Smart Mining of Drug Discovery Information: 1. A web service and workflow infrastructure

Xiao Dong, Kevin Gilbert, Rajarshi Guha, Jungkee Kim, Marlon Pierce, Geoffrey C. Fox and David J. Wild*

Indiana University School of Informatics, Community Grids Laboratory, and Chemical Informatics Cyberinfrastructure Collaboratory
901 East Tenth Street, Bloomington, Indiana 47408

ABSTRACT

The vast increase of pertinent information available to drug discovery scientists means that there is strong demand for tools and techniques for organizing and intelligently mining this information for manageable human consumption. At Indiana University, we are developing techniques for "smart mining" of this information, based on web services, workflows, and a variety of client interfaces. In this paper, we introduce our model and describe how workflows of web services can be used to achieve the first steps of this vision, including bridging chemical and biological information.

INTRODUCTION

Recent technological developments such as high-throughput screening, microarray assays and combinatorial chemistry have vastly increased the amount of scientific information available without necessarily providing a clear way of using it effectively (something which has been dubbed data overload[1], and which is a great concern of managers and scientists in the pharmaceutical industry). Scientists have to deal with large volumes of many kinds of information coming from diverse sources, with limited experience of how this information can most effectively be interpreted and applied. This is not helped by the fact that research computing in drug discovery is, unlike systems under regulatory oversight, very fragmented and heterogeneous[2].

At Indiana University's School of Informatics, we are engaging in a wide-ranging effort to address the issues of handling large volumes and diverse sources of chemical information for academia and industry, using web service technologies. In particular, we envision a new model of data mining that pushes relevant information to pharmaceutical scientists based on straightforward expressions of needs, using whatever tools and techniques are needed to answer the queries, rather than relying on them stumbling upon it using traditional tools and databases.

WEB SERVICE TECHNOLOGIES

Recently developed web technologies include standards for the communication of metadata and meaning (XML and derivatives), and the access of applications remotely

* Corresponding Author: djwild@indiana.edu

(Web Services, including SOAP and WSDL). These technologies are considered part of the next wave of internet usage known as the *Semantic Web*.[3, 4]

XML[5] (eXtensible Markup Language) is a markup language similar to HTML, but which conveys metadata (i.e. information about the data). XML tags can be included in HTML documents, wrapping around different kinds of data and describing its meaning: for example, a person's name might be represented in XML as <NAME>Fred Bloggs</NAME>, the NAME tags encapsulating the data and describing its type. In this way, information relevant to a particular application or domain can be automatically extracted from an HTML or a pure XML document. XML is designed to allow domain-specific subsets, which can be defined by *XML schemas*. Further recent developments of XML include languages for describing rules and ontologies on the web, thus enabling complex forms of knowledge representation.

A byproduct of XML is RSS,[6] that popularly stands for *Really Simple Syndication*, although the origin of the acronym is disputed. RSS is a simple system for information aggregation, which involves websites creating a simple HTTP accessible XML file that describes the articles available on the site. RSS Aggregators can then run on users' machines, and scour these XML files for the addition of new articles or pages which may be of interest to the user (e.g. by looking for keywords). RSS is interesting in that it gives the appearance of a push model (i.e. pro-actively finding and presenting information to a user) but it operates using a browsing model by repeatedly pulling the XML files from sites of interest. Most recent versions of web browsers are including RSS aggregators. Other languages of interest are OWL, which allows the development of ontologies, and RDF, which allows the relationships between resources to be described.

The use of XML-derived standards for chemistry and chemical informatics applications including the development of Chemical Markup Language (CML), an XML schema for Chemistry, and applications of RSS, has been documented in a series of papers by Peter Murray-Rust and Henry Rzepa[7-12].

Web services are an emerging way of aggregating and integrating data sources and software. Web services allow software applications and data sources to be published on the internet (or on intranets), thus making tools and data widely available with a standardized interface, and facilitating the construction of applications that employ distributed resources and data to solve complex tasks. Three standards have emerged for creating web services: Web Services Description Language (WSDL) is an XML-based standard for describing web services and their parameters; Simple Object Access Protocol (SOAP) "wraps around" existing applications to map abstract interfaces in WSDL to their actual implementations; Universal Discovery, Description and Integration (UDDI) effects the publishing and browsing of web services by user communities.

The idea of using of web services is just beginning to take hold in the life sciences[13]. Much of the initial groundwork has been in bioinformatics (to the extent that several service providers such as the EBI[14] now offer their tools as web services). There has been less adoption in chemoinformatics, but this is now starting to change, particularly with

the introduction of CML and InChI[15] representations for chemical structure information. In particular, the Murray-Rust group at Cambridge[16] has carried out much of the foundational research in this area.

The emergence of workflow tools that can employ chemoinformatics functionality exposed through SOAP, in particular Scitegic's Pipeline Pilot and Inforsense KDE, has resulted in some degree of exploitation of web service workflows in the pharmaceutical industry, particularly for simplifying the task of using applications together. The use of Pipeline Pilot has been documented in a number of recent papers, which particularly focus on the use of the descriptors and Beyesian classifier supplied with the tool for virtual screening.[17-21]

A SMART MINING MODEL

Our approach to harnessing large volumes of information from diverse sources is based around a three-layer model, depicted in Figure 1. First, we have a *web service layer* which provides a set of databases and computational tools that are exposed with a SOAP & WSDL interface. These databases and tools can be hosted on any internet-connected machine, and so this layer is a conglomeration of services that we have created at Indiana, and services at other locations throughout the world. We are constantly adding new services to add new functionality. We have adopted a minimal set of standards for the input and output of data to and from services, including the use of CML and/or SMILES for chemical structure information, VOTables for purely numeric data, and the use of Uniform Resource Indicators (URI's) for the passing of data (i.e. a link to the data is passed, not the data itself). Second, we are implementing an *aggregation layer* in which web services are packaged together to perform more complex tasks, usually as workflows. We are using the open source Taverna package, being developed as part of the UK's eScience initiative, to create workflows. Although Taverna is primarily an interactive tool, these workflows can be run in a non-interactive execution environment, and thus can themselves be exposed as aggregate web services. Third, we plan to implement an *interaction layer*, in which smart clients, email clients and portlets will be used to allow scientists to employ these services and aggregated workflows effectively. Further, semantic web languages such as RDF and OWL will permit the mapping of concepts in natural language and other human-derived representations with those in workflows.

WEB SERVICE INFRASTRUCTURE

We are developing web service wrappers around as much chemoinformatics functionality as we can, in order to maximize flexibility in creating workflows. These can be considered in four categories as listed below. Except where noted, all our web services are available to be used by other members of the academic community.

Database Services. We maintain a Linux server running the PostgreSQL database system, with the gNova CHORD cartridge installed to allow chemical structure searching. Several databases are exposed through web service wrappers, in particular we

maintain a copy of the NCI Discovery Therapeutics Program Human Tumor Cell Line dataset (henceforth referred to as the TCL set) which contains approximately 200,000 compounds, around 40,000 of which have associated screening results for 60 tumor cell line assays. We are currently using this database as a surrogate for high-quality High Throughput Screening data. We also keep a local "sandbox" copy of the PubChem database which is chemical structure searchable through the web service interface and which is regularly updated.

Commercial Chemoinformatics Services. We have created web service wrappers around several commercial chemoinformatics tools that we have generously been permitted to use by OpenEye Inc. and Digital Chemistry Ltd. These services are currently only available within the Indiana University environment and include OpenEye FRED (for docking), OMEGA (for 2D to 3D conversion), FILTER (for property calculation and filtering), and Digital Chemistry Divisive K-Means (for clustering).

Services from Cambridge University. We have a close working relationship with the Murray-Rust group at Cambridge University that is one of a small number of sites that has pioneered semantic web approaches to chemoinformatics. We have implemented several of their web services locally, including InChIGoogle, InChIServer, CMLRSSServer, and OSCAR3 (for automatic detection and conversion of chemical structure names).

Other services. Other services we have implemented include wrappers around Chemistry Development Kit (CDK) functionality (contributed by one of the authors, Rajarshi Guha), a variety of services relating to the R statistical package, and a special modified web service implementation of ToxTree for toxicity flagging. We have also implemented web services that allow conversion of tabular information to and from the VOTABLES format and visualization of tabular information using VOPLOT.

WORKFLOW EXAMPLES

Below we describe three of the workflows that we have developed using our current selection of chemoinformatics and related web services: these three were chosen because they represent, respectively, examples of replicating existing tools with web service workflows, using workflows to bring together functionality in new ways, and packaging functionality in a way which may prove directly useful for scientists. Web services were mainly implemented locally at Indiana University, although there is no technical requirement that this be so. Workflows were implemented using the Taverna package.

Simple HTS data flagging and organization. For this workflow, we simply aimed to replicate some of the steps normally applied in post high-throughput screen chemistry follow-up decision making, namely compound flagging/filtering, organization, and visualization. This is the kind of workflow one might commonly find applied in industry using a tool like Pipeline Pilot. The steps in this workflow are:

1. Extract the compounds and data from the TCL set for a particular screen using the PostgreSQL web service
2. Calculate Rule-of-Five parameters (molecular weight, LogP, acceptors, donors) and Polar Surface Area using the FILTER service
3. Create toxicology flags using the TOXTREE service
4. Organize the structures and data into clusters using the Divisive K-means clustering service
5. Visualize the compounds and data by cluster using VOPLOT.

Relating cellular screening and docking results. In this workflow, we attempted to bring together chemoinformatics techniques that might not normally be used together. Specifically, we were interested in whether docking results could be correlated with compounds' cellular biological screening results in such a way that possible mechanisms of action might be proposed. We have created a workflow that enables the following steps:

1. A protein-ligand complex suspected to be related to the cellular assay of interest is selected (e.g. from the PDB)
2. The ligand is extracted from the complex, and is used as a query for a 2D similarity search on the TCL set
3. The most similar compounds in the TCL set are converted to 3D conformers (using the OMEGA service) and docked into the original protein (using the FRED service).
4. The docking scores were then passed into R services to test for correlations with cellular activity in the TCL set
5. In a separate path, the 3D docked complexes were made available for visualization using the open source JMOL visualization tool in a web page.

This workflow is depicted in Figure 2. We are currently examining the scientific usefulness of this workflow specifically for finding relationships between Kinase inhibition properties of compounds and activity in tumor-related cellular screens.

Mining the scientific literature for docking. This workflow was motivated by a desire to attempt to answer a specific question a scientist might answer: given a 3D protein structure, what compounds in the literature might bind to that protein? As a proof-of-concept, we assembled all of the PubMed abstracts for the 2005-2006 year (around half a million abstracts) and fed them into the OSCAR3 service. OSCAR3 searches for chemical structure names in text documents, and attempts to convert them to machine-readable SMILES format. Having created a database of SMILES found in abstracts, our workflow converts these SMILES to 3D (with OpenEye OMEGA) and docks them into a protein of interest, then allowing them to be visualized in a Google-like interface. We are currently expanding this workflow to search a number of databases (including PubChem) as well as a wider range of abstracts and available full-text articles.

NEXT STEPS

We aim to develop web service interfaces to as much chemoinformatics functionality as we can, and so we are continuing to create more services, exposing algorithms and software developed at Indiana University and in the wider chemoinformatics community. To this end, we are actively participating in open source chemoinformatics software movements such as Blue Oblelisk and the Chemistry Software Development. We are also committed to making all of our services publicly available except where we are bound by commercial licenses.

To aid in the development of workflows that are meaningful and useful to the scientific community, we are engaging chemists and other life scientists from a variety of disciplines to identify specific life science information problems that are amenable to be solved using our web service infrastructure, and to evaluate the scientific usefulness of the output of the workflows. We hope to be able to demonstrate that by combining methods in workflows that might not previously have been used together (such as literature searching and docking), we can provide supplemental information that is useful in guiding drug discovery and chemistry projects.

We have also begun work on interaction-layer tools that allow scientists to interact with workflows and data. In particular, we are interested in "active computation" that employs workflows to generate information that is considered to be useful to scientists without necessarily direct input from scientists. We are currently developing a variety of tools and methodologies to enable scientists to access and interact with this information.

ACKNOWLEDGEMENTS

This work was financially supported by the NIH through their Exploratory Centers for Cheminformatics Research funding and through a Microsoft Smart Clients for eScience grant. We would like to thank OpenEye software and Digital Chemistry for allowing us to use their programs. We would also like to thank Gary Wiggins at Indiana University, and Tom Doman, Mic Lajiness, and Dan Robertson at Eli Lilly in Indianapolis for helpful input into this work.

REFERENCES

1. Mullin, R., Dealing with Data Overload. *Chemical & Engineering News* **2004,** 82, (12), 19-24.
2. Gardner, S. P.; Flores, T. P., Integrating information technology with pharmaceutical discovery and development. *Trends in Biotechnology* **1999,** 18, (Supplement 1), 2-5.
3. Berners-Lee, T.; Hendler, J.; Lassila, O., The Semantic Web. *Scientific American* **2001,** 284, (5), 34-43.
4. Hendler, J.; Berners-Lee, T.; Miller, E., Integrating Applications on the Semantic Web. *Journal of the Institute of Electrical Engineers of Japan* **2002,** 122, (10), 676-680.
5. http://www.w3.org/XML, accessed May 25, 2006

6. A good introduction to RSS can be found at http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html, accessed May 25, 2006.
7. Murray-Rust, P.; Rzepa, H. S., Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *Journal of Chemical Information and Computer Sciences* **1999,** 39, (6), 928-942.
8. Murray-Rust, P.; Rzepa, H. S., Chemical Markup, XML, and the Worldwide Web. 2. Information Objects and the CMLDOM. *Journal of Chemical Information and Computer Sciences* **2001,** 41, (5), 1113-1123.
9. 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. *Journal of Chemical Information and Computer Sciences* **2001,** 41, (5), 1124-1130.
10. Murray-Rust, P.; Rzepa, H. S., Chemical Markup, XML, and the Worldwide Web. 4. CML Schema. *Journal of Chemical Information and Computer Sciences* **2003,** 43, (3), 757-772.
11. Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L., Chemical Markup, XML, and the Worldwide Web. 5. Applications of Chemical Metadata in RSS Aggregators. *Journal of Chemical Information and Computer Sciences* **2004,** 44, (2), 462-469.
12. Holliday, G. L.; Murray-Rust, P.; Rzepa, H. S., Chemical Markup, XML, and the Worldwide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *Journal of Chemical Information and Modeling* **2006,** 46, (1), 145-157.
13. Curcin, V.; Ghanem, M.; Guo, Y., Web services in the life sciences. *Drug Discovery Today* **2005,** 10, (12), 865-871.
14. http://www.ebi.ac.uk, accessed May 18, 2006
15. http://www.iupac.org/inchi, accessed May 18, 2006
16. http://wwmm.ch.cam.ac.uk, accessed May 18, 2006
17. Klon, A. E.; Glick, M.; Davies, J. W., Combination of a Naive Bayes Classifier with Consensus Scoring Improves Enrichment of High Throughput Docking. *Journal of Medicinal Chemistry* **2004,** 2004, (47), 4356-4359.
18. Klon, A. E.; Glick, M.; Davies, J. W., Application of Machine Learning to Improve the Results of High-Throughput Docking against the HIV-1 Protease. *Journal of Chemical Information and Computer Sciences* **2004,** 44, (6), 2216-2224.
19. Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W., Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking. *Journal of Medicinal Chemistry* **2004,** 200, (47), 2743-2749.
20. Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D., Classification of Kinase Inhibitors using a Bayesian Model. *Journal of Medicinal Chemistry* **2004,** 47, 4463-4470.
21. Yoon, S.; Smellie, A.; Hartsough, D.; Filikov, A., Surrogate docking: structure-based virtual screening at high throughput speed. *Journal of Computer-Aided Molecular Design* **2005,** 19, (7), 483-497.

	Purpose	Technologies
Interaction Layer	Interactive software for creative access and exploitation of information by humans	Microsoft Smart Clients, portlets, Java applets, email and browser clients, visualization technologies. RDF descriptions, OWL ontologies
Aggregation Layer	Workflows and data schemas customized for particular domains, applications and users	BPEL, Taverna and other workflow modeling tools, aggregate web services, RDF descriptions, OWL ontologies
Web service layer	Comprehensive data and computation provision including storage, calculation, semantics and meta-data exposed as web services	Apache web services, SOAP wrappers, WSDL, UDDI, XML, Microsoft .NET

Figure 1. A three-layer model for a web service infrastructure applied to chemoinformatics

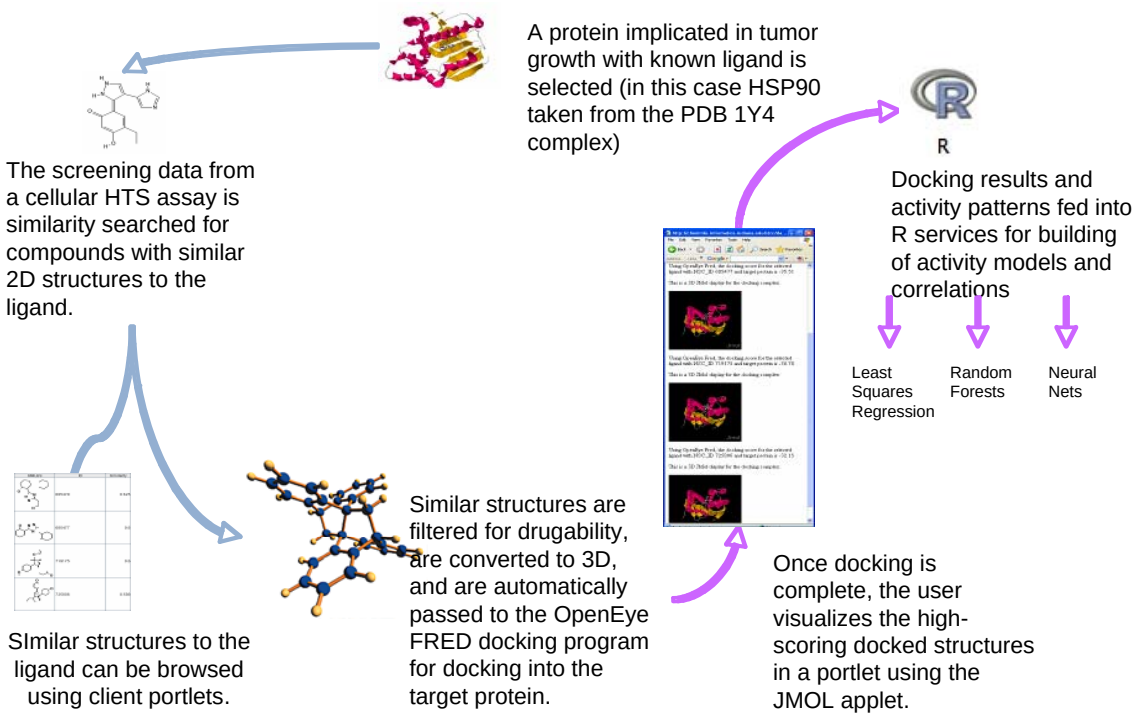

Figure 2. An example workflow to correlate docking results with cellular assay results