Cheminformatics for chemists: Improving usability and accessibility of cheminformatics techniques and data

Rajarshi Guha, Gary D. Wiggins, David J. Wild, Mu-Hyun Baik, Marlon E. Pierce, and Geoffrey C. Fox*

Affiliations:
Guha: Indiana University School of Informatics
Wiggins: Indiana University School of Informatics
Wild: Indiana University School of Informatics
Baik: Indiana University Department of Chemistry and School of Informatics
Pierce: Indiana University Pervasive Technology Laboratories, Community Grids Lab
*Fox (corresponding author): Indiana University School of Informatics, Department of Computer Science, Department of Physics, Pervasive Technology Laboratories, Community Grids Lab. Address: Community Grids Lab, 501 N. Morton, Showers 224, Bloomington, IN 47405 E-mail: gcf@indiana.edu Phone: 812-856-7977 FAX: 812-856-797.

ABSTRACT

Some of the latest trends in cheminformatics, computation, and the world wide web are reviewed with predictions of how these are likely to impact the field of cheminformatics in the next five years. The vision and some of the work of the Chemical Informatics and Cyberinfrastructure Collaboratory at Indiana University are described, which we base around the core concepts of e-Science and cyberinfrastructure that have proven successful in other fields. Our chemical informatics cyberinfrastructure is realized by building a flexible, generic infrastructure for cheminformatics tools and databases, exporting "best of breed" methods as easily-accessible web APIs for cheminformaticians, scientists, and researchers in other disciplines, and hosting a unique chemical informatics education program aimed at scientists and cheminformatics practitioners in academia and industry.

1. HOW FAR HAS CHEMINFORMATICS COME?

The practice of cheminformatics has come a long way since the start of the field in the area of chemical information and structure representation. In its current form cheminformatics has become an encompassing field that includes areas such as structure representation and searching, prediction of molecular properties, and visualization of molecular structures and properties. Naturally cheminformatics is very multidisciplinary in nature, utilizing tools and techniques from computer science, mathematics, statistics and visualization. The goal of cheminformatics is to make sense of the large amount of information that is derived from chemical structures and processes. The results must not be presented only as mathematical models or tables of numbers, but in terms of chemistry. Fundamentally, cheminformatics must be able to manage and process the large amounts of data that are characteristic of today's chemical enterprises, and transform the data into usable chemistry. We believe that in the last fifty years, cheminformatics has made significant progress towards this goal, though many problems still exist.

The explosion of data in cheminformatics has come about because of the large increase in the numbers of chemical compounds that have been synthesized or which can be shown to be synthesizable, and a plethora of properties (biological and chemical) that can be measured or calculated for these compounds. One of the largest repositories of such data is PubChem which currently has 10 million chemical structures as well as results from nearly 500 bioassays on about 80,000 compounds. In addition to public resources, there are many proprietary repositories of such information, which are generally not publicly accessible. In addition to these databases, alternative forms of information are becoming increasingly available. Such information may take the form of electronic journal articles, web pages containing structural formulae and other representations as well as discussion forums related to chemistry. These non-traditional sources of

information may not contain an explicit chemical structure. In many cases, text and images must be parsed to identify relevant chemical items, which can then be processed to generate traditional structures and other information. We include algorithms as possible data sources. Given the large number of algorithms and software packages available for a variety of cheminformatics tasks, we believe that it is important to be able to organize these types of resources, so that one can use them in a uniform manner allowing for automation. Thus the first fundamental feature of cheminformatics is the ability to handle multiple and disparate sources of chemical data.[52]

With cheminformatics in its fifth decade (although not under that title[18]), it is interesting that the scope of the term and goals of the field are not well agreed upon. We believe a fundamental goal is to *provide useful information to chemists to help them do better chemistry* in whatever sphere they work; that is, to make use of chemical information to identify the best molecules for a specific purpose and explain the behavior of a given molecule from its structural features. Given the deluge of data described above, today's scientists require not only that answers be quickly available but also that answers can be quickly updated when new data is available. This approach implies that in many cases, one may not be able to get an exact answer. Depending on the situation, an approximate answer may be sufficient and in a number of situations, an approximate answer may be all that is available. In other words a fundamental feature of cheminformatics is its ability to handle large amounts of data efficiently and its ability to provide an approximate answer to the questions: W*hich are the molecules I want?* and W*hat do these molecule do?*

The cheminformatics literature is filled with approaches that attempt to answer these two questions using a variety of techniques. Obviously, if one were able to calculate everything from first principles in reasonable time, we would not need to consider simplified or transformed versions of chemical structures as much of cheminformatics does. Since such first principle calculations are still not feasible we must resort to abstractions that allow us to manage large amounts of chemical data in reasonable time frames. Not only must we perform these analyses in reasonable time, we must remember that the goal of cheminformatics is to provide *useful* information to the practicing chemist so all information must be contextualized for the human and scientific pursuit.[40] We believe that the issue of extracting chemistry from cheminformatics techniques is still open and that much remains to be done to unify cheminformatics resources such that they can be accessed uniformly and combined in innovative ways.

In the following sections we describe some of the areas of cheminformatics that we are tackling in the Indiana *Chemical Informatics and Cyberinfrastructure Collaboratory* (CICC) using a *cyberinfrastructure* approach. By cyberinfrastructure we mean the integration of advanced instruments, computing systems, data storage facilities, visualization environments, software, and people connected by advanced networks to enable scientific discoveries that would otherwise not be possible. Arden Bement, director of the National Science Foundation, has called cyberinfrastructure[10] the engine for change that will drive "a second revolution in information technology, one that may well usher in a new technological age that will dwarf, in sheer transformational scope and power, anything we have yet experienced in the current information age."[17] For reviews and summaries of Grids and cyberinfrastructure, see Berman et al[14] and Foster[26].

We advocate the adoption of an open distributed system architecture that is based on Web Services[11]. This architecture is illustrated conceptually in Figure 1. Web Services provide a universal means for wrapping databases, analysis codes, computational methods, and miscellaneous capabilities such as format converters with well-defined XML interfaces that may be invoked using network messaging-based approaches. The primary advantage of this approach is that it enables us to provide an open, loosely coupled, easily extensible architecture. Other collaborators and partners can easily pick and choose the services they want and integrate these into their environments. Similarly, they can provide back services to the community. Higher level capabilities, such as workflow systems[27] and science gateway portals[9] can be built on this foundation. Further, as we discuss also in this paper, this is no fundamental difference between the so-called Web 2.0 approaches[28] and cyberinfrastructure, and we attempt a blending of these two approaches in detail. By

adopting these approaches, we hope to make our efforts compatible with the general trends in network computing.

2. EXPLAINING CHEMISTRY WITH CHEMINFORMATICS

One of the main focuses of the CICC is to devise and improve techniques for identifying molecules that can be used for a specific purpose. Successful approaches generally involve a range of predictions, which may include protein binding, absorption, distribution, metabolism and excretion (ADME) and toxicity and solubility prediction, *inter alia*. The success of these methods requires a high degree of prediction accuracy, as well as a high degree of accessibility to and interpretability by the chemist.

The use of machine learning methods to predict the properties of molecules from their structural features has a long history, starting with Hansch in the early 60's.[36] This approach is also termed Quantitative Structure-Activity Relationships (QSAR) and has been a vigorous area of research for a long time. The basic idea underlying these approaches is that in many cases, it is difficult or impossible to directly predict the property of a molecule from the molecular structure. As a result we take an indirect route, whereby we consider a set of molecular structures which have measured values of the property of interest. We then derive *numerical descriptors* (also termed *features*) of the molecular structures and attempt to correlate these descriptors to the observed property via a mathematical model. The resultant model is then used to predict the property of a new molecular structure. It should be noted that this approach is very general. Thus, a predictive model can be built for any property that can be measured and there are numerous examples in the literature covering a wide variety of properties.[12, 13, 19, 42]

Though this area of research is well established there are a number of issues that still remain. In particular, we seek to investigate methods that allow one to extract chemical sense from predictive models as well as identify when one can use a model sensibly (i.e., obtain a valid prediction) and when one should not rely on the results of a model.

We first discuss the issue of extracting chemistry from predictive models. Traditionally predictive models have been judged in terms of numerical quantities. These include the statistical methods R^2, q^2 and Root Mean Square Error (for regression models) and the percentage correct classification and area under the curve (AUC) (for classification models). Though these are certainly necessary to judge whether a model is usable, they do not really describe the structure-activity trends that are encoded by the model. Thus one of the aims of the CICC is to devise methods that allow us to explain the encoded structure-activity trends. We have considered a number of approaches to this problem, focusing on individual modeling techniques. Thus for linear regression models we have applied a technique based on partial least squares (PLS)[46] which allows one to explain the effects of individual descriptors on the predicted property. Though this can be achieved to some extent by simply considering the appropriately scaled regression coefficients, the PLS approach allows much more detail when considering the effects of individual descriptors on specific molecules. This approach has been used to successfully extract detailed descriptions of structure-activity trends involved in the anti-malarial activity of artemisinin analogs[33] and the inhibitory activity of piperazyinylquinazoline analogs against PDGFR.[32]

We have also considered the interpretation of neural network models. For this case we devised a broad interpretation method[31] which essentially ranks the descriptors in order of their importance to the model's predictive ability and is based on the concept of descriptor importance used in random forest models.[16] We also devised a detailed interpretation protocol[34] which allows us to analyze the effects of individual descriptors on the predicted output in a manner analogous to the PLS interpretation approach for linear regression models. This approach has been shown to correctly extract the encoded structure-activity trends from neural network models developed for a variety of properties such as boiling point and skin permeability.[34] The disadvantage of this approach is that it *linearizes* the network, thus losing a good deal

of the non-linear encoding of the structure-activity trends. Current efforts in this area involve improving the method to avoid the linearizing approximation as well as make better use of the bias terms in the neural network model.

The result of these approaches has been that rather than simply providing a table of numbers or a scatter plot, we are now able to explain why a certain molecule is predicted to be active (or toxic or mutagenic and so on) based on specific structural features. Naturally, one can obtain similar types of information using other methods (quantum mechanics, docking). Given that predictive models can be used on large chemical datasets, the approaches discussed here allow us to fully utilize the information encoded in these types of models, rather than simply using them for their predictive ability.

One feature of our investigations in this area is that they have mainly focused on the interpretation of regression models. Future directions for this area of research include development of methods that can be used to interpret classification models. Currently, the descriptor importance approach for neural networks and random forest models can be applied to classification problems, but as noted above, this method does not lead to an in-depth understanding of the encoded structure-activity trends.

From the above discussion it is clear that the process of extracting encoded structure-activity trends is dependent on the nature of the descriptors that are used to characterize molecular structure. Currently, there are a wide variety of packages[47, 49] that can be used to calculate molecular descriptors. As a result, one can calculate many thousands of descriptors. However, many of these descriptors are quite abstract (such as topological descriptors) and it is difficult to go from the descriptor values to a real molecular feature. On the other hand, a variety of descriptors are available that make physical sense; that is, it is clear how the descriptor relates to the molecular structure. Thus when interpreting a predictive model it is beneficial to have a model made up of such physically interpretable descriptors. Given that we can calculate thousands of descriptors, we must choose a small subset when developing a predictive model, so as to avoid overfitting.[37] The problem of choosing a small subset of descriptors from a large pool is termed *feature selection* and has been extensively studied[38, 39, 41] in the machine learning and data mining community.

At the CICC we have considered the problem of feature selection in the context of multiple models. In the field of QSAR modeling it is traditionally seen that one builds a linear model for its interpretability and simplicity and a non-linear model for its improved predictive accuracy. However, each model is developed using a different set of descriptors, which is optimal for each individual model. Thus it is possible that the individual models may encode slightly different structure-activity relationships. However if one is able to select a set of descriptors that is *simultaneously* optimal for multiple types of models (say a linear regression model and a non-linear neural network model) then one of the models can be used for its greater interpretability and the other model for its higher accuracy. This approach would also be useful when developing models for ensemble predictions.[8, 22] In either case, by using the same set of descriptors in multiple types of models, one may expect that they now consistently encode the same structure-activity relationships. Our approach[25] has been to use a genetic algorithm in which the objective function is a linear combination of the error functions of the individual model types. Naturally, by forcing different types of models to use the same descriptor subset, the models may not be optimal in terms of their predictive ability. However our results[25] indicate that the decrease in predictive ability, as a result of the constraint of using the same descriptor subsets, is usually less than 10% for both regression and classification models. Interpretations of the models indicate that both models encode more or less the same structure-activity trends.

3. WHEN IS A MODEL USEFUL?

Though interpretability is a very important aspect of predictive models in cheminformatics, it is equally important to understand when a model returns valid predictions. That is, a model that is developed using a collection of molecules (termed the *training set*) will be able to give reliable predictions for new molecules

that are similar in nature to the training set that the model was built with. When a model is asked to predict the property of a new molecule that is significantly dissimilar to the training set, it will return a predicted value. However, can we be sure that the model has given us a *reliable* prediction, given that, by definition, it has never seen the features present in the new molecule? This problem is formally stated as identifying the domain of applicability of a model. This problem has become increasingly important as the use of *in silico* predictive models has increased, especially within regulatory agencies. A number of workers have addressed this issue[44] using techniques ranging from similarity to the training set to probability contours.

Our previous investigation[30] in this area has focused on the use of an auxiliary model, which was used to predict whether a new molecule would have a high or low prediction error. The auxiliary model was developed using the training set residuals of the model, where the residuals were arbitrarily divided into two classes – high and low. The auxiliary model was then used to predict whether a new molecule would have a high or low residual. We considered a variety of methods, and a neural network model was able to predict the class of the residual correctly more than 90% of the time. The approach does exhibit a number of disadvantages such as being focused on a binary classification of residuals. More importantly, the use of an auxiliary model leads to a recursive solution to the problem. That is, how does one measure the applicability of the auxiliary model? Clearly this is not an optimal approach. The underlying feature of this approach and a number of approaches described in the literature is that they focus on the model space. Thus the domain of applicability is defined as the space within which the training set exists. As a result, if a new molecule does not lie within this space it is expected that its predicted property will not be reliable. Our current lines of investigation in this area attempt to better quantify the chemical space and the relationships between the training set and a new molecule in this space. In this context we are investigating the use of external descriptor sets that allow us to avoid being restricted in the chemical space of the model and investigating density based methods which consider the population density in different regions of the chemical space, rather than distances between objects in the space.

We believe that providing more rigorous approaches to extracting structure-activity relationships from QSAR models and quantifying the domain applicability of these models will not only make these mathematical models more accessible to the practicing chemist, but will also increase their reliability and utility. In addition to providing chemical insight into mathematical models of structure-activity relationships, the techniques being investigated should allow us to more reliably identify promising leads when such models are used in virtual screening protocols.

4. EXPLORING CHEMICAL SPACES

As noted above, one can calculate many thousands of molecular descriptors. Given a large pool of descriptors one usually performs some form of feature selection to choose an information rich and relevant set of descriptors. This subset is then used to perform modeling, searching or some other task and represents a chemical space. In many situations it is beneficial to understand the distribution of compounds in such a space. Exploring a chemical space can lead to various types of useful information. The most familiar would be the fact that compounds located close to each other in a space will exhibit similar properties. Alternatively, identification of compounds that are located in a very sparse region of a space might lead to interesting new leads and thus serve as the starting point for lead hopping[21]. Our goal in this are has been to develop intuitive methods to characterize chemical spaces and distributions of compounds in these spaces, as well as pursuing methods which identify interesting, related structures to probes or query compounds. We recently described a method,[29] termed RNN curves that uses a nearest neighbor approach to characterize the spatial location of each point of a dataset in a given chemical space. It is important to note that rather than consider the k-nearest neighbors we consider the number of neighbors within a series of radii. Traditionally this approach has a running time that is quadratic with the number of points in the dataset, though this can be improved by use of data structures such as kd-trees. We also investigated[24] the use of an approximate nearest neighbor detection technique that runs in sublinear time, allowing this approach to be applied to large

datasets. Our initial investigation focused on nearest neighbor detection using the Euclidean metric. However by considering weighted distance metrics, whereby different dimensions of the chemical space are given different weightings, we can bias the nearest neighbor search to relevant descriptors. Our current investigations focus on the use of this approach to identify the natural numbers of clusters in a dataset, identifying good lead hopping candidates, as well as applying it to the problem of the domain applicability of QSAR models. Finally, we are further developing 3D scaffold searching techniques based on atom mapping techniques recently applied by one of the authors.[15]

5. CHEMINFORMATICS AND THE WEB

As noted previously, an increasing amount of chemistry information is available on the internet and intranets, often in forms other than traditional, searchable databases (electronic journal articles, web pages, online documents, blogs, RSS feeds and so on). In addition to these forms of data sources, current models of usage of chemical information also include Internet based access in addition to traditional forms such as local desktop clients.

One of the fundamental problems that faces the cheminformatics community is the fact that though there are many novel algorithms being developed as well as new data sources being created, they more or less exist as islands. That is, the implementation of these algorithms and the databases containing new data, require that one visit a web page and either download a collection of source files or enter data into a web form which then returns results in the form of an HTML page or in some cases an email. Though this is certainly better than not being able to access these resources at all, such modes of access create bottlenecks for interoperability.

This problem is being tackled in the web community in general by the use of *web services* which allow computational procedures to be accessed through standard web protocols, and the *semantic web* which promises to help interoperability of services and information through the use of standardized markup and ontology languages. We believe that this approach is highly appropriate for cheminformatics. The cheminformatics community has developed a number of algorithms for a variety of problems such as structure searching, descriptor calculation, similarity and so on. Many of these are bound to a specific web page or require one to download a program. It would be better if all these algorithms were accessible in a uniform manner (at least those algorithms that perform similar functions) without binding a user to a specific web page interface. Additionally, the development of workflow tools such as Pipeline Pilot[3] and Knime[2] has led to a significant increase in the usability of cheminformatics tools when handling large datasets. In such a setup, having to go outside the environment to generate data which must then be imported back into the environment severely hampers the whole idea of a workflow environment. Now, it is certainly possible that a given algorithm might be implemented in a specific workflow tool. However this requires that the developer of the algorithm have access to the workflow program and be familiar with it. Furthermore, this would require reimplementation for each new workflow tool. Clearly, a standardized approach to accessing new algorithms and data sources would easily allow arbitrary workflow tools to include them. Finally, access to a wide variety of algorithms and data sources in a standardized and distributed manner will lead to novel uses of such resources that may not have been considered by their designers. For example, structure searching algorithms have a well defined function, but when coupled to new data sources (such as journal articles or blogs), one can envision new resources that go beyond the traditional view of searching structure databases for similar compounds. The explosion of innovation around the Google Maps API should serve as an inspiration as to what is possible when computational capabilities are made highly accessible.
The vision for cheminformatics described above depends on the presence of an infrastructure. At this point, such an infrastructure is not widely available, though many of the underlying technologies have been available for some time. In the following sections we describe approaches taken by the CICC to achieve the vision of a distributed collection of cheminformatics resources that cover both algorithms and data sources.

6. A WEB SERVICE INFRASTRUCTURE FOR CHEMINFORMATICS DATABASES AND TOOLS

With the aim of making a wide variety of algorithms and data sources available in a distributed and standardized manner, we have developed a variety of web services.[23] A web service is essentially a remote function call that allows a user to send an argument (which may be a simple SMILES string or something more complex such as a 2-dimensional data matrix) and obtain a result (which again, may of different types). In this sense a web service is much the same as a traditional library call. However a web service provides a number of advantages. First, web services are generally language agnostic. Thus one may write a web service in Java, but be able to access it in any language (Python, C, Java, Ruby, etc.) that can communicate with web services. Second, by allowing a web service to reside anywhere on a network (which may be the Internet or a local network), we are not required to have the algorithm or data reside locally. This may be important when the developer of the resource does not want to make the actual source code public, but does want to allow users to access the functionality. Third, by presenting arbitrary functionality as a web service, one is no longer restricted to the manner of access as defined by the original developer. Thus for example, consider an algorithm that evaluates the 3D similarity between two structures. The developer of an algorithm could implement a command line or GUI program which a user would then download and run locally. However this would not be very useful if the user was using a workflow environment. Alternatively, the developer could design a web page where a user would upload two structures and get back an HTML page that printed the similarity. Though useful, this is not helpful if one needs to process many structures in an automated fashion. By exposing the similarity function as web service, the user can now access the functionality in a variety of ways. The user could design a client that is simply a web page. However, the user is not bound to such an interface. If the user's workflow tool supports web services, he could simply include the similarity web service as another entry in the workflow tool's palette. Alternatively, the user could write his own command line or GUI program which would utilize the similarity web service.

The above discussion has focused on the presentation of specific algorithms as web services. However we should also point out that web services are not restricted to this. That is, arbitrary programs (which have a command line interface) can also be wrapped and presented as web services. Furthermore, the web services approach can also include commercial programs and algorithms by the use of appropriate authentication mechanism. In addition, databases can be exposed as web services. This is especially useful since one is no longer restricted to using SQL queries or a static web form. We describe these aspects in more detail below.

In addition to the above advantages, the use of web services means that a given user does not need to maintain a variety of software packages and databases. As long as standard interfaces are designed and standard protocols are used for communication, the user need only worry about what to send to a service and what he will get back.

Though web services are not common in the field of cheminformatics, they have existed for some time in other fields. As a result, the design of web services is rather straightforward in terms of communication protocols and hosting environments. The bulk of the services developed and hosted by the CICC are written in Java and hosted in a Tomcat application container. However it should be noted that one is not restricted to Java for the development and deployment of web services and that numerous other languages can be used. The source code for the web services developed by the CICC is freely available and can be obtained from the Subversion repository located at http://sourceforge.net/svn/?group_id=163501. In addition, we provide complete Javadocs for each service as well as Junit tests, which provide working examples of the usage of the web services.

The underlying protocol that allows arbitrary clients to communicate with web services written in arbitrary languages is the *Simple Object Access Protocol* (SOAP). This protocol is based on XML and allows a program to send data to a service and receive data from a service without having to worry about service-specific encodings. SOAP allows one to handle primitive data types (int, string, float) as well as composite types such as arrays, hash maps and so on.

Given a mechanism by which clients can communicate with services, the next requirement is to identify what types of arguments must be sent to a service and what type of return value will be received from a service. The input / output specifications for a service are described in a *Web Services Description Language* (WSDL) document. This document will define how many and what types of arguments a service will accept, what type of value it will return to the caller and what (if any) exceptions may be generated by the service. Thus for a client to be able to use a service, it must first access the WSDL for the service. This leads to a broader question: given a multitude of services at different locations how does a client know what is available? This is an open question, though the web service community has made a number of advances. The CICC provides this information in three different ways. First we collect the links to the WSDL for each service on a web page. One can visit this page and copy the links to access the web services. In addition to web services developed by the CICC we also collect WSDL links to services provided by other organizations such as Cambridge University and NCI. However visiting a web page to retrieve links does not lend itself to automation. As a result we also host a UDDI registry which allows automated discovery of available web services. Finally, we also provide an RSS feed of web services (available at http://www.chembiogrid.org/cheminfo/wsrss/wsdlrss/getFeed) . RSS feeds have traditionally been used to syndicate news items. We expect web services to grow in importance. As new web services are developed, an RSS feed that syndicates the WSDL for each web service provides an easy and low-cost way to keep track of what services are available. Unlike a UDDI registry which requires special client software to interact with it, an RSS feed of web services can be viewed in any standard RSS viewer.

The CICC currently hosts a number of web services which include services that provide specific functionality in a specific area as well as services that represent complete applications. Table 1 lists the web services currently hosted by the CICC. As can be seen we provide services that cover core cheminformatics functionality, database access, and statistical methods.

The core cheminformatics services include functionality to perform fundamental cheminformatics tasks such as evaluating 2D and 3D similarity between molecules, evaluating molecular descriptors, generating 2D structure diagrams and 3D coordinates from SMILES strings, file format conversions and so on. It is clear that on their own many of the individual services do not provide a complete cheminformatics platform. However that is not the aim. The idea underlying these services is that they can be used as components on the client side, which may be a standalone program, a web page or a workflow tool. In general most of the cheminformatics services accept a SMILES string, though some services such as the descriptor and 3D coordinates services also accept an SDF formatted file as input. Depending on the nature of the service the output may be a simple number or a complete structure in some specified format. The bulk of these services are implemented in Java and are based on the Chemistry Development Kit,[47] a Java library for cheminformatics. As a result, much of the underlying code has already been written and the service is simply a wrapper around the relevant functions from the CDK library.

A number of the CICC services are actually more involved than simple core functions. Examples are a toxicity prediction service and a docking service. The former was originally a GUI program developed by the IdeaConsult[4] which used a decision tree algorithm to predict the toxicity class of a given molecule based on the approaches of Cramer et al[20], Verhaar et al.[50] and Walker et al.[51] We were able to extract the core functionality of the program and wrap it in a web service, thus allowing access to the program without having to use the original interface. In the case of the docking service, we use the FRED docking program developed by OpenEye[5]. This is a command line program that requires a number of parameters to be set. The program was wrapped in a Java web service which creates the command line invocation of the program and returns the resultant docked poses back to the caller of the web service. It should be noted that this is a commercial program and thus the service is not freely available to the public. In general, proprietary programs can be easily presented as web services and proper authentication mechanisms can be employed to ensure that only the authorized users can access such a service.

Given the above services, it is natural to ask for clients that can make use of the services. As mentioned above, the source code for the services includes examples of their use in clients. However, we have stressed the fact that one can write clients that take on a variety of forms, ranging from command line applications to web pages. Thus many of the cheminformatics core services are used by web page clients. These clients appear to be standard web forms utilizing CGI. However underneath the web pages the clients make the appropriate calls to the web services to obtain the results. The web page at http://www.chembiogrid.org/projects/proj_applications.html provides a number of links to such clients. One of the examples is a simple form that allows you to specify two SMILES strings and evaluate the Tanimoto similarity between them. More involved examples include evaluating molecular descriptors and generating 3D coordinates. Of course, clients are not restricted to web pages and we have written examples of command line clients as well as incorporated web services in workflow tools.

The next class of services revolves around database access. The CICC maintains a local mirror of the PubChem database which is updated on a monthly basis. The goal of this mirror is to provide easy access to compound and bioassay data for various projects that the CICC is involved in. A number of these projects aim to add value to the data that is present in PubChem. We discuss two such projects and how web service interfaces provide easy access to the data that is being generated. The first project is aimed at providing a database of coarse-grained docking results for a 960,000 compound drug-like subset of PubChem. The long term goal of this project is to dock this subset into all protein structures that are available in the PDB, whose binding site is known. The total number of such protein targets is approximately 1700. We are using OpenEye's FRED to perform the docking using a variety of scoring functions. The docked ligands and the individual components of the scoring functions are stored in a PostgreSQL database, which is searchable by SMARTS using the gNova cartridge. Thus one can construct SQL queries which will identify the best fitting ligands for a given protein, identify proteins to which a given ligand has been docked and so on. It should be noted that we are not performing high quality docking. Rather we view the results as a coarse filter that will allow us to ignore molecules that do not obviously fit a given protein target. Given that the database is accessible via SQL queries one might think that that is all there is to it, but many users of the database will not know SQL. Thus the natural step is to provide a web page frontend to the database and this is available at http://www.chembiogrid.org/cheminfo/dock/. However this web form does not allow easy inclusion of the results of queries into other applications. Thus we provide web services that essentially encapsulate a number of SQL queries. The web services allow one to retrieve ligand and protein structures, score values, perform SMARTS searches and so on, using well defined interfaces that do not involve SQL.

The second database that we maintain is a collection of 3D structures derived from the 2D structures stored in PubChem. Currently PubChem has about 10 million unique chemical structures, of which we have been able to generate 3D coordinates for approximately 9.99 million. The coordinates represent a single low energy conformer generated using the MMFF94 forcefield[35]. The structures are then stored in a PostgreSQL database, which is searchable by SMARTS using the gNova cartridge. In addition, we also allow queries based on 2D and 3D similarity. As before, the database is accessible via raw SQL queries as well as a web form (http://www.chembiogrid.org/cheminfo/p3d/). We also provide web services that encapsulate a variety of queries. Therefore one does not necessarily have to know SQL or use the simple web page to access the data.

An interesting application of these web services is to use them to generate RSS feeds for database searches. For example, one may query the structure database for the term "thiol". This would ordinarily return a static web page containing a list of structures whose name contains the word "thiol". Instead we can generate an RSS feed for the search, which can then be viewed in any RSS reader. Whenever the user refreshes the RSS feed (which can be performed at predefined intervals if desired), any new molecules that match the original query will show up in the feed. We provide RSS feeds for both of the databases described above, specifically, CMLRSS[43] feeds which include 2D and 3D (if available) structure information. Although a traditional

RSS reader will only be able to view the name and other textual information for the element of the feed, a CMLRSS-aware reader such as Bioclipse[45] will also be able to view the structures embedded in the feed.

The next class of services consists of those that provide statistical and mathematical functionality. These services are based on the R software package[48] and include services that provide access to model building methods (linear regression, neural networks, random forest and linear discriminant analysis), clustering methods (k-means), plotting methods (2D scatter plots and histograms) as well as miscellaneous statistical functions (sampling distributions and statistical tests). It should be noted that the design of these services is based on the Rserve package which results in the fact that the underlying computation engine can be located on a different machine than the one hosting the web services. This allows us to host the web service interfaces on one machine, but perform heavy duty computations on a more powerful, separate machine. Another result of the use of R as our computation engine is that it is possible to present arbitrary R code as a web service. We have provided a number of examples of this feature. For example, we were able to convert the PkCell[55] program to R code, which was then made available as a web service. We then developed a web page client (http://www.chembiogrid.org/cheminfo/pkcell/) that allows the user to evaluate pharmacokinetic parameters for arbitrary molecules.

The fact that the R-based web services provided access the full functionality of R allows us to approach the problem of model deployment. Traditionally, when one develops a QSAR model using a specific statistical package, the user must be familiar with the statistical package to be able to obtain new predictions from the model. In some cases, it is possible to wrap the model in a web form so that a non-expert does not need to learn the mechanics of a statistical software package, but this means the user is still restricted to using the web page frontend. Since R is capable of storing arbitrary models, we are able to develop a variety of models and deploy them within the R web service infrastructure. This allows us to implement a client in the form of a traditional web page, but in addition include these models in workflow tools, command line programs or any other type of client program. This approach clearly expands the utility and usability of models and also extends their use beyond expert users of statistical software. Examples of models deployed in this manner include an ensemble of random forest models that predict the toxicity of a given compound using BCI fingerprints and a set of random forest models that predict the anti-cancer activity of a compound against the 60 cell lines hosted by the NIH Developmental Therapeutics Program. Each of these models can be accessed via a web service and Figure 2 shows a screen shot of the web page client for the anti-cancer models.

Of course the R-based web services do have a number of downsides at this point. One of main disadvantages is that it restricts us to developing models in R. Though there are many good reasons to do so, it is evident that many people will prefer other packages for their statistical work. At this point, our infrastructure does not allow us to include models from other packages. Another issue is one of standardization. Though it is easy to save and deploy an R model, we cannot easily evaluate arbitrary descriptors for use in a model. That is, a model developed by the CICC will use a set of descriptors accessible to the CICC. Thus when a deployed model is to be called we are able to evaluate the descriptors and generate predictions. For an arbitrary model supplied by an outside organization, it may not be easy and in some cases may be impossible to evaluate the required descriptors. This leads to the issue of standardization, which is discussed in the following section.

7. STANDARDIZATION ISSUES

Standardization is an important topic in various areas of cheminformatics ranging from structure representations to structure names. Currently the CICC is focusing on standardization issues related to the deployment of predictive models and associated molecular descriptors. As noted above, our current web service infrastructure requires that models be stored in an R binary format. In addition to storing the model itself, it is equally important to store metadata associated with the model. Examples of such metadata include the author of the model, the date of development (or contribution), the algorithms employed, data sources and

descriptors used. Certain attributes can be obtained by interrogating the model object, though this is specific to the R platform. A more general approach that we are considering is to associate each model with an external XML document that will allow us to add metadata using well known standards such as RDF. This has two advantages. First, we are no longer limited by the R platform as the metadata document can be used to describe models generated using arbitrary packages. Second, given that we employ an XML format for the document we are easily able to include other XML resources within such a document. This implies that we do not need to reinvent mechanisms for attribution (which is handled by the Dublin Core specification) and so on. This is especially useful when we consider the issue of how to note what descriptors were used in a model. This is a potentially troublesome aspect of a standardized model infrastructure, as different programs may implement the same descriptor using slightly different names and may even use slightly modified versions of the original algorithm. As part of the Chemistry Development Kit, we are involved in the development of an RDF based descriptor dictionary which allows us to unambiguously identify each descriptor implemented in the CDK. Since this dictionary is an XML document it is easy to incorporate the relevant information within the model metadata document mentioned above. Of course, this does not solve the problem entirely since the descriptor dictionary is specific to those descriptors implemented in the CDK. However we believe that this approach has significant benefits in terms of interoperability and extensibility and hope to encourage the inclusion of external descriptors within such a dictionary. Naturally, such an approach must involve the community at large and we hope that the approach used in the CDK descriptor dictionary will provide a starting point for some sort of standardization of molecular descriptors.

8. CHEMINFORMATICS EDUCATION THROUGH LOCAL AND DISTANCE LEARNING

The Indiana University School of Informatics has dynamic, innovative programs in several areas of science informatics, including cheminformatics. There are chemoinformatics learning opportunities for both on-site undergraduate and graduate students, as well as for people outside the academy through distance education (DE).[54] At present, Indiana University is the only academic institution in the United States with formal graduate degree programs in cheminformatics.[53] This is confirmed by the continuously updated survey of informatics programs from the University of North Carolina (Chapel Hill) School of Information and Library Science.[6] Furthermore, IU has developed a number of courses and workshops in grid computing and Web services. The marriage of these two areas strengthens both.

It is a common perception that the US is not keeping up with work going on in the rest of the world in e-Science (cyberinfrastructure), and the same could be said for cheminformatics, since most of the major academic advances of the last few decades have emanated from Europe. One need look no farther than the Obernai Declaration which was adopted by 100 scientists in mid-2006 to recognize that Europe is not going to stand still in cheminformatics.[7] The declaration seeks an increase in funding for cheminformatics research and teaching in Europe and makes the following point:

"The further development of chemistry in general and chemoinformatics in particular needs an increase in teaching of chemoinformatics. This is necessary:

- to provide chemoinformatics specialists for academia and industry
- to train chemists in the use of chemoinformatics methods in all areas of chemistry."

Our efforts to combine cheminformatics with cyberinfrastructure are helping to bridge the gap between the US and Europe, as the trend toward distributed, interoperable digital object representations on the grid and data intensive science gains momentum. The Indiana University School of Informatics and the IU Community Grids Laboratory are developing plans to spread the knowledge of their respective areas of expertise among the chemistry and life science communities in the US and potentially throughout the world. In terms of formal degree programs, Indiana University offers the BS in Informatics with a specialization in chemistry, the MS in Chemical Informatics, and the Ph.D. degree with a track in chemical informatics. We

want to broaden cheminformatics education at IU to the full continuum—from one-day courses to the PhD—with training components that complement the formal education programs, including summer schools, tutorials, re-training, and continuing education. Chemistry departments are traditionally very pure in what they think belongs in the chemistry curriculum. Therefore, they need service centers in cheminformatics to supplement their efforts. We will link participating schools via distance education and develop curriculum packages that they can use, while serving as a reference and referral center for expert guidance on the teaching of cheminformatics.

9. FORMAL CHEMINFORMATICS EDUCATIONAL ACTIVITIES AT INDIANA UNIVERSITY

BS in Informatics with Cheminformatics Specialization

The Bachelor of Science in Informatics requires a cognate (specialization) in a subject discipline. For a cheminformatics cognate, a student currently takes the equivalent of an undergraduate minor in chemistry (17 semester credit hours) and two courses that are undergraduate versions of I571 and I572 (see below). A minimum of 35 credits of Informatics courses is also required.

MS in Chemical Informatics

The M.S. in Chemical Informatics Program admitted its first students at IUB in 2002. This is a 36-semester-hour degree, with both core informatics classes and core cheminformatics courses among the 30 hours of required coursework. The final 6 hours of credit are spent on a capstone project that demonstrates the student's mastery of the tools and techniques learned. Specific requirements for the chemical informatics MS are:

- I501 Introduction to Informatics (3 cr. hrs.)
- I502 Information Management (3 cr. hrs.)
- I571 Chemical Information Technology (3 cr. hrs.)
- I572 Computational Chemistry and Molecular Modeling (3 cr. hrs.)
- Electives: 18 hrs., including (strongly recommended):
    - I519 Bioinformatics: Theory and Application (3 cr. hrs)
    - I529 Bioinformatics in Molecular Biology and Genetics: Practical Applications (4 cr. hrs.).

Ph.D. in Informatics: Chemical Informatics Track

This is the standard 90-hour Ph.D. program offered through the Graduate School of Indiana University. The first group of Ph.D. students entered the program in August 2005. In addition to the core courses for the M.S. degree, students must take at least two advanced seminar courses in cheminformatics. The goal of the Ph.D. program is to produce graduates with a multidisciplinary education. Graduates should be comfortable discussing enzymology, organic chemistry, quantum mechanics, databases, e-Science and programming, while being specialists in areas such as grids, artificial intelligence, and data mining of gigantic datasets, such as PubChem.

Continuing Non-Degree Students

The Graduate Certificate Program in Chemical Informatics is awarded to those who complete I571 and I572 plus the following courses:

• I573 Programming for Science Informatics (3 cr. hrs.)
• I553 Independent Study in Chemical Informatics (3 cr. hrs.)

The certificate is available both to on-site and DE students. Among the out-of-state students who have enrolled for the I571 class are a patent specialist and a chemist in pharmaceutical companies, and a researcher at a major bio/cheminformatics software company.

Our experience in teaching cheminformatics at a distance goes back to 2001 when the first cheminformatics undergraduate classes were taught by Polycom teleconferencing link between Indianapolis and Bloomington.[54] The DE program was expanded to include students from around the country and to offer graduate courses by using Macromedia Breeze to view the slides and phone connections for the audio portion. One of the target groups in this endeavor is people moving from an earlier career into the field of cheminformatics. Typically, they lack any kind of formal training in the area, yet they are highly motivated. As one of our external advisers to the NIH grant project recently put it, “We have yet to see an example of a cheminformatician going back to the laboratory, but there are plenty of examples of chemists going the other way.” This underscores the need for more outreach to provide a sound educational foundation to those who are making the move from the laboratory to the computer end of chemistry.

An area that people new to the field, whether students in the academy or outside, are likely to have little inkling of is the vast and rich resources for chemical information provided by commercial and society publishing houses over many decades. In order to provide an easy introduction to such resources, we have developed the Chemical Information Sources Wiki.[1] Included are chapters on such topics as chemical structure searching, searching by chemical names, finding synthesis routes, etc.

Enrollment

In the past two academic years, 124 people have been exposed to cheminformatics instruction through our teaching efforts at Indiana University, including 30 distance education students. With four PhD students and eight MS students currently enrolled in the graduate cheminformatics programs, we are assured of a healthy core of research help as we continue our research and teaching efforts in the fast growing field of cheminformatics.

9. SUMMARY

In this paper we have attempted to highlight some of issues that the field of cheminformatics currently faces. Fundamentally, cheminformatics attempts to identify molecules with specific properties as well as explain why molecules exhibit certain properties. In many ways these aspects are similar to traditional computational chemistry. The difference arises in the sources of data. In the field of computational chemistry, methods focus on specific molecules and proteins and try to derive molecular properties based on first principles. This results in a high computational cost. On the other hand the techniques of cheminformatics are, in general, able to work with collections of molecules ranging from a few hundred to many thousands and even millions. Naturally, cheminformatics methods do not always give an exact answer to the question we ask. However, even an approximate answer can winnow these large collections of molecules to a smaller set that can then be investigated in depth by more computationally intensive methods. However, cheminformatics methods do not focus exclusively on the evaluation of molecular properties. Recent developments in web technologies have given rise to a large number of alternative data sources which include electronic journal articles, RSS feeds and so on. The field of cheminformatics attempts to handle these disparate resources and provide an intuitive and uniform mode of access to the chemical information contained in them.

Given the above characteristics of cheminformatics, we must not forget where the field has come from and what is its primary purpose. The role of cheminformatics in today's research is to enable chemists to better use the huge amounts and wide variety of information that is available to them. That is, the results of cheminformatics must help a chemist make better decisions in terms of chemistry. This has been true in the past and there are documented cases where cheminformatics techniques have led more quickly to better

drugs. However, we believe that there is much more to do. Though cheminformatics methods have played important roles in various chemical enterprises, these methods are still the domain of experts. We do not expect that non-experts will play a role in the development of new cheminformatics techniques, but we do expect that any chemist should be able to easily use these techniques and extract chemical meaning from the results. It is to that end that our educational web services activities are in part devoted.

To achieve this vision, we have focused on four tracks. The first track is the development of a distributed infrastructure that allows us to integrate a variety of data sources and computational methods. We have focused on the use of web service as an underlying technology as this leads to a high degree of flexibility. Coupled with recent developments in thin clients and mash-ups, this infrastructure allows us to combine a wide variety of functionality in ways that the original designers might not have thought of. However, not only can we combine resources, the infrastructure also allows us to present information in novel ways as well as to include the resources and information in preexisting frameworks (such as workflow tools and so on). The second track involves the investigation of methods that extract chemical meaning from mathematical models. Thus rather than consider numerical values, we are devising methods that allow a chemist to easily interpret chemical trends encoded in these models. It is also equally important to provide a chemist with some idea of when a model can be used and when it cannot. We believe this is vital to prevent the use of models in scenarios that they were not meant to handle. Given that much of cheminformatics is based on a variety of representations of chemical structures, resulting in a multiplicity of chemical spaces, we are also investigating intuitive ways to explore and characterize these spaces. The common thread underlying the approaches in this track is that we use state of the art modeling techniques, developing them in such a way that they do not remain as numerical black boxes, but rather, yield useful chemistry. The final track focuses on the use and integration of disparate sources of information. Thus we aim to extract chemistry from journal articles, from web pages, RSS feeds and so on. A major effort in this area is the addition of chemical semantics to a variety of data sources, including web services, descriptors and so on. Underlying this effort is the development of ontologies. There is much evidence for the utility of ontologies in a variety of fields and we believe that the development of ontologies in chemistry will lead to a significant increase in usability and interoperability of chemical data resources. Finally, we are creating a comprehensive, local and distributed chemoinformatics education program to expand knowledge of cheminformatics tools and techniques.

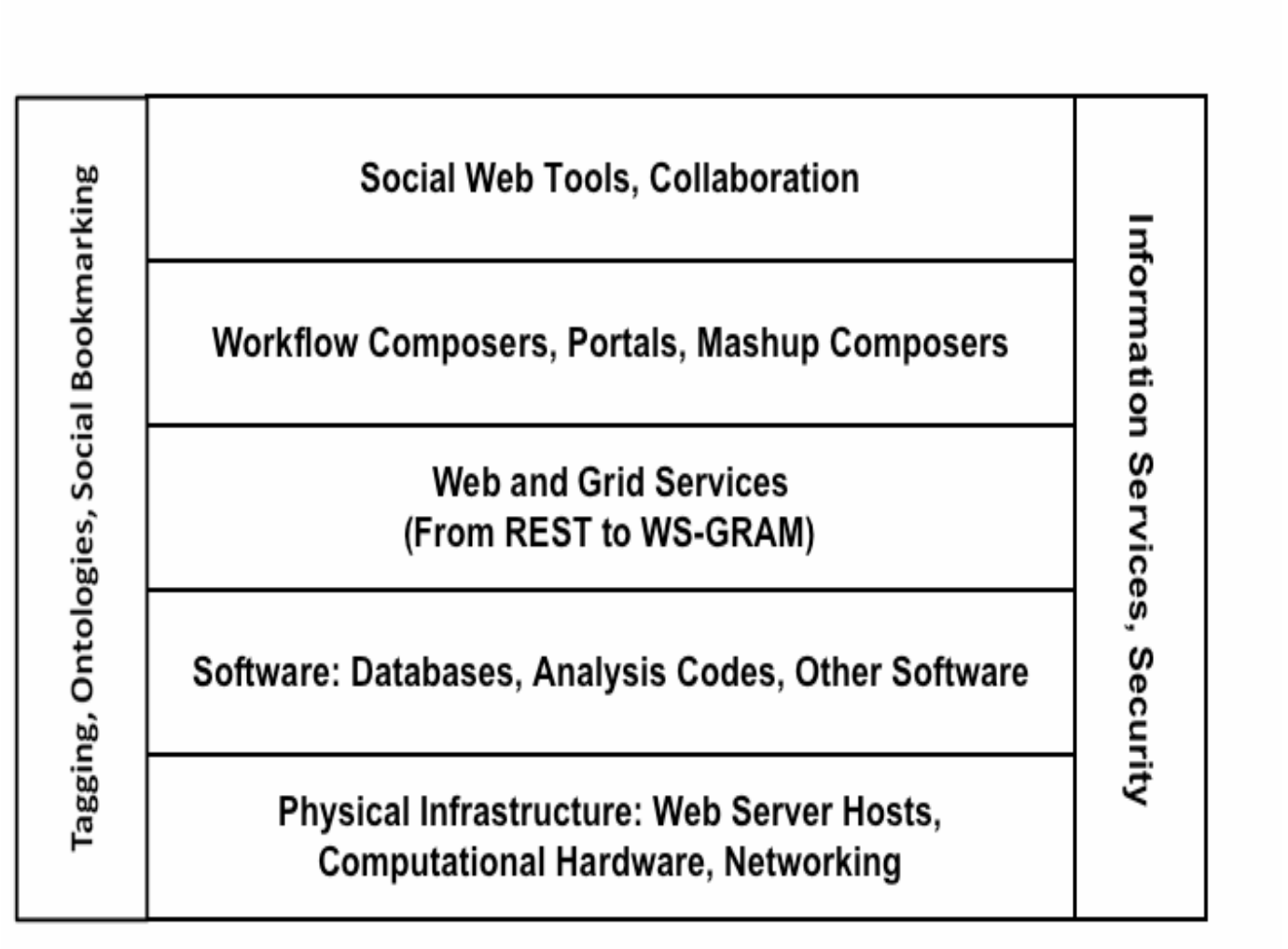

Figure 1 illustrates the general cyberinfrastructure "layer cake" that can be adapted to Chemical Informatics. In going from the base to the top, we proceed from physical infrastructure to software to Web enabled capabilities (Web Services) to service client environments to scientific collaboration tools. The individual layers are loosely coupled: the same web services can be used by many different workflow composers, for example.

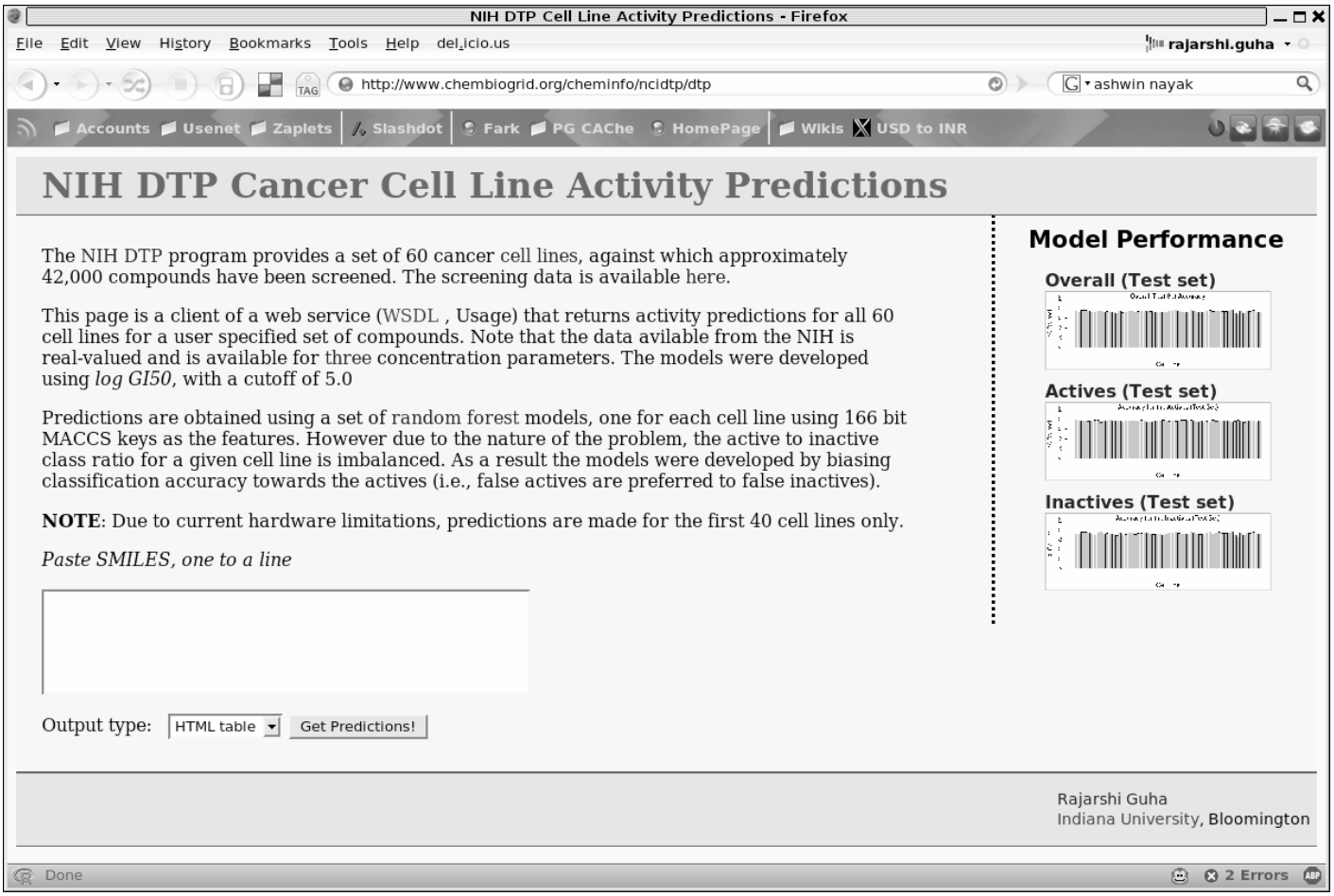

Figure 2. A screen shot of a web page client that provides access to a set of models that predict the anti-cancer activity of a compound, given its SMILES, against the 60 cancer cell lines hosted by the NIH NCI Developmental Therapeutics Program.

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 and 3D similarity.
	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
	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

REFERENCES

[1] Chemical Information Sources Wiki, http://cheminfo.informatics.indiana.edu/cicc/cis (Accessed April 17, 2007)
[2] Knime, http://www.knime.org (Accessed April 17, 2007)
[3] Scitegic Pipeline Pilot, http://www.scitegic.com (Accessed April 17, 2007)
[4] IdeaConsult Ltd., http://ecb.jrc.it/qsar/home.php?CONTENU=/qsar/qsar-tools/qsar_tools_toxtree.php (Accessed Apr 17, 2007)
[5] OpenEye, http://www.eyesopen.com (Accessed April 17, 2007)
[6] UNC Informatics Survey, http://ils.unc.edu/informatics_programs/ (Accessed April 17, 2007)
[7] Obernai Declaration, http://infochim.u-strasbg.fr/chemoinformatics/Obernai_declaration.php (Accessed April 17, 2007)
[8] Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S., On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002,** 42, 903--911.
[9] Alameda, J.; Christie, M.; Fox, G. C.; Futrelle, J.; Gannon, D.; Hategan, M.; Kandaswamy, G.; von Laszewski, G.; Nacar, M. A.; Pierce, M. E.; Roberts, E.; Severance, C.; Thomas, M., The Open Grid Computing Environments collaboration: portlets and services for science gateways. *Published Online: 10 Oct 2006 DOI: 10.1002/cpe.1078. Available from http://www3.interscience.wiley.com/cgi-bin/fulltext/113391287/PDFSTART* **2006**.
[10] Atkins, D. *Report of Blue-Ribbon Advisory Panel on Cyberinfrastructure*; National Science Foundation: 2004.
[11] Atkinson, M. P.; De Roure, D.; Dunlop, A. N.; Fox, G. C.; Henderson, P.; Hey, A. J. G.; Paton, N. W.; Newhouse, S.; Parastatidis, S.; Trefethen, A. E.; Watson, P.; Webber, J., Web Service Grids: an evolutionary approach. . *Concurrency and Computation: Practice and Experience* **2005,** 18, (10), 1009-1019.
[12] Avery, M. A.; Alvim-Gaston, M.; Rodrigues, C. R.; Barreiro, E. J.; Cohen, F. E.; Sabnis, Y. A.; Woolfrey, J. R., Structure Activity Relationships of the Antimalarial Agent Artemisinin. The Development of Predictive in Vitro Potency Models Using CoMFA and HQSAR Methodologies. *J. Med. Chem.* **2002,** 45, 292-303.
[13] Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D., Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2001,** 41, 671-678.
[14] Berman, F.; Fox, G. C.; Hey, T., *Grid Computing: Making the Global Infrastructure a Reality*. John Wiley & Sons: Chichester, England, 2003.
[15] Bohl, M.; Dunbar, J.; Gifford, E. M.; Heritage, T.; Wild, D.; Willett, P.; Wilton, D. J., Scaffold Searching: Automated Identification of Similar Ring Systems for the Design of Combinatorial Libraries. *Quantitative Structure-Actitivy Releationships* **2002,** 21, (6), 590-597.
[16] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J., *Classification and Regression Trees*. Chapman & Hall/CRC: Boca Raton, FL, 1984.
[17] Bremet, A. L., Cyberinfrastructure: the Second Revolution. *The Chronicle of Higher Education* **2007,** 53, (18), B5.
[18] Brown, F., Chemoinformatics - a ten year update. *Current Opinion in Drug Discovery and Development* **2005,** 8, (3), 298-302.
[19] Butina, D.; Gola, J. M. R., Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003,** 43, 837-841.
[20] Cramer, G. M.; Ford, R. A.; Hall, R. L., Estimation of Toxic Hazard--A Decision Tree Approach. *Food. Cosmet. Toxicol.* **1978,** 16, (3), 255--276.
[21] Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D., Lead Hopping. Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004,** 47, 6777--6791.

[22] Dietterich, T. G., The Handbook of Brain Theory and Neural Networks. In Arbib, M. A., Ed. MIT Press: Cambridge, MA, 2002.
[23] Dong, X.; Gilbert, K. E.; Guha, R.; Heiland, R.; Kim, J.; Pierce, M. E.; Fox, G. C.; Wild, D. J., A web service infrastructure for chemoinformatics. *Journal of Chemical Information and Modeling* **2007,** Submitted.
[24] Dutta, D.; Guha, R.; Jurs, P. C.; Chen, T., Scalable Partitioning and Exploration of Chemical Spaces using Geometric Hashing. *J. Chem. Inf. Model.* **2006,** 46, 321--333.
[25] Dutta, D.; Guha, R.; Wild, D.; Chen, T., Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models. *J Chem Inf Model* **2007,** ASAP.
[26] Foster, I.; Kesselman, C., *The Grid 2: Blueprint for a new computing infrastructure*. Morgan Kaufmann: 2004.
[27] Fox, G. C.; Gannon, D., Special Issue: Workflow in Grid Systems. *Concurrency and Computation: Practice and Experience* **2006,** 18, (10), 1009-1019.
[28] Graham, P., Web 2.0. *http://www.paulgraham.com/web20.html (Accessed April 17, 2007)* **2005**.
[29] Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T., R-NN Curves: An Intuitive Approach to Outlier Detection Using a Distance Based Method. *J. Chem. Inf. Model.* **2006,** 46, (4), 1713--1722.
[30] Guha, R.; Jurs, P. C., Determining the Validity of a QSAR model--A Classification Approach. *J. Chem. Inf. Model.* **2005,** 45, (1), 65--73.
[31] Guha, R.; Jurs, P. C., Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance. *J. Chem. Inf. Model.* **2005,** 45, 800-806.
[32] Guha, R.; Jurs, P. C., The Development of Linear, Ensemble and Non-linear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004,** 44, 2179-2189.
[33] Guha, R.; Jurs, P. C., The Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comp. Sci.* **2004,** 44, 1440-1449.
[34] Guha, R.; Stanton, D. T.; Jurs, P. C., Interpreting Computational Neural Network QSAR Models: A Detailed Interpretation of the Weights and Biases. *J. Chem. Inf. Model.* **2005,** 45, 1109-1121.
[35] Halgren, T. A., Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comp. Chem.* **1996,** 17, 490-519.
[36] Hansch, C., A Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1969,** 2, 232-239.
[37] Hawkins, D. M., The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004,** 44, (1), 1--12.
[38] Kohavi, R.; John, G. H., Wrappers for Feature Subset Selection. *Artif. Intell.* **1997,** 97, (1-2), 273--324.
[39] Liu, Y., A Comparative Study on Feature Selection Methods for Drug Discovery. *J. Chem. Inf. Model.* **2004,** 44, (5), 1823--1828.
[40] Martin, Y. C., What Works and What Does Not: Lessons From Experience in a Pharmaceutical Company. *QSAR & Combinatorial Science* **2006,** 25, (12), 1192-1200.
[41] Miller, A., *Subset Selection in Regression*. Chapman & Hall/CRC: Boca Raton, FL., 2002.
[42] Mosier, P. D.; Counterman, A. E.; Jurs, P. C.; Clemmer, D. E., Prediction of Peptide Ion Collision Cross Sections from Topological Molecular Structure and Amino Acid Parameters. *Anal. Chem.* **2002,** 74, 1460-1370.
[43] Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L., Chemical markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comput. Sci.* **2004,** 44, (2), 462--469.
[44] Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C., Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *ATALA* **2005,** 33, 155--173.
[45] Spjuth, O.; Helmus, T.; Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. S., Bioclipse: An Open Source Workbench for Chemo- and Bioinformatics.

BMC Bioinformatics **2007,** 8, 59.
[46] Stanton, D. T., On The Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput Sci.* **2003,** 43, (5), 1423-1433.
[47] Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L., Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006,** 12, 2110-2120.
[48] Team, R. D. C., *A language and environmnet for statistical computing.* . Foundation for Statistical Computing: Vienna, Austria, 2006.
[49] Todeschini, R.; Consonni, V.; Pavan, M., DRAGON. In 2005.
[50] Verhaar, H. J.; SolbÈ, J.; Speksnijder, J.; van Leeuwen, C. J.; Hermens, J. L., Classifying Environmental Pollutants: Part 3. External Validation of the Classification System. *Chemosphere* **2000,** 40, (8), 875--883.
[51] Walker, J. D.; Gerner, I.; Hulzebos, E.; Shlegel, K., The Skin Irritation Corrosion Rules Estimation Tool (SICRET). *QSAR. Comb. Sci.* **2005,** 24, 378--384.
[52] Wild, D. J., Strategies for Using Information Effectively in Early-Stage Drug Discovery In *Computer Applications in Pharmaceutical Research and Development*, Ekins, S., Ed. John Wiley & Sons: Hoboken, NJ, 2006; pp 229-246.
[53] Wild, D. J.; Wiggins, G. D., Challenges for Chemoinformatics Education in Drug Discovery. *Drug Discovery Today* **2006,** 11, (9/10), 436-439.
[54] Wild, D. J.; Wiggins, G. D., Videoconferencing and Other Distance Education Techniques in Chemoinformatics Teaching and Research at Indiana University. *Journal of Chemical Information and Modeling* **2006,** 46, (2), 495-502.
[55] Zhang, X.; Shedden, K.; Rosania, G. R., A Cell-Based Molecular Transport Simulator for Pharmacokinetic Prediction and Cheminformatic Exploration. *Molecular Pharmaceutics* **2006,** 3, (6), 704-716.