameworks that are cheap and simple to deploy. KNIME provides a company with a workable solution to share its models with organizations that cannot afford a complete, commercial modeling suite.

KNIME bridges the gap between the expensive commercial solutions that big pharmaceutical companies can afford, and solutions such as Taverna at the other extreme. KNIME does have some free cheminformatics nodes such as CDK, but Taverna is truly free and its developers have a strong preference for the open source philosophy, although there is, in practice, no barrier to interoperability with closed source systems (such as in the system built with Schrödinger nodes). Vendors of cheminformatics components claim that there is little demand from their customers for wrapping the components up in Taverna implementations but there are certainly some users of Taverna in the pharmaceutical industry, perhaps without formal management backing. Changing requirements, such as big pharma's need to share components and workflows with collaborating

organizations, and the need to lower some of the barriers between bio- and cheminformatics applications, suggest that the workflow paradigm has an interesting future.

Acknowledgment

References

1. InforSense. http://www.inforsense.com (accessed November 30, 2007).
2. SciTegic. http://www.scitegic.com/ (accessed November 30, 2007).
3. Spotfire. http://www.spotfire.com (accessed November 30, 2007).
4. Brown, R.; Varma-O'Brien, S.; Rogers, D. Data Pipelines and Virtual Screening: Automating the Process. *QSAR Comb. Sci.* 2006, *25*(12), 1181-1191.
5. Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L.. Global Mapping of Pharmacological Space. *Nature Biotechnology* **2006**, *24*, 805-815.
6. ChemAxon http://www.chemaxon.com (accessed November 30, 2007).
7. Daylight Chemical Information Systems. http://www.daylight.com (accessed November 30, 2007).
8. Symyx (acquired MDL October 2007). http://www.symyx.com (accessed November 30, 2007).
9. Molecular Networks. http://www.molecular-networks.com) (accessed November 30, 2007).
10. Tripos. http://www.tripos.com (accessed November 30, 2007).
11. May, M. Working Out the Flow. http://www.bio-itworld.com/issues/2006/sept/cover-story (accessed November 30, 2007).
12. KNIME. http://www.knime.org/(accessed November 30, 2007).
13. Eclipse. http://www.eclipse.org/ (accessed November 30, 2007).
14. Schrödinger. http://www.schrodinger.com (accessed November 30, 2007).
15. THINK. http://www.treweren.com/ (accessed November 30, 2007).
16. Nodes for KNIME. http://sourceforge.net/projects/nodes4knime/ (accessed November 30, 2007).
17. The Kepler project. http://kepler-project.org/ (accessed December 6, 2007).
18. SOMA2. http://www.csc.fi/soma) (accessed November 30, 2007).
19. Lehtovuori, P. T.; Nyronen, T. H. SOMA - Workflow for Small Molecule Property Calculations on a Multiplatform Computing Grid. *J. Chem. Inf. Model.* **2006**, *46*(2), 620-625.
20. Murray-Rust, P.; Rzepa, H. S. *Chemical* Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928-942.
21. Murray—Rust, P.; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. 2. Information Objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1113 -1123.
22. Gkoutos G. V.; Murray-Rust, P.; Rzepa, H. S.; Wright, M. Chemical Markup, XML, and the Worldwide Web. 3. Toward a Signed Semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124-1130.
23. Murray-Rust, P.; Rzepa, H. S.; Wright, M. Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content. *New J. Chem.*, **2001**, 618-634.
24. Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. 4. CML Schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757-772.

25. Murray-Rust, P.; Rzepa, H. S.; Williamson, J.; Willighagen, E. L. Chemical Markup, XML and the Worldwide Web. 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 462-469.
26. Holliday, G. L.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. 6. CMLReact; An XML Vocabulary for Chemical Reactions. *J. Chem. Inf. Model.* **2006**, *46*, 145-157.
27. Taverna. http://taverna.sourceforge.net (accessed November 30, 2007).
28. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P. Taverna: a Tool for the Composition and Enactment of Bioinformatics Workflows. *Bioinformatics* **2004**, *20*(17), 3045-3054.
29. CDK-Taverna. http://www.cdk-taverna.de (accessed Deember 6, 2007).
30. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK). An Open-source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*(2), 493-500.
31. Steinbeck C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L.. Recent Developments of the Chemistry Development Kit (CDK) - an Open-source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*(17), 2111-2120.
32. Guha, R. A Flexible Web Service Infrastructure for the Development and Deployment of Predictive Models. *J. Chem. Inf. Model.* in press.

Dr. Wendy A. Warr, Wendy Warr & Associates (wendy@warr.com, http://www.warr.com), December 2007