J Comput Aided Mol Des (2012) 26:801–804
DOI 10.1007/s10822-012-9577-7

WARR'S PIECE

Scientific workflow systems: Pipeline Pilot and KNIME

Wendy A. Warr

Received: 19 April 2012 / Accepted: 1 May 2012 / Published online: 27 May 2012

There are many examples of scientific workflow systems [1, 2]; in this short article I will concentrate only on cheminformatics applications and the workflow tools most commonly used in cheminformatics, namely Pipeline Pilot [3] and KNIME [4]. Workflow solutions have been used for years in bioinformatics and other sciences, and some also have applications in so-called "business intelligence" and "predictive analytics". Readers can find details of Discovery Net, Galaxy, Kepler, Triana, SOMA, SMILA, VisTrails, and others on the Web. Kappler has compared Competitive Workflow, Taverna and Pipeline Pilot [5]. Taverna has been widely used in bioinformatics but is also used with the Chemistry Development Kit (CDK) [6, 7]. CDK-Taverna workflows are made freely available at myExperiment.org [8]. (myExperiment.org also includes KNIME workflows.)

DiscoveryNet was one of the earliest examples of a scientific workflow system; its concepts were later commercialized in InforSense Knowledge Discovery Environment (KDE). My 2007 review [1] centered on Pipeline Pilot and InforSense KDE; KNIME was then a relative newcomer. In 2009 the loss-making InforSense organization was acquired by IDBS and KDE has made progress in translational medicine [9]. InforSense's ChemSense [10] used ChemAxon's JChem Cartridge, and ChemAxon chemical structure, property prediction, and enumeration tools. ChemSense's three major pharmaceutical customers have turned to other solutions. The InforSense Suite lives on but it not seen as a "personal productivity tool"; rather it is integrated into the IDBS ELN platform. KNIME and Pipeline Pilot are now the market leaders in personal productivity in cheminformatics.

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

Comparison

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 [11]. KNIME has a graphical user interface for combining "nodes" [12]. Collections of nodes are known as "extensions". KNIME is based on the Eclipse [13] open source platform, and Java. Java is part of the foundation of Pipeline Pilot and programmers can create new components with the Java components API or write new clients against the Java SDK. In addition, Pipeline Pilot has its own scripting language (for "non-programmers"); it has much more cheminformatics technology built in and scripting is more concise. There were initially very few chemistry nodes for use with KNIME, and adding a new one required Java programming, but many more nodes are now being added.

KNIME uses a workflow methodology in which 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 (as in Pipeline Pilot), task 1 is completed on compound 1 and the data are passed to task 2. Task 1 can then start on the next compound. In short, the data stream from 1 to 2. 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. The table-by-table processing of KNIME offers benefits such as multiple iterations over the same data (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. The penalty is the need to store the data somewhere, but it is easier to cache the data at the end of each task. In data pipelining, a cache of all the data can be added as a "finish here and resume" component.

KNIME came into the market from a data mining background while Pipeline Pilot came from cheminformatics. In practice, Pipeline Pilot and KNIME are complementary. In some markets they do not compete at all (Pipeline Pilot is not aimed at non-scientific applications) and Pipeline Pilot has a separate role to play within Accelrys' software portfolio. As a gross generalization, users say that Pipeline Pilot is very expensive but easy to use, while KNIME is free (or less expensive), and less easy to use. KNIME would counter by saying that "ease of use" is a subjective criterion, and familiarity with another system may have a bearing on it. Accelrys would argue that on a total cost of ownership basis (including factors such as IT costs, developer costs, and support levels), rather than initial purchase cost alone, the differential between Pipeline Pilot and KNIME is not as great as it first appears. Pipeline Pilot is very memory efficient; KNIME is not as scalable (although that issue is being addressed, as we shall see later). Some might say that Pipeline Pilot is "professional" while KNIME suffers from its non-commercial background, but others actually prefer the open source nature and community spirit of KNIME.

KNIME

At the outset in 2004, KNIME was not aimed specifically at cheminformatics but it was initially taken up by the cheminformatics community. Nowadays fifty percent of users are in other disciplines; KNIME is a "business intelligence" or "predictive analytics" product. *The Economist* uses KNIME in customer relationship management. A major telecom uses it in social networks and text mining. Private banks in Zürich and the Grand Casino in Lucerne use it. The Pasteur Institute is adding sequencing extensions. Small biotechs and pharma are using KNIME. In all there are 9,000 registered users or organizations and over 500,000 copies have been downloaded [14].

KNIME is not sold: it is free. The business model [15] involves licensing enterprise components allowing users to exchange workflows and build Web portals. KNIME already has a free reporting engine, for example, but for corporate-wide use, the KNIME server adds value with the WebPortal. Users who do not pay for KNIME can call Web Services but enterprise users can make better use of Web Services. The company was formed in 2006 and moved to Zürich in 2008. It is small but profitable: it has 15 staff, most of them involved in technical development. There are currently three openings for new staff to relieve developers of commercial pressures, but KNIME does not need a big sales and marketing resource: companies call up KNIME spontaneously. Often they are already using the free software.

KNIME has a non-exclusive technology partnership with Perkin Elmer Informatics which supports the enterprise KNIME solution and is a global distributor. (There are five other local distributors.) Fifteen KNIME partners (e.g., Schrödinger, Tripos, Infocom (for ChemAxon), BioSolveIT, Chemical Computing Group, Cresset, Dotmatics, Molecular Discovery, and Molegro) have added cheminformatics tools. There is a KNIME–Spotfire bridge that allows users to call KNIME workflows from within Spotfire. From 2010, a community aspect [16] has been growing. CDK [17], Indigo [18] and RDKit [19] nodes have been added, the RDKit ones through Novartis.

KNIME has a broad spectrum of contributors from software vendors, academia and pharma. Collaborations between pharmaceutical companies are becoming more commonplace: OpenPHACTS [20] and the Pistoia Alliance [21] are examples. There is an informal KNIME precompetitive pharma group that includes teams within AstraZeneca, Boehringer Ingelheim, Evotec, Lilly, Novartis, Pfizer, Sanofi-Aventis, Syngenta and Vernalis. Through Erl Wood Informatics, the Lilly group has made 30 nodes open source. These include format converters, fingerprinting, docking, viewers, R-group analysis, matched pairs, scoring and ranking, multi-objective optimization, reaction vectors [22] and activity cliffs.

Many proprietary nodes are available in-house at Lilly. Using KNIME, Lilly can present a common interface for BioSolveIT and Cresset software; cheminformatics and data mining tools can all be mixed in one desktop environment. Companies use KNIME in different ways. Lilly uses it to deliver applications to chemists while other companies use it more in computational chemistry. Novartis has developed a collection of KNIME nodes for working with internally developed algorithms and services. Beyond computational chemistry, KNIME is used within the Novartis Research IT department for tasks like tracking and reporting on the utilization of high-performance computing resources. At Boehringer-Ingelheim both Pipeline-Pilot and KNIME are used. Automated calculation engines are partly deployed via KNIME; nodes developed in-house are used to interface with the underlying infrastructure. Some companies publish protocols in Pipeline Pilot and consume them in KNIME, a solution that introduces complexity, for instance, in terms of communication between both tools, but does allow users to combine the advantages of both tools in a tailored fashion.

Recently KNIME has added the facility to pipeline data, useful in applications such as image processing where there is too much data to store. Distributed processing is also being addressed: RushAnalyzer [23] is a combination of KNIME and Pervasive DataRush, for faster processing of large data sets, focusing particularly on analysts without skills in parallelism.

Pipeline Pilot

Pipeline Pilot competes with KNIME where scientists need personal productivity tools, but a separate big opportunity for Accelrys is use of Pipeline Pilot to build applications such as chemical registration and search. Accelrys is building a next generation search and decision support application based on a Pipeline Pilot middle tier (in a three-tier architecture). The new system will be more open than earlier software: in the new development environment, programming expertise will not be needed to insert new modules or bolt on in-house code. Enterprises with their own software will be able to write a Pipeline Pilot protocol incorporating it, or they could use the Pipeline Pilot integration collection to integrate it. An example is Symyx' Cheshire business rules for chemical structures. Users need to have a high level of confidence in the security of their corporate compound collections; Pipeline Pilot is designed to address such security issues.

It is Accelrys' intention that Pipeline Pilot becomes the underlying architecture for all the company's software solutions. Pipeline Pilot is already integrated in Isentris: Isentris users can thus benefit from more sophisticated calculations, for example, without realizing that they are not working in Isentris. Accelrys has also benefited from technologies developed by Symyx. (Symyx and Accelrys have been one company since 2010.) Symyx' enhanced chemical representation is being built into Pipeline Pilot, together with high quality structure rendering.

Initially Pipeline Pilot was used mainly in cheminformatics; now it is used in other scientific fields such as next generation sequencing (NGS) and imaging. Oxford Nanopore Technologies has developed a disposable DNA sequencing device the size of a USB memory stick whose low cost, portability and ease of use are designed to make DNA sequencing universally accessible. The company chose Pipeline Pilot and its NGS Collection because Pipeline Pilot can scale in "big data" deployments such as this.

There are many computational chemistry publications the authors of which used tools such as fingerprinting, Murcko scaffolds, and Bayesian modeling supplied with Pipeline Pilot. Since the main theme of these papers is not actually Pipeline Pilot itself, and it is not possible to do a comprehensive search in SciFinder, I have chosen not to make a selection here. Some companies have written their own nodes for Pipeline Pilot because they preferred other methodologies. Pipeline Pilot is reportedly more powerful than KNIME in reporting: one customer is very satisfied with interactively analyzing data and reporting in HTML, with useful features such as selecting parts of plots and viewing the molecules in a "Spotfire-like" environment, without Spotfire.

Spotfire and Accelrys coupled DecisionSite's interactive visual analytics with Pipeline Pilot so that researchers can embed Pipeline Pilot computations in DecisionSite (without any scripting or programming), create their own Guides, and deploy these throughout the enterprise so that DecisionSite users can run analyses in Pipeline Pilot without leaving DecisionSite. Alternatively, Pipeline Pilot users can view protocol results in the Spotfire Viewer. Another component, the Spotfire Filter, allows users to view intermediate protocol results in Spotfire, select a subset of the data, for example, and then continue processing them in Pipeline Pilot. Similar technology works with the version of Tibco Spotfire that is replacing DecisionSite.

ChemAxon and Accelrys compete in terms of chemical cartridges but ChemAxon is also an independent software vendor partner of Accelrys and has a very good component collection for Pipeline Pilot which provides access to 16 different ChemAxon tools, is free of charge, and is developed and supported by ChemAxon.

Conclusion

It is interesting to contrast the concept of "you get what you pay for" with the current trend towards open source software, pre-competitive alliances, and community contributions. I do not think that KNIME will go out of business, and much of its business is already outside cheminformatics. I doubt that Pipeline Pilot will become a lot cheaper in the near future; moving from a specialist high tag item into a commodity product with broad appeal raises challenges. KNIME thinks that the data backbone really needs to be commodity and the science (which may or may not be proprietary) plugs into that. What often happens when a market leader discovers a strong competitor is that users benefit from the new features that the vendors introduce in response to competition. In so far as KNIME and Pipeline Pilot are in competition, rather than complementary, this could be good news for us all.

Acknowledgments I am grateful to Alex Allardyce, Michael Berthold, Michael Bodkin, Robert Brown, Val Gillet, Robert Glen, Johannes Kirchmair, Jan Kriegl, Greg Landrum, Andrew Lemon, Phil McHale, Chris Molloy, Cameron Neylon, Krisztian Niesz, Jens Schamberger, Chris Swain, and Ton van Daelen for helpful advice and useful discussions.

References

1. Warr WA (2007) Workflow and Pipelining in Cheminformatics. http://www.qsarworld.com/qsar-workflow1.php. Accessed 15 April 2012
2. Shon J, Ohkawa H, Hammer J (2008) Scientific workflows as productivity tools for drug discovery. Curr Opin Drug Discov Dev 11(3):381–388
3. Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/. Accessed 15 April 2012
4. KNIME. http://www.knime.com. Accessed 15 April 2012
5. Kappler MA (2008) Software for rapid prototyping in the pharmaceutical and biotechnology industries. Curr Opin Drug Discov Dev 11(3):389–392
6. Kuhn T, Willighagen EL, Zielesny A, Steinbeck C (2010) CDK-Taverna: an open workflow environment for cheminformatics. BMC Bioinformatics 11:159
7. Truszkowski A, Jayaseelan KV, Neumann S, Willighagen EL, Zielesny A, Steinbeck C (2011) New developments on the cheminformatics open workflow environment CDK-Taverna. J Cheminform 3:54
8. myExperiment. http://www.myexperiment.org/. Accessed 15 April 2012
9. IDBS Omics Workbench http://www.idbs.com/products-and-services/inforsense-suite/omics-workbench/ Biomolecular Hub http://www.idbs.com/products-and-services/inforsense-suite/biomolecular-hub/ and Enterprise Translational Medicine Solution http://www.idbs.com/solutions/healthcare/enterprise-translational-medicine/. Accessed 15 April 2012
10. ChemSense. http://www.idbs.com/products-and-services/inforsense-suite/chemsense/. Accessed 15 April 2012
11. Hassan M, Brown RD, Varma-O'Brien S, Rogers D (2006) Cheminformatics analysis and learning in a data pipelining environment. Mol Divers 10(3):283–299
12. KNIME tutorial. http://macinchem.org/reviews/knime-tutorial.php. Accessed 15 April 2012
13. Eclipse. http://www.eclipse.org. Accessed 15 April 2012
14. Taylor J (2012) http://jtonedm.com/2012/01/23/first-look-knime-analytics-workbench-update/. Accessed 15 April 2012
15. Riehle D (2009) The single-vendor commercial open source business model. http://dirkriehle.com/publications/2009-2/the-commercial-open-source-business-model/. Accessed 15 April 2012
16. KNIME community contributions. http://tech.knime.org/community. Accessed 15 April 2012
17. CDK. http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page. Accessed 15 April 2012
18. The Indigo cheminformatics toolkit. http://ggasoftware.com/opensource. Accessed 15 April 2012
19. RDKit. http://www.rdkit.org/. Accessed 15 April 2012
20. OpenPHACTS. http://www.openphacts.org/. Accessed 15 April 2012
21. The Pistoia Alliance. http://www.pistoiaalliance.org/. Accessed 15 April 2012
22. Patel H, Bodkin MJ, Chen B, Gillet VJ (2009) Knowledge-based approach to de novo design using reaction vectors. J Chem Inf Model 49(5):1163–1184
23. Taylor J (2012) http://jtonedm.com/2012/03/29/first-look-pervasive-rushanalyzer/. Accessed 15 April 2012