Advances in Cheminformatics Methodologies and Infrastructure to Support the Data Mining of Large, Heterogeneous Chemical Datasets

Rajarshi Guha[§], Kevin Gilbert[§], Geoffrey Fox*[§], Marlon Pierce*, David Wild[§], Huapeng Yuan*
[§]School of Informatics, Indiana University, Bloomington, IN 47408.
*Community Grids Labs, Indiana University, Bloomington, IN 47408

Abstract

In recent years, there has been an explosion in the availability of publicly accessible chemical information, including chemical structures of small molecules, structure-derived properties and associated biological activities in a variety of assays. These data sources present us with a significant opportunity to develop and apply computational tools to extract and understand the underlying structure-activity relationships. Furthermore, by integrating chemical data sources with biological information (protein structure, gene expression and so on), we can attempt to build up a holistic view of the effects of small molecules in biological systems. Equally important is the ability for non-experts to access and utilize state of the art cheminformatics method and models. In this review we present recent developments in cheminformatics methodologies and infrastructure that provide a robust, distributed approach to mining large and complex chemical datasets. In the area of methodology development, we highlight recent work on characterizing structure-activity landscapes, QSAR model domain applicability and the use of chemical similarity in text mining. In the area of infrastructure, we discuss a distributed web services framework that allows easy deployment and uniform access to computational (statistics, cheminformatics and computational chemistry) methods, data and models. We also discuss the development of PubChem derived databases and highlight techniques that allow us to scale the infrastructure to extremely large compound collections, by use of distributed processing on Grids. Given that the above work is applicable to arbitrary types of cheminformatics problems, we also present some case studies related to virtual screening for anti-malarials and predictions of anti-cancer activity.

1. Introduction

The field of cheminformatics is multi-disciplinary and is an amalgam of chemistry, mathematics and computer science. Cheminformatics focuses on the development of methods and tools to address problems in the management and analysis of chemical information. Such information can be of a variety of types including chemical structure (in various formats such as SMILES, SDF, CML and so on) and derived aspects of chemical structure (such as number of atoms and various descriptors of structure). With the advent of large public repositories of chemical and biological data, there has been an explosion in the type and amount of data that is now available for cheminformatics analysis. Thus, we can now access chemical structures of millions of compounds, as well as the biological activities of many of these structures in a variety of biological assays. Public literature databases now allow one to move from chemical structure to specific documents and in some cases vice versa.

It is clear that there is a need to be able to connect chemical structure with biological information (which may include structures of biologically relevant macromolecules and biological activities). Given the rise of chemical biology and chemogenomics [Bredel2004], we must also look beyond information organization and handling issues and develop efficient tools and interfaces that will allow us to analyze chemical and biological information in an integrated fashion and obtain robust predictions of the biological effects of small molecules. This aspect is very important since it is now easy to collect information from a variety of high throughput experiments and vendor catalogues. As a result, though access to the data in a variety of modes (such as SQL queries, web service based queries and so on) is possible, it is important to consider that analysis tools and pipelines should also be able to access the data in a uniform and

standardized manner and in the longer term, without human intervention.

These requirements are not unique to cheminformatics. Indeed computer scientists and software engineers have developed a wide array of systems that support the interlinking of geographically distributed computational resources. The idea of the modern Grid computing was first described by.[Foster1999]. This architecture aims to connect geographically distributed individual computers, clusters of computers, and more specialized parallel computing and mass storage resources, providing services and client libraries for remote job execution, distributed file management, resource information management, and integrated security. This architecture has merged with the more general Web service architecture. Generally all such systems are examples of Service Oriented Architecture, in which network accessible services with well-defined access (or programming) interfaces communicate with client applications using XML-encoded messages transmitted over HTTP. Typically the service interfaces are defined with the Web Service Description Language (WSDL) and messages are wrapped in SOAP. Alternative versions of this architecticture have also become popular. The Representation State Transfer (REST) style of Web services replaces WSDL with the basic HTTP operations GET, POST, PUT, and DELETE, which operate on URLs. The actual content of the URLs may be SOAP messages but also RSS and Atom XML formats are popular. Non-XML formats like JavaScript Object Notation (JSON) can also be used for messages. In any case, the general architecture approaches are similar, and the distinctions do not concern us here. The advantages to such an architecture are obvious: one is no longer restricted to a local machine and submitted jobs can be moved to the most appropriate machine (in terms of hardware configuration or availability). The NSF-funded t TeraGrid and the NSF and DOE-funded Open Science Grid are the two most prominent examples of Web service-based Grids for scientific computing within the United States.

There are a number of important aspects of such distributed infrastructures. First, the broad goal is to enable collaboration between different groups in an easy and efficient manner. Such collaborations will involve communication between people as well as machines, and will exchange a wide variety of data types (documents, chemical structures, assay readouts and so on). Second, to achieve efficiency, standards for interoperability are critical. Such standards apply to various parts of such an infrastructure including communication standards, data storage and annotation standards as well as authentication standards. Third, given the compute intensive nature of many chemical problems, easy access to powerful machines such as supercomputers and computational grids becomes important. Such infrastructures need to adhere to standards developed in the Web and Grid computing communities if they are to make use of such resources. In the area of cheminformatics, Guha et al [REF] have described the “Blue Obelisk”, an umbrella organization of various open source projects, whose aim is to define and promulgate community-approved standard for various aspects of cheminformatics. There are many different features of distributed infrastructures and while a number of efforts in the chemical sciences community will be discussed in this paper, it should be noted that, all the projects do not aim to address all the features noted here. Indeed, this is the utility of such infrastructures – as long as they follow standards, they should be able to interoperate. We also note that a variety of other fields have developed distributed infrastructures and example include geography, biology and high energy physics and we do not discuss them in this paper and rather, focus on the chemical sciences.

It should be noted that a distributed infrastructure is not simply about linking computational resources. Indeed, the most common feature of such an infrastructure is the ability to connect data sources in a seamless fashion. This is especially important given the fact that we are faced with a deluge of data from a variety of sources. Clearly, no one person or organization is in a position to effectively analyze it all. As a result, such data sources must be made easily accessible. But it is not sufficient to allow easy connections to the sources. Flexible approaches must be provided to allow queries and combine results from different data sources. Though it is certainly true that simply having large amounts of data does not lead to insight, it is also true that no insight will be derived from such data collections if they cannot be analyzed in a flexible and efficient manner. This observation leads on to the fact that once data is accessible, analytic methods must be applied. The cheminformatics literature is replete with methods for various

tasks. Yet, analysis of real world data demands that robust and efficient *implementations* be available. Given the heterogeneity of today's computing infrastructure, it is natural to expect that analysis tools and methods will be accessible in a variety of ways. Traditionally, command line tools have been used by scientists, giving way to GUI's. However, both are generally restricted to local functionality, though many of today's tools can make use of resources on the Internet. As will be described later on, making such tools available as distributed services leads to a much more flexible analysis infrastructure – one that can be accessed from anywhere and that can be incorporated into any thing, be it a command line, GUI or workflow tool. Adopting this open service architecture is vital if we want academic chemical informatics research to prosper, as it enable new tools and pipelines to be developed from existing parts. Figure 1 pictorially demonstrates the evolution of distributed infrastructures from both a hardware and software perspective.

2. Outline

The paper is arranged as follows. In Section 3 we discuss previous work that has developed distributed infrastructure for various aspects of chemistry including ab-initio computation, crystallography, virtual screening and cheminformatics. We also briefly discuss the use of various standards that make such infrastructures viable. In Section 4 we discuss the efforts at Indiana University to develop a cheminformatics cyber-infrastructure that addresses a wide range of topics in cheminformatics ranging from ore cheminformatics methods and predictive modeling methods to specific tools and data storage. In this section, we highlight how each aspect of the infrastructure is designed to provide flexibility in terms of access and usage, allowing one to develop novel applications in a rapid fashion. In Section 5 we present two case studies where we have utilized portions of our infrastructure to address real-world problems. Finally, Section 6 summarizes our research and discusses future work.

3. Previous Work

Though cheminformatics has been in existence since the 1960's, it is only recently that large collections of chemical information have highlighted the need for infrastructures capable of handling them. At the same time, the ability to connect computers in networks as well connect networks themselves provides us with vast computing power, that in many cases is also cheap. The use of compute clusters, which are usually a collection of machines located in a single geographical location, is common in many areas of high-performance computing such as computational chemistry and biology. More recently, the ability to seamlessly utilize geographically dispersed computers and networks has led to a number of large, *distributed*, projects such as Folding@Home [Pande2003]. In both these cases, the focus is on enhancing computational power, by virtue of parallelization. The combination of distributed computer systems, large collections of chemical and biological information, and novel computational methodologies presents us with a new way to enhance the usage and accessibility of chemical information. By virtue of well-known standards to represent data and communicate between programs, we now have the ability to "mix and match" methods and data in an unprecedented fashion.

Various subfields of chemistry have developed and employed distributed infrastructures for a variety of purposes. A number of these efforts focus on compute-intensive tasks such as *ab-initio* calculations whereas others represent infrastructures to store and disseminate chemical data (such as crystal structures). It is only recently that such infrastructures have been developed and deployed for problems in cheminformatics in general. In this section, we briefly survey past work that have employed distributed infrastructures in the chemical sciences. At this point, we should note that traditional definitions of cheminformatics can cover all the work described in this section. Thus, a cyberinfrastructure for *ab-initio* calculations could be described as a cheminformatics project. We choose to be a little narrower and consider individual infrastructures in terms of their application area. In this sense, a cheminformatics cyberinfrastructure would relate to tasks in the domain of cheminformatics such as predictive

modeling, diversity analysis, virtual library generation and so on. Such an infrastructure would be distinct from say one designed to support crystallographic information, though of course, the two could interoperate with each other.

We first consider a distributed infrastructure designed to store and share crystallographic information. Cole et al [REF, REF] described an infrastructure, termed the ECSES environment, to share high-throughput crystallographic information that connected users, spread across the Internet, to instrumentation. The infrastructure allowed users to remotely monitor crystallographic experiments and perform the relevant analysis. For example, a user could provide a sample to the crystallography service and then view images from a diffractometer and evaluate unit cell parameters in real time, communicate with the instrument operator and so on. On completion, the crystallographic data is stored in a local database. The environment then allows a user to search databases remotely for crystal structures that are similar to the one generated. The ECSES environment was constructed using Grid (specifically, Globus [REF]) and web service technologies that allowed secure two-communication between users and crystallographic facilities. By use of secure web services, the environment provided access to the Cambridge Structural Database for queries. The use of Grid technologies allowed the ECSES environment to support computationally intensive operations such as melting point calculations. A similar project that enables control of crystallographic instrumentation and distributed analysis of experimental data is the CIMA project [REF,REF]

Another distributed infrastructure for crystallography was described by Hunter et al [REF]. They developed AnnoCryst, a system that allowed users to collaboratively annotate crystal structures and share such annotations in a secure and asynchronous fashion. The system is designed such that annotations for a given crystal structure are stored in a secure fashion and can easily be linked to actual crystal records, which may be stored in arbitrary databases. An important focus of the system is the ability to allow *groups* of scientists to collaborate on annotations. Given that annotations are stored, security played an important role, to allow authentication of such annotations. The infrastructure was developed using a variety of open source components and made extensive use of published standards. For example, they employed Shibboleth [REF] to enable identity management and XACML to define and enforce access control policies. They also made extensive use of social networking standards such as FOAF to describe relationships between contributors and semantic web tools (such as Annotea for the storage of annotations) and standards such as RDF. The latter aspect is important since it allowed them to link to other semantic resources such as Connotea and the Protein Ontology as well as allow sophisticated semantic searches (using SPARQL) over the annotation collection. This project highlights that distributed infrastructures need not be restricted to computational services or data storage. Indeed, the AnnoCryst project combines data storage with collaborative interaction, allowing multiple distributed users to enhance entries in a wide range of crystal structure repositories.

Paolini and Bhattacharjee [REF] described the development of a web service infrastructure to access thermochemical data and calculations. They aggregated a variety of thermochemical data sources, such as NIST, NASA and CHEMKIN [REF], into a relational database. In addition to storing reported data, they also performed a number of calculations to evaluate properties such as entropy and hat capacities. Access to the database is provided via SOAP based web services. They also highlighted the flexibility of these web services by incorporating them into Matlab and Microsoft Excel to generate formatted thermochemical data tables.

The field of computational chemistry has seen a number of grid computing initiatives. An early example is that of the Computational Chemistry Grid (CCG) [REF]. This project was a three-tiered system consisting of a client side GUI, a middleware service and a resource layer. The middleware layer was based on Globus [REF] and its role was to distribute jobs across the Grid, based on a variety of parameters such as allocations and availability. The primary use of this grid was to run computational chemistry calculations such as Gaussian and NWChem. In a similar fashion, Truong et al [REF] developed a cyberinfrastructure for scientific computing termed CSE-Online. This project consisted of a centralized web portal that was the point of entry for users. Via this portal, a user would be able to access individual programs, upload datasets and extract results form prior runs. The actual computational services were provided

by multiple, possible remote, servers hosting individual applications. The infrastructure also provided resource management tools such as job scheduling via PBS and job submission to the TeraGrid [REF] Communication between the portal and individual compute servers is performed via web services. In addition to the online portal, the project also developed a desktop interface to the portal which provided a variety of capabilities such as relational database for storage, visualization of results, generation of input scripts and so on. In terms of functionality, the infrastructure was oriented towards ab-initio calculations and molecular simulation. Sanna et al [REF] described the development of a Grid infrastructure for quantum chemistry computations using Gaussian [REF]. Their infrastructure was based on a three-tier architecture comprised of an *interface tier* (responsible for communications between the grid infrastructure and user clients), the *grid tier* (responsible for management of grid resources and scheduling) and finally the *resource tier* (responsible for the actual computations) and has been described by De Meo [REF]. It is interesting to note that the quantum computing grid described by Sanna was adapted from a Grid infrastructure developed for a bioinformatics application – highlighting the generality of this approach.

Finally, we consider efforts to develop distributed infrastructure specifically in the area of cheminformatics. An early effort at a distributed infrastructure for chemical information is the World Wide Molecular Matrix (WWMM) developed by Murray-Rust and co-workers at Cambridge University. The focus of this project was the development of web services that would provide access to data sources (usually some form of relational database) and methods (such as format conversions). As part of this project, they also investigated the use of XML databases (specifically XIndice) for efficient storage and querying of CML [Murray-Rust1999] documents.

In 2005, the NIH funded several groups (including Indiana University) under the Molecular Libraries Initiative (MLI) [Austin2004]. These groups were termed Exploratory Centers for Cheminformatics Research (ECCR) and were tasked with extending the state of the art in cheminformatics research. Though much of the research focused on methodology development, a significant aspect of their work was in providing access to tools, methods and data. For example, the University of North Carolina have developed *CarolineChemBench* (http://ceccr.unc.edu/) which is a web interface to a wide variety of tools, primarily focusing on QSAR modeling. As the name suggests, the application aims to be a web-based workbench for QSAR modeling. It provides access to a number of molecular descriptors and machine learning methods. A similar application was developed at the Renssalear Polytechnic Institute, which focused on a descriptor development and machine learning methods for QSAR modeling (http://reccr.chem.rpi.edu/). As with the UNC efforts, their tools were exposed in the form of a web application that allowed users to upload data and develop predictive models. Characteristic of both these efforts is that access to their tools and methods is strictly via a web application. In other words, a user must navigate to their web page and use the functionality provided by the developers. For many cases, this is sufficient. However, it is can be an inflexible approach at times and does not allow one to “mix and match” methods. North Caroline State University was also designated an ECCR and their focus was on machine learning methods and validation methods for statistical models. They developed a web application in which users could upload data and automatically develop a series of predictive models. As part of their work, they developed a tool called PowerMV [REF] which is a standalone tool to examine SD files and generate a wide variety of molecular descriptors. This tool was employed by their web application, in the backend, to provide input to the machine learning algorithms.

Apart from the NIH funded groups mentioned above, a number of other workers have developed infrastructure for various cheminformatics tasks. For example, Sild et al [Sild2006] have described a Grid infrastructure termed OpenMolGRID which connects geographically distributed computational resources to provide seamless access to data and applications. Their infrastructure was comprised of three main components: a data warehouse, data mining module and molecular engineering module. These high level modules are based on the UNICORE [Romberg2002] Grid middleware. More specifically, they incorporated tools such as CODESSA Pro and MOPAC, which were linked into workflows using MetaPlugin [Sild2005] allowing users to automate the QSAR modeling process across the Grid.

Karthikeyan et al [REF] described a distributed infrastructure, termed ChemSTAR based on Java Remote Method Invocation (RMI) to distribute molecular descriptor calculations over multiple machines. They employed this infrastructure to rapidly evaluate molecular descriptors for 11 million molecules. This infrastructure differs from OpenMolGRID in that it does not use a standard middleware such as Globus or UNICORE. In this sense, the use and deployment of ChemSTAR is not as seamless as some other infrastructures. It should also be noted that other projects such as VCCLAB [REF] have also employed a Java RMI approach to connect geographically distributed resources to provide access to various cheminformatics and data mining tools via a single point of entry. As with OpenMolGrid, it employs a three-tier architecture in which users employ a local client (written as a Java applet), which communicates with a "SuperServer" acting as the middleware between the client applet and the actual compute servers.

Though the Semantic Web [Goble2006] is not a distributed infrastructure per se, developments in this area are becoming key to efficient, automated and seamless use of distributed infrastructures. A number of efforts have been reported on the incorporation of semantic technologies into chemical information sources. A number of technologies have been developed to represent, search and manipulate semantic information including RDF, SPARQL, and FOAF. A number of workers have co-opted these technologies to develop novel applications in cheminformatics. This is not surprising since chemical information is extremely diverse (ranging from structural information, patents and journal articles to drug targets, side effects and pathway information) and traditional methods of storage and manipulation do not always make the links between individual concepts apparent.

Much of the work in this area, as applied to chemistry builds on the seminal work by Murray-Rust and co-workers in developing an XML format to represent chemical data, termed Chemical Markup Language (CML) [REF]. An early application of CML to represent the semantics of chemical information was CMLRSS [REF], which was used to enhance RSS feeds with chemical data. Casher et al [REF] described an ontology, termed SemanticEye, to enhance the extractability of chemical information from various documents (primarily PDF) and store such information in an RDF repository. Some examples of the information they extracted included document metadata (encoded as Adobe XMP within a PDF), Digital Object Identifiers (DOI) and structures represented as InChI's. They also developed a browser-based interface allowing users to explore the metadata repository and examine individual documents or even aggregate documents based on their semantic relationships (such as similarity based on the InChI's contained within the documents). Taylor et al [REF] also described the use of RDF to represent and store chemical information. An important aspect of their presentation was the ability to incorporate provenance information into the repositories. They also examined the use of triple-stores to store chemical information and their behavior with respect to size of the collection. They performed benchmarks such as chained RDF queries and insertion of RDF files into the triple-store and showed that they could obtain reasonable performance with up to 36 million triples (representing the unique molecules from the ZINC [REF] database).

Another project in the area of semantic chemistry being carried out by the Royal Society of Chemistry and is termed Project Prospect. The goal of this project was to markup journal articles in the field of chemistry such that one could link articles to other source of information in a semantic manner (as opposed to direct URL's). Thus, they employed the Gene Ontology [REF] and Gold Book [REF] to identify terms within the documents, highlighting them on the fly. In addition, they were able to extract chemical entities using OSCAR3 [Corbett2006] and display chemical structures for the entities that were successfully extracted. Though restricted to chemistry journal articles, this project highlights the utility of semantic markup, since searches over a document collection can employ not only exact or fuzzy textual matches, but also perform semantic queries that make use of the extracted metadata.

4. A Cyberinfrastructure for Cheminformatics

In this section, we present the cheminformatics cyberinfrastructure developed by the Cyber Infrastructure Cheminformatics Collaboratory (CICC) at Indiana University. This project was funded under the NIH MLI initiative and was tasked with developing a distributed infrastructure that would incorporate a wide variety of cheminformatics data sources and methodologies. The overarching goal of the project was to provide multiple, flexible approaches to access cheminformatics methods and data. In this section we discuss various aspects of this project including methodology development and cheminformatics web services.

4.1. Novel cheminformatics methodologies

Key to a cyberinfrastructure for cheminformatics is the development of novel cheminformatics methodologies. In this context, we have been developing methods that address a number of areas in cheminformatics ranging from characterization of chemical spaces and quantitative structure-activity relationships to analysis of chemical documents.

Chemical spaces are defined for a set of molecules in terms of a collection of molecular descriptors. Given that one can generate thousands of descriptors [REF], chemical spaces can have high dimensionality. Given a suitably selected chemical space, a common task is to characterize the distribution of the molecules in the chemical space. This is formally termed diversity analysis [Martin2001] and a number of methods have been described in the literature [REFS]. We recently described a method termed RNN curves that allows one to characterize the spatial distribution of molecules in arbitrary chemical spaces by measuring the density of the chemical space around a given molecule. An immediate application of the method is to identify outliers in a dataset [REF]. The method has also been extended to the problem of *a priori* evaluation of the number of clusters in a data set. Traditionally, one has had to perform multiple clusterings to identify the "best" number of clusters. In this context, "best" is determine using one of the many cluster quality measures such as the silhouette width [Kaufman1990], Dunns index [Dunn1974] and so on. The RNN-curve method allows one to analyze the dataset to identify the number of clusters, without having to perform the clustering itself. In many ways this method is similar to the “gap statistic” described by Tibshirani et al [Tibshirani2001], but is not based on any distributional assumptions. Though the method does have some drawbacks related to specific spatial distributions, experiments on real and synthetic datasets [REF] indicate that it is able to correctly identify the number of clusters in a dataset. The observations were also confirmed by comparing the results with those obtained using the silhouette width based on multiple clusterings.

We have also applied the deterministic annealing algorithm [Rose1990, Rose1998] to the problem of clustering large collections of chemical compounds. The specific method employed combined deterministic annealing and Generative Topographic Mapping (GTM) [Bishop1998]. In comparison to simulated annealing, deterministic annealing does not involve any Monte Carlo steps and rather, optimizes the energy function iteratively, as the annealing temperature is reduced. This procedure compares well with the traditional k-means clustering algorithm. Given that the clustering is usually performed in a high dimensional space, the use of GTM allows us to map this space to a lower dimensional form (usually 2D) such that the results of the clustering can be easily visualized. The motivation for this work was to investigate to what extent the algorithm could be parallelized on multi-core processors. We performed experiments [Qiu2007] on a 4-core and a 8-core compute cluster and considered a collection of 40,000 compounds taken from PubChem, characterized using 1052 bit BCI keys (Digital Chemistry, UK). Our experiments indicated that, depending on the number of clusters specified, we could achieve a speed of 7X on an 8-core system. While this approach requires that one specify the number of clusters *a priori*, the results indicate that clustering large compound collections can make good use of modern multi-core systems.

In the area of quantitative structure-activity relationships we have focused on both applications and methodologies. For example, we developed QSAR models to predict the anti-cancer

activities of compounds tested against the 60 NCI cancer cell lines by the Developmental Therapeutics Program. The random forest was employed to predict the anti-cancer activity of a compound based on MACCS keys as structural descriptors. The performance of the models ranged from 75% to 80% correct over the 60 cell lines. We also employed random forest models to predict the ability of compounds to inhibit cell proliferation in a variety of human cell lines. These cell lines have been used in high-throughput screening (HTS) assays [Xia2008], which tested approximately 1300 compounds. This data was used to train the individual random forest models, which were then used to predict the cytotoxicity of an external set of compounds. The challenge in this study was the fact that the assay results for a number of the cell lines were severely unbalanced. We investigated several methods to alleviate this problem and developed a method termed "bit spectra" to allow us to compare a new dataset with the training set to determine whether the model will lead to reliable predictions for the new dataset [Guha2007].

We have also performed methodological investigations into a number of fundamental aspects of QSAR modeling. We developed a method termed "ensemble feature selection" [REF] whose goal is to identify a suitable set of descriptor such that it optimizes the predictive performance of multiple, disparate models. Traditionally, one uses a feature selection algorithm (such as a genetic algorithm) to identify a suitable set of descriptors that optimizes the performance of single model. In general, this set of descriptors will not be optimal for a different type of model. As a result, most studies have used different set of descriptors for different models, leading to the possibility that the individual models encode different aspects of the structure-activity relationship (SAR). The ensemble feature selection employs a genetic algorithm whose objective function is function of the performance of multiple models (say, linear regression and neural network models). The result of the procedure is to identify a single set of descriptors that maximizes the predictive performance of the individual models. Clearly, the models are not as good as they would have been if they had been optimized separately. Yet, our results indicate that the loss in performance is minimal.

More recently, we have described a novel approach to characterizing "activity cliffs" in a dataset [REF]. Activity cliffs are defined as pairs of compounds that are structurally very similar, yet exhibit large differences in activity. Such cliffs represent problems for traditional machine learning based QSAR methods. At the same time, these cliffs represent very important aspects of the structure-activity relationships in a dataset. As a result, it is useful to be able to explore the cliffs in a dataset. We described the Structure Activity Landscape Index (SALI) [REF] that allows one to numerically characterize activity cliffs of varying degrees. The approach generates a symmetric matrix termed the SALI matrix, which can be used to easily summarize the nature of the “landscape” represented within the dataset. However, the matrix is a rather broad view and so we developed a novel network visualization of the activity cliffs in a dataset. This view is derived from the SALI matrix, in which we consider each compound a node and two nodes are connected if they have a SALI value greater than a user defined cutoff. An example of such a SALI network is shown in Figure 2. The key feature is that it highlights the most significant activity cliffs and allows one to intuitively explore the changes that lead to increased activity. We have also employed the SALI to measure the ability of structure-activity relationship models to encode the SAR's present in a dataset [REF]. This is based on the observation that since activity cliffs represent the most “interesting” parts of an SAR, a model encoding the SAR should be able to capture these cliffs. In other words, a model that identifies more cliffs than another model could be considered the “better” model. In this context, a model “identifies” a cliff by predicting the ordering of the nodes in a SALI graph. We have shown that the method is very general in that it can be applied to any type of model that aims to encode an SAR such as QSAR docking, CoMFA and pharmacophore models. Retrospective experiments indicate that it is indeed able to differentiate between statistically similar models by considering the SAR’s encoded by the model, as opposed to pure numerical metrics.

4.2. Cheminformatics Tools

Though a large component of cheminformatics research revolves around the development of

efficient algorithms to analyze large and diverse chemical datasets, it is vital that such research not be restricted to theoretical analyzes. In other words, implementation of cheminformatics research is arguably as important as algorithm development. To this end, we have developed a number of tools implement many of the methodologies described in Section 4.1 and allow users to access both methods and data in a variety of environments. In this section, we provide a brief overview of the various end-user tools that have been developed as part of the CICC at Indiana University.

With the growing importance of the Internet as a means of information presentation and exchange, the number of Web-accessible chemical information sources has increased significantly. Examples include online journals, chemical databases [Williams2008], blogs and Wiki's. By definition, most such sites focus on their core competencies. Thus, online journals will not maintain large chemical structure databases. Similarly blogs that talk about chemistry will usually link to other resources. It is clear, however, that aggregation of the various different types of information can enhance the experience of a user visiting such sites. For example, if a user visits the Table of Contents page of an online journal, it would be useful to let the user know of any web pages that have commented on a specific article. Similarly, when performing a structure search on PubChem, it can be useful to display a 3D structure of a molecule. However, PubChem does not provide such structures, so these must be obtained from somewhere else.

The above scenarios, which focus on aggregating information from multiple sources, can be achieved using Userscript technology. Essentially, this is a browser technology (specifically for Mozilla based browsers) that allows a user to run a Javascript program on a webpage during load time. The script itself is free to access any resource on the Internet. However, the key feature of these scripts is that they can *rewrite* the webpage that the user is viewing in a transparent manner. That is, the user sees the rewritten page, rather than original page. We have recently described the application of this technology to cheminformatics applications [REF]. One of the userscripts we reported allows one to view 3D structures for molecules obtained via searches on the PubChem website. Thus, if one searches for the term "aspirin" and then clicks on the link for CID 2244, one is taken to a page summarizing the information for that molecule. With the userscript installed, the page is rewritten to include two new links. The first link brings up a Jmol window displaying the 3D structure obtained from the Pub3D database at Indiana University. The second link allows one to download the 3D structure in SD format. Note that this procedure is entirely transparent to the user and no manual interaction is required. A second example is a userscript that operates on Table of Contents pages for ACS journals. A number of blogs are now available that refer to various journal articles using their Digital Object Identifier (DOI). The blog posts are aggregated at http://cb.openmolecules.net/, a resource which can be queried by DOI to identify which blogs refer to a given journal article. If the userscript is installed, then when the user navigates to a Table of Contents web page, the page is modified to provide new links in the article summary that refer the user to those blog posts. A modification to this script allows the user to click on the link, which will cause a "balloon" to popup containing the commentary in question. An example is shown in Figure 5. Note that this procedure requires no intervention on the part of the user. More importantly, it requires no support from the journal website.

However, the browser is not the only environment in which users access chemical information. A number of cheminformatics tasks are based on mathematical and statistical modeling techniques (such as QSAR). A common environment for this is the R package [REF]. This is primarily a statistical programming environment, and thus does not directly support cheminformatics. We have described an add-on to R called *rcdk* [REF], which interfaces the CDK [REF] (a Java library for cheminformatics) to R. As a result, one can directly load chemical structures in the R environment and perform various cheminformatics operations such as similarity calculations, fingerprint generation and evaluation of molecular descriptors. A related package called *rpubchem* [REF] allows direct access to the PubChem database from within R. Thus, one can perform keyword searches and down structure and property information from both the PubChem compound and bioassay collections. Together with the *rcdk* package, it allows the user to directly access and manipulate chemical information within the R

environment.

In addition to the browser based tools and R packages, we have developed a number of standalone tools for a variety of tasks. Though the standalone tools by themselves do not represent a distributed infrastructure, they provide core functionality, allowing groups outside of Indiana University to replicate our infrastructure. Many of these are based on the CDK library and examples include a GUI for calculation of molecular descriptors and fingerprints and a command line tool that allows one to search 3D structure collections using pharmacophore queries [REF]. In addition, the research on structure-activity landscapes described in Section 4.1, has been implemented in a program which performs the SALI analysis and allows one to interactively explore SALI graphs. We also recently developed a program to generate 3D structures from SMILES strings, termed smi23d (http://www.chembiogrid.org/cheminfo/smi23d/), that is able to convert a SMILES string to a 3D structure in SD format. The program uses a linear embedding scheme to generate an initial set of "rough" 3D coordinates, which are then optimized using the MMFF94 force field [REF]. The use of this force field implies that compounds containing atoms that have not been parameterized in MMFF94 cannot be handled. This is contrast to programs like Corina [REF] that use a fragment-based approach and thus are able to handle a wider range of compounds. It should also be noted that smi23d generates a single conformer – which may not necessarily be the lowest energy (or even low energy). This program was used to construct the Pub3D database, discussed in Section 4.4. All the tools discussed here have been publicly released under Open Source licenses allowing them to be freely reused and incorporated into a variety of applications.

4.3. Distributed access to data and methods

Given that many groups, distributed across the world, have developed a wide variety of cheminformatics tools and algorithms, an important issue that must be addressed is the *deployment* of such tools in a manner that makes them accessible to experts and non-experts. Traditional approaches to tool deployment have been to simply make binaries (or provide the actual source code) available on a web site and have potential users download them. This approach assumes a level of software management and engineering expertise on the part of the user. Furthermore, one may also have to investigate the possible dependencies and obtain them before the actual cheminformatics tool can be used. This implies that the downloaded codes must be maintained. This approach is certainly useful in some instances, but is not an elegant solution for users who are not experts in the field of cheminformatics and simply want to use the tool. Moving away from locally running software is to provide web interfaces to tools. Wide varieties of cheminformatics methods are available in this form (such as VCCLab [Tetko2005], CECCR and RECCR). Though such a mode of deployment is useful for the non-expert user, the flexibility of such an approach is limited to that defined by the developer. In addition, web page front ends are not always amenable to programmatic access - each tools' web page being slightly different from the others.

Our approach to the problem of deployment has been to focus on the use of web services, which allow us to provide a consistent, distributed, programmatic interface to arbitrary methods, tools and data-sources.

Specifically, we have created web service interfaces for several cheminformatics methods, standalone tools and databases. Though these services are hosted at Indiana University, they can be hosted anywhere on the Internet. Each such service is accessible in the form of function calls. As a result, one is able to combine multiple services in a single program. Clients can be written in virtually any language that supports the Simple Object Access Protocol (SOAP) [REF]. As a result, one can generate the traditional web page interfaces to these services, which communicate with them via SOAP calls. However, if alternative interfaces (such as GUI's, command line programs or even inclusion into workflow tools such as Pipeline Pilot and Knime) are desired, one can easily combine multiple services. Table 1 summarizes the various services that are currently provided by the CICC and we provide a brief description of the services below.

We classify the services into four categories. First, we have a set of core cheminformatics services. These services represent fundamental cheminformatics operations such as fingerprint generation, similarity calculations, molecular descriptors and so on. Clearly, the bulk of these operations will not be used in a standalone manner. Rather, we expect that these will be used in combination with other services. An example might be an application that performs database queries and would like to display 2D depictions of chemical structures. Rather than integrate a standalone cheminformatics library in the application, the developer can simply make a function call to the web service and obtain a PNG of the 2D depiction.

The second class of services, represent statistical services. We specifically developed these services since predictive modeling is a common cheminformatics application. The underlying statistics engine is based on R. Individual R features such as sampling methods, model building (linear regression, neural networks and so on) and feature selection are presented as SOAP web services. The architecture of these web services has been described in detail previously [Dong2007]. The R web service infrastructure goes beyond simply exposing specific R methods as web services. Given the importance of predictive models in cheminformatics, we provide an infrastructure that allows one to deposit predictive models built in R (stored in the R binary format) and then *deploy* the model via the web service infrastructure. In this context, “deploy” implies that the model is publicly available and can be used to obtained predictions for new molecules via the web service. The infrastructure also allows one to retrieve the models, so they can be used locally if desired. An example of models deployed in this manner are a set of random forest models developed to predict anti-cancer activity on the NCI 60 cancer cell line dataset [Wang2007].

The third class of web services provides access to data sources. For example, we have created a number of databases that can be accessed via traditional SQL queries. Examples of these databases include mirrors of the PubChem compound, substance, synonym and bioassay tables as well as derivatives of PubChem, such as a 3D structure database (see Section 4.4). Traditionally, one would access such databases via SQL queries or web pages that execute said SQL queries. In contrast, we have wrapped a number of common queries, and presented them as SOAP based web services. As a result, clients can access these database using simple function calls, thus avoiding the construction of possibly complex SQL queries. At this point we don not provide a web service method that allows arbitrary SQL queries due to security issues. Even with this restriction, such an approach is still useful, since now the database can be accessed from arbitrary environments that support SOAP calls. As a result, there is no need for clients to construct SQL or scrape web pages to extract the results of queries.

The final class of services represent full fledged applications that have been wrapped as SOAP web services. This approach can be useful when certain applications have not been designed as a library. As a result, a web service essentially runs them as standalone programs and returns the output (possibly reformatted) to the user. From the users point of view, the command line details are hidden and only the functionality provided by the application is visible. The result of this approach is that arbitrary programs written in any language can be wrapped using Java web services and made to appear as simple function calls. As noted in Section 3, a number of workers have deployed computational chemistry programs in this manner. As part of our cyberinfrastructure, we have developed web service interfaces to a number of cheminformatics applications including ToxTree (http://ambit.acad.bg/toxTree) and a program design to evaluate pharmacokinetic parameters [Zhang2006]. One aspect of wrapping application as web service is that certain applications can take a long time to compute a result. In general, this is not as common a case in cheminformatics as in some other areas such as *ab-initio* calculations. A number of protocols have been devised to address the issue of asynchronous web service methods. WS-Notification and WS-Eventing have been put forward by the Web service community to provide call-back and notification mechanisms for long-running services. WS-Notification is included in the Globus Web Service stack. More recently, the Amazon Web Services team has put forward the Simple Queue Service. This is standard WSDL-described Web Service that allows other applications to update and query the state of long-running applications. Finally, RSS and Atom feeds, although normally used to syndicate human-readable news items, can also be used to convey state-machine updates in machine-to-

machine communications. Microsoft uses an extension of this approach, called FeedSync, in its LiveMesh software.

The utility of a collection of web services providing a variety of cheminformatics functionality is key to developing novel applications that go beyond traditional command lines and GUI's. As described in Section 4.2, developing of user scripts to enhance and modify web page content is dependent on the ability to communicate with arbitrary web services for 2D depictions and structure databases. Similarly, we have been exploring the use of Ubiquity (https://wiki.mozilla.org/Labs/Ubiquity), which is an extension to Firefox allowing one to easily develop mashups. An example of such functionality is the ability to select a piece of text in a web page and display the SMILES and associated structure if it is a chemical compound. This ability hinges on the use of a web service interface to our local PubChem mirror that allows us to search for compound synonyms. With the increasing number of web services being made available in the life sciences, the ability to "mashup" functionality to provide novel representations of chemical information becomes not only easy but, we expect, commonplace.

4.4. Adding Value to PubChem

The PubChem project is a component of the Molecular Libraries Initiative [Austin2004] and was designed to be a central repository of chemical structure and biological assay information. As of August 2008, it contained 2D structures (along with a number of calculated properties) for approximately 17 million compounds and 40 million substances. In addition to chemical structures, the database also stores synonyms for the compounds. Since one of the goals of the MLI was to identify chemical probes that could be used to investigate biological systems, the screening centers created by the MLI have performed 1157 assays and the results of these have also been deposited in PubChem. Clearly, the repository provides a wealth of information on the effects of chemical structure in biological systems and thus presents a ripe opportunity for chemical data mining.

While an extremely useful resource, one aspect that has not been addressed by the project is that of 3D chemical structures. A number of cheminformatics tasks such as CoMFA, docking and pharmacophore searches require that one generate 3D structures. Thus, if one were to employ compounds from PubChem, one would need to use some external software to generate such structures. Another advantage of 3D structures is the ability to perform searches based on shape similarity. For example, the PDBCal [Li2008] database reports thermodynamic properties for a number of receptor-ligand complexes. The database provides 3D structures and it would be useful to be able to search PubChem for molecules with shapes, similar to those stored in PDBCal. Though PubChem provides structure and substructure search capabilities, such 3D shape searching is not currently possible.

To address the lack of 3D structures, we have developed a derivative database of PubChem containing single conformers for 99.9% of the entire collection. The database is termed Pub3D and is available at http://www.chembiogrid.org/cheminfo/p3d. We used the smi23d program (Section 4.2) to generate the 3D structures for the PubChem collection. After generation of the 3D structures, we then created a Postgres database storing the structure files as text fields. We then evaluated shape descriptors using the CDK for each molecule, using the technique described by Ballester and Graham Richards [Ballester2007] and implemented in the CDK. These descriptors consist of 12 values that represent the first three moments of the distributions of distances between the atoms in a molecule and three four points. In contrast to more rigorous shape representation methods such as volume overlaps (implemented in ROCS, OpenEye Inc.), this method is cruder. However, since it represents a molecule as a 12-element vector, shape comparisons can be performed extremely rapidly. Specifically, the shape similarity between two molecules in this representation is given by

$$S_{i,j} = \left(\sqrt{\sum_{k=1}^{12} \left(X_{i,k} - X_{j,k} \right)^2} \right)^{-1}$$

where $X_{i,k}$ represents the k'th descriptor for the i'th molecule. Note that this definition of

similarity is different from that given by Ballester et al, where they employ the reciprocal of the normalized Manhattan distance between two shape descriptor vectors. For our purposes, the use of such a representation allowed us to store these vectors in the database and then index shape descriptors using an R-tree [Guttman1984] index, which is a spatial index that partitions a space and allows efficient nearest neighbor queries. Given that similarity queries can be formulated as nearest neighbor queries, this allows one to identify molecules with similar shapes in a rapid fashion. Figure 3 presents an example of the query performance, with a shape database of 10 million structures.

While our benchmarks indicate that shape searching in very large databases can be performed efficiently there are two significant drawbacks. Firstly, the database only stores a single conformer. In general, this limits the utility of the database. As a result, we are in the process of generating multiple conformers for the compounds in Pub3D, using in-house code for conformer generation. We expect that with multiple conformers, the database will be of general purpose utility. This however leads to the second problem. Assuming an energy window of 3.5 Kcal, our initial runs indicate that the average number of conformers per molecules is approximately 12. For the current collection of 17 million structures, this will create a database of more than 100 million structures. For such a large database, the size of the index becomes sufficiently large, that simply searching the index can take a long time. This is further exacerbated by the fact that in an ideal case the index would be stored in RAM. However, for such a large database, the amount of RAM (greater than 10GB based on current estimates) required to host it in memory, is not economical. As a result, much slower disk accesses will be required to search the index, negating the utility of the index.

To address this issue we have developed a distributed infrastructure that will allow scaling of the database to hundreds of millions of rows, yet achieve good performance. The infrastructure is shown schematically in Figure 4. We first performed a clustering of the database using the simulated annealing clustering method described in Section 4.1. The result of this procedure is to generate several clusters ranging in size from 200K to 1.5 million molecules. A key aspect of the clustering is that similar molecules are generally contained within the same cluster (though molecules on the edges of a cluster may be ambiguous). Each cluster of molecules is then hosted as a separate database, with it's own disk. We then implemented a broker that sits in front of the individual databases. The broker acts as an interface between user queries and actual query submission to the individual databases. The broker provides a web service interface, providing the advantages described in Section 4.3. When the broker receives an SQL query, it distributes the query to each of the individual databases asynchronously. It then collates the responses from each database and returns the result set to the user. The advantages of this infrastructure are obvious. As the original database grows in size, one need only add more clusters. Obviously, the clustering process is time consuming, but is a relatively infrequent process. Another bottleneck is the collation of results. If each cluster returns many thousands of results, then the collation procedure may become slow. However, for many practical cases, searches with such large result sets may not be very useful and hence the current approach appears to handle common scenarios quite efficiently.

Another area where we have attempted to add value to PubChem using a distributed infrastructure is in a service that allows one to identify possible promiscuous compounds within the PubChem bioassay collection. Promiscuous compounds are those that exhibit activity over a wide range of assays, due to factors not related to actual target binding [Feng2005, Feng2006]. Examples of phenomena that can lead to promiscuous behavior include aggregation [Feng2007] and redox cycling [Brisson2005]. Since such compounds are not actually active, it is useful to exclude them from further analysis. Though groups have reported the development of predictive models to identify promiscuous compounds [Doman2005, Roche2002], a simple way to highlight compounds that might be promiscuous is to count the number of times that they are reported as active in all the assays they are tested in. We developed a simple web application that utilizes a web service interface to a local mirror of the bioassay database and allows users to provide PubChem compound ID's and obtain a summary of the assays in which they have been noted as active. Clearly, this approach can be fooled – for example, if a compound is tested in few assays, it may be difficult to state that it is promiscuous. In such cases, the

"frequentist" approach can be enhanced with the use of heuristics. For example, one could employ a list of functional groups that are known to be reactive and hence could lead to promiscuous behavior.

A final aspect of adding value to PubChem is the use of RSS feeds. Traditionally, RSS feeds have been used by website to notify subscribers of new content or updates. It is natural step to apply this to databases, which are updated on a regular basis. To investigate the use of RSS feeds for PubChem, we constructed a mirror of PubChem at Indiana University. We then utilized a local web service interface to the mirrored database that allows one to execute a variety of queries (identify similar structures, retrieve structures by compound ID, substance ID or synonym and so on). The web service was used to generate an RSS feed for user-defined queries. For example, the user could search for the term "Viagra" at http://www.chembiogrid.org/cheminfo/rssint.html. The result of this search is an RSS document, rather than the usual web page. The RSS document is intended to be viewed in an RSS reader (such as Sage or Google Reader) and each entry of the feed is an individual hit from the database that matches the query. More importantly, the user can now bookmark the field and when viewed at a future time, will be updated with any new database entries that match the original query. Furthermore, we construct the feed using CMLRSS [Murray-Rust2004], which allows one to embed chemical information using CML [Murray-Rust1999] within the RSS feed. As a result the feed generated from the above website will contain 2D and 3D (if available) chemical structures of the compounds matching the query. These structures can be visualized in any CMLRSS capable reader such as Bioclipse [Spjuth2007]. Though we currently only provide an RSS feed for synonym based searches it is quite easy to extend it to arbitrary searches against the PubChem databases. Examples include searches for compounds satisfying certain property ranges (molecular weight, heavy atom count etc.) or new assays which tested a compound of interest.

5. Case Studies

In this section, we present two case studies where our infrastructure has played a role in real-world applications. These applications were chosen since they highlight the utility of our infrastructure in rapidly developing new applications by using a "mix and match" approach, without having to worry about specific tools and libraries.

5.1. Virtual screening for anti-malarial drug candidates

Malaria is a disease endemic to much of South America, South East Asia and Africa and is reported to affect nearly 300 to 500 million individuals every year [Snow2005]. Though anti-malarial drug research has been carried out for over 70 years, drug resistance has become a worrisome issue. There has been much focus on identifying novel targets in the genome of *P. falciparum*, as well as new compound classes, that exhibit increased potency and higher efficacy. We have been collaborating with Prof. Jean-Claude Bradley of Drexel University in the development of novel anti-malarial compounds. Prof. Bradley is a synthetic chemist and is focusing on the Ugi reaction to generate drug candidates. Since the Ugi reaction consists of four reagents, it is easy to show that one can synthesize millions of compounds depending on how many kinds of each reagent are available. As a result, we have focused on utilizing our computational infrastructure to generate virtual libraries and then prioritize compounds within such a library.

The first step in the procedure was to generate a virtual library, given a set of candidate reagants. The Ugi reaction requires the use of an amine, carboxylic acid, ketone and isonitrile and when the compound structures are provided in the form of SMILES, it is relatively easy to implement a procedure to generate all possible combinations of the reagents. We developed a web application that utilized the CDK based cheminformatics web services to validate structures and perform substructure searches that allowed us to using string manipulation methods to generate the final library. The application also made use of OpenBabel to provide "clean" SMILES back to the user. The result of the application is that, non-experts in cheminformatics

can easily construct arbitrarily large virtual combinatorial libraries.

Given that the application was used to generate virtual libraries of sizes ranging from 71,000 to 550,000 compounds, it would be infeasible to actually synthesize and test them all. Thus, the next step was to prioritize the library and suggest a small number of compounds to be synthesized. The focus of the study was to identify inhibitors for the falcipain-2 enzyme, a cystiene protease that plays a central role in the erythrocytic stage of the malarial parasite [Sijwali2006]. A crystal structure of this enzyme was available in the PDB (1YVB) and we performed a docking study to identify which of the Ugi products from the virtual library would behave as possible inhibitors. We first generated conformations for a 71,000 member virtual library using Omega 2.2.1 (OpenEye Inc.). Given that conformer generation can be time consuming, we performed the calculation on a Condor cluster. Each node in the cluster was a dual core Intel Xeon (2.4GHz, with 4MB cache) with 4 GB of RAM. By running the conformer calculation process in parallel, we achieved a speed up of 28 over a single processor calculation. Given a set of conformers, we then proceeded to dock them in the active site of falcipain-2. As before, we performed this calculation on the Condor cluster, utilizing 6 (dual core) CPU's and achieved similar speedups. The result of this procedure was that were able to rapidly screen the virtual library and provide a ranking of the compounds. These rankings were then provided to Prof. Bradley who then selected 10 compounds for synthesis and assay. These compounds were then tested in a cell-based assay and four compounds were identified with micro-molar activities *(in vitro* and *in vivo)* and five compounds were identified with similar activities *in vivo* (unpublished data).

5.2. Prediction of anti-cancer activities

The NCI Developmental Therapeutics Program (DTP) provides a 60-tumor cell line dataset, which is a valuable resource for the development of new anti-cancer agents. The dataset consists of assay results that measure the ability of 44,653 compounds to inhibit one or more cancer cell lines. Note that this is a subset of the entire collection (consisting of approximately 250,000 compounds) and represent the compounds for which screening data is available. The cell lines represent a variety of cancers (melanomas, leukimeas, cancers of the breast prostate and so on) and have been collected over a period of 18 years. The readouts represented in the dataset include growth inhibition (GI_{50}), lethal dose (LD_{50}) and total growth inhibition (TGI). In addition to such dose related information, the untreated cell lines have been analysed using micro-arrays providing gene expression information.

We recently described a study [Wang2007] that attempted to perform large scale data mining over the entire 60 cell line dataset using a variety of techniques ranging from a SMARTS based pattern extraction algorithm to decision trees and random forests. In this paper we briefly describe the random forest [Breiman1984] models that were developed as they directly involved our distributed infrastructure. The random forest was selected due to a number of favorable features such as the lack of dependence on feature selection and the ability to avoid overfitting the data. We generated 166 bit MACCS keys to characterize the molecules. We then built the a random forest model for each cell line using the implementation in R [REF]. The models were developed on a local machine, not hooked into our infrastructure. Each model took approximately 16.5 min to build and we employed a simple parallelization scheme allowing us to develop all the models in under 8 hours. Given that the data for each cell line was quite unbalanced (in favor of inactives) we employed a biasing scheme, where we specified that each tree in the ensemble would sample preferentially from the active subset of a given cell line. The performance of the models was reasonable, with the worst model amongst al 60 exhibiting 67% correct predictions overall and the best model exhibiting 77% correct predictions overall. When the active class was considered the performance ranged from 74% to 79% correct predictions over the 60 models.

After developing the models they were deployed into out R web service infrastructure. The 60 models were stored in binary format, generating 3 model files each storing 20 models. These

files were deposited within a remote R server. At startup, the R server identifies available model files and loads them into memory. This is especially useful for the NCI random forest models, since they are quite large in terms of memory usage. As a result, loading them on the fly would be time consuming. Once loaded into memory, we can then access them using the SOAP based web service interface. Specifically, we developed a simple Java class that would accept a SMILES string and would automatically generate the MACCS keys for the input molecule. The class would then connect to the R server via a web service and obtain a prediction. Note that it is also possible for a user to bypass this helper class and directly connect to the R server via the public web service interface. In such a case, they would have to provide their own MACCS keys. The R server returns the predictions of all 60 cell lines along with a confidence measure (defined in terms of the number of trees that predict the compound as being active). The data can then be displayed in any manner. We have provided a simple web page interface, which allows users to paste SMILES strings and obtain a variety of textual and graphical representations of the results. The application can be found at http://www.chembiogrid.org/cheminfo/ncidtp/dtp,

6. Conclusions

In this paper we have presented a review of the start of the art in distributed systems designed to handle chemical information. With the rise in automated, high throughput systems, the large amounts of data that are being generated impose a number of requirements on a storage and analysis framework. First, the resultant data must be easily accessible using standardized protocols. While this is not always the case, the use of relational databases and standard web service protocols (such as SOAP) go a long way to allowing uniform and well defined access methods. Second, it is vital that we have methods that can analyze such data, to generate insight into the workings of the underlying chemical or biological systems. We have presented recent work carried out at the Indiana University Chemical Informatics Cyberinfrastructure and Collaboratory (CICC) that has focused on developing cheminformatics methodologies for a variety of tasks including QSAR, clustering, outlier detection and model domain applicability. In addition to methodology development, we have designed a web service based deployment infrastructure that is suitable for the large chemical datasets characteristic of modern drug discovery efforts. Many of these services provide convenient interfaces to both data sources and computational methods. In the latter case, we have provided SOAP-based web service methods to both cheminformatics and statistical methods. Along with developing methodoligies and infrastructure development, we have created data source derived from PubChem. Notable amongs them, is the Pub3D resource that provides single conformer 3D structures, for 99% of PubChem. The database allows retrieval of structures based on compound ID's or by 3D shape similarity searches and can be accessed by direct SQL or via SOAP web services. We have also described several case studies that have enabled collaboration between cheminformatics experts and non-experts, allowing the latter to make use of cheminformatics technologies without requiring expertise in cheminformatics or distributed systems.

The development of distributed infrastructures stresses the collaborative nature of scientific discovery. In the field of drug design, such an approach is vital given that multiple domains of expertise (chemistry, biology, computation) are required to fully understand chemical and biological systems. A distributed infrastructure, such as described here, allows multiple groups, with different domains of expertise to easily access and analyze a wide variety of chemical and biological data.

7. Acknowledgements

We would like to thank Prof. Jean-Claude Bradley for synthesizing the Ugi products identified in our docking procedure and Prof. Philip Rosenthal (UCSF) for performing the assays.

Table 1: A summary of the various web services currently hosted by the CICC. The services include those developed at the CICC as well as services that are derived from external software packages.

Class	Service	Functionality
Cheminformatics	Similarity	2D similarity using the Tanimoto coefficient and 3D similarity using the descriptor-based method described by Bellester et al[REF].
	Molecular descriptors	TPSA, XlogP and other descriptors
	2D structure diagrams	Generates 2D diagrams from a SMILES string
	Drug-likeness	Currently returns the number of Lipinski failures
	Utility	Fingerprints, file conversion
	CMLRSSServer	Generates an RSS feed from CML formatted molecules
	InChIGoogle	Search Google using an InChI
	OSCAR3	Extract chemical structures from text
	ToxTree	Obtain toxicity predictions
	3D Coordinates	Generate 3D coordinates from SMILES strings
Databases	PubDock	Obtain ligand structures and associated score values by PubChem compound ID or SMARTS based similarity searches
	Pub3D	Obtain a low energy 3D conformation of PubChem structures by compound ID, SMARTS based similarity or by 3D similarity searching
Statistics	Modeling	Builds regression (OLS, CNN, RF) and classification (LDA) models. Perform clustering using k-means
	Feature Selection	Select descriptor subsets for linear regression using backward or forward stepwise regression
	Plots	Generate 2D scatter plots and histogram plots
Applications	Toxicity	Ensemble of random forests that provide predictions of animal toxicity
	Mutagenicity	Single random forest that predicts mutagenicity of a compound given a SMILES string based on data studied by Kazius et al [REF]
	Anti-cancer activity	A set of random forest models that predict anti-cancer activity against the 60 cell lines managed by the NIH DTP program
	Pharmacokinetic parameters	Modified version of the pkCell calculator [REF]

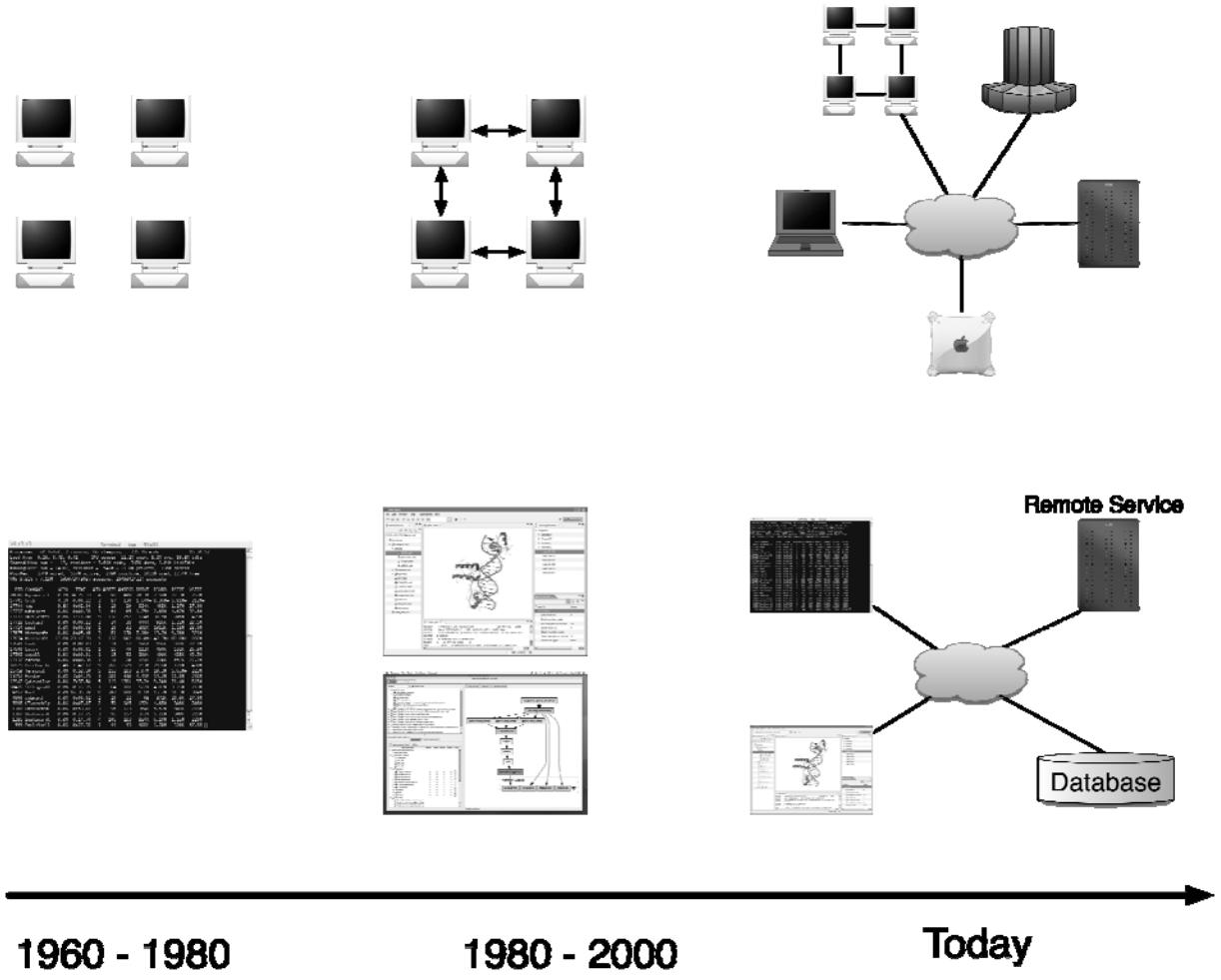

Figure 1. A schematic timeline highlighting the evolution of distributed infrastructure. The upper row highlights the evolution of hardware from single, isolated computers to a grid type infrastructure. The lower row highlights the evolution of software from standalone programs to those capable of interacting with a distributed infrastructure.

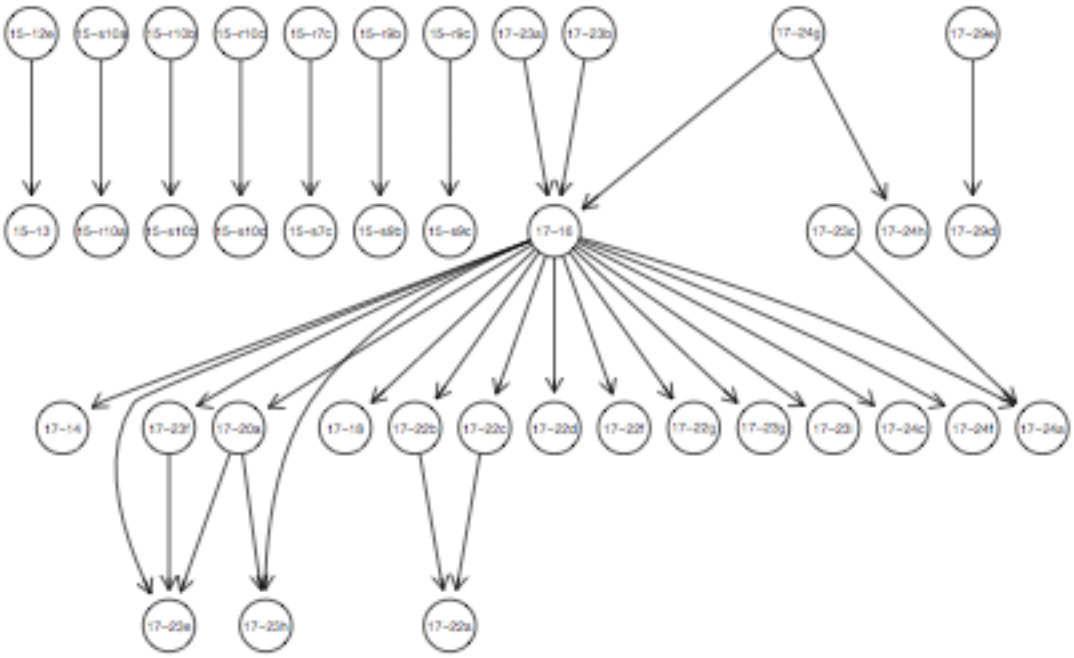

Figure 2. A SALI network derived for a set of melanocortin-4 receptors at a 50% cutoff. Each node represents a compound. Nodes are connected by an ordered edge (compound with lower activity at the tail and with higher activity at the head) if their SALI value is greater than the user specified cutoff.

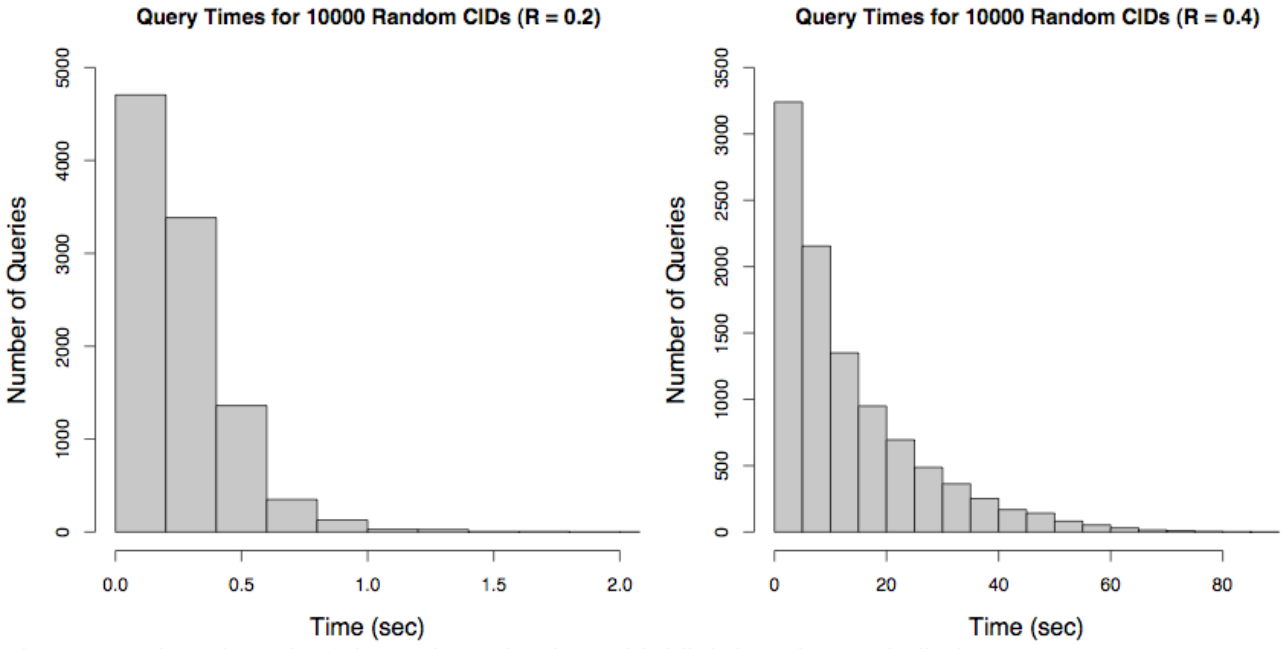

Figure 3. A benchmark of the Pub3D database highlighting shape similarity query performance. The database contained 10 million structures. R represents a radius, larger radii implying more dissimilar molecules. CID's represent PubChem compound identifiers.

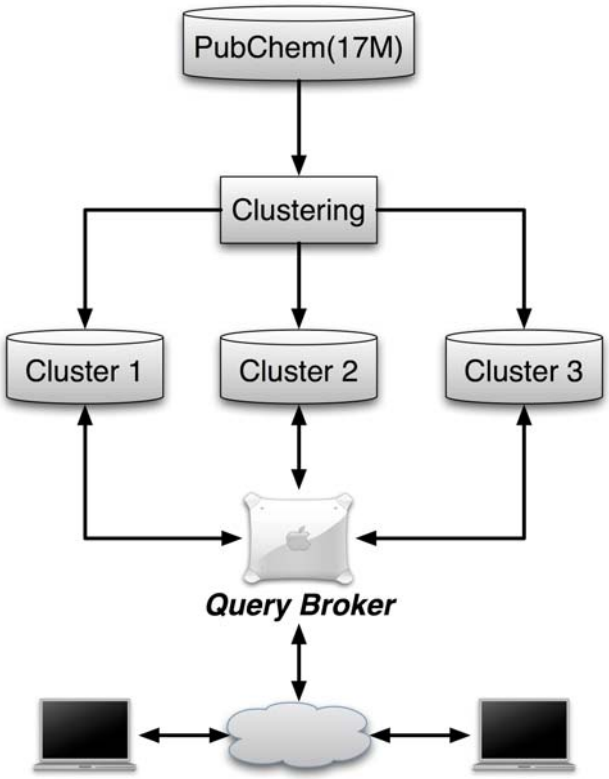

Figure 4. A schematic diagram highlighting the decomposition of the original, monolithic Pub3D database into clusters, which are queried in parallel.

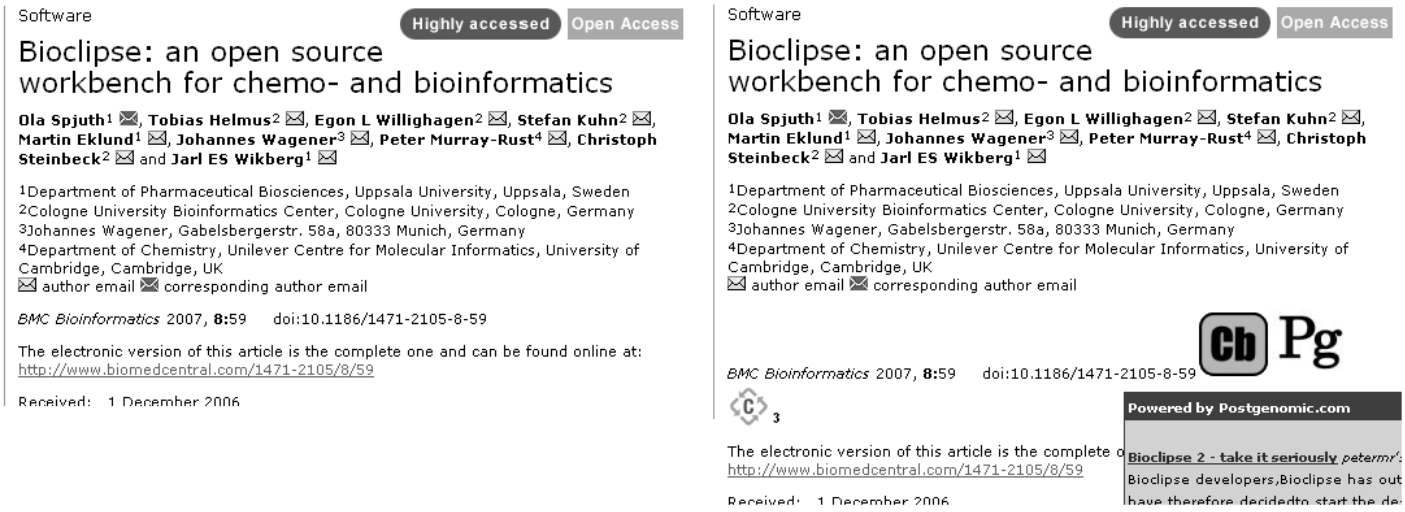

Figure 5. An example of a userscript that modifies online journal web pages to include links to commentary located elsewhere on the Internet. (a) displays the web page in the absence of the userscript. (b) shows the web page when the user script is active.