

Chemical Informatics and Cyberinfrastructure Collaboratory

A project funded by the

National Institutes of Health

under the

NIH Roadmap Molecular Libraries Initiative

for

Exploratory Centers (P20) for Cheminformatics Research

NIH Grant No. 1 P20 HG003894-01 (Fox)

Geoffrey C. Fox, Principal Investigator

October 1, 2005 – September 30, 2007

Chemical Informatics and Cyberinfrastructure Collaboratory

A. Executive Summary

Chemical and life science research is capable of generating terabytes of data each day. Answers to health-related problems are buried in the data, and the computer techniques of informatics can help unearth sound medical solutions for the benefit of all. In conjunction with other units of Indiana University and external advisors, the School of Informatics and the Pervasive Technology Laboratories at Indiana University will use a grant from the National Institutes of Health to establish an exploratory center for cheminformatics research, the Chemical Informatics and Cyberinfrastructure Collaboratory (CICC). Over the next two years, the CICC will investigate novel approaches to assist in the discovery of new chemical compounds.

Using powerful computer simulation and visualization environments, an integrated Chemical Informatics Cyberinfrastructure based on modern distributed service architectures will be constructed. The project will utilize the emerging high-capacity computer networks, powerful data repositories, and computers that comprise the Grid. This will ensure scalability, computational efficiency, and interoperability among heterogeneous components and will lead to the development of specific data-mining tools and applications for chemistry. The tools and infrastructure will be tested in real-life industry and academic science situations, so the approach can be assessed by practicing scientists, ultimately contributing to the design of a new kind of integrated discovery environment that will vastly speed up scientific research.

To develop generally applicable and widely usable technologies, a team of researchers and educators from Informatics, Chemistry, Biology, and Computer Science has been assembled at IU. They will be assisted by a group of external advisors from the pharmaceutical, software, and high-throughput screening industries, as well as experts from other academic institutions. The participants will closely collaborate to work on three major areas:

- Adapting and further developing Web Services and other grid technology for chemistry research
- Applying the computer techniques to real chemical research problems with the potential to aid in the treatment of diseases such as Alzheimer's Disease and cancer
- Improving the educational efforts in cheminformatics to incorporate the lessons learned in the project.

Specific Aims

Developing ways to use the High Throughput Screening data from public databases, especially the data flowing into PubChem, the NIH's small molecule database of publicly available chemical and biological data, will be the main target of one of the experiments. In addition, in the early stages of the project, a prioritized list of critical chemical informatics challenges will be generated. These will form the basis of new academic courses. Eventually, an application for a full NIH-funded Center for Cheminformatics Research (CCR) will be submitted. The Indiana University School of Informatics aims to establish itself as an international forum for the exchange of ideas and focused conferences in chemical informatics technology and applied chemical research fields that utilize chemical informatics tools.

I. Adapting and further developing Web Services and other grid technology

The ultimate goal is to devise an integrated Cyberinfrastructure composed of diverse and easily expandable databases, simulation engines, and discovery tools that fully exploit PubChem. Requirements for implementing common service and data interface specifications that will allow flexible customization of the discovery environment to meet the needs of a broad range of chemical research projects will be investigated in three specific prototyping activities:

- **Design of a Grid-based distributed data architecture** compatible with PubChem and other databases and capable of good performance with high volume data
- **Development of tools for HTS data analysis and virtual screening**, integration of data with both parallel and distributed simulation engines, metadata and annotation generation, navigation, and visualization
- **Design of a novel database for quantum mechanical simulation data that will complement the experimental libraries.**

II. Applying the computer techniques to real chemical research problems

Practicing chemists will use and assess the tools in their investigations of state-of-the-art academic and industry challenges in:

- **Novel routes to the discovery of enzymatic reaction mechanisms**
- **Mechanism-based drug design**
- **Data-inquiry-based development of new methods in natural product synthesis.**

III. Cheminformatics education

There is a need for a new generation of scientists who can integrate modern methods of information technology and fundamental chemical expertise into the solution of complex life science problems. The Indiana University School of Informatics Masters and Ph.D. programs, supported by distance education courses, will produce such scientists. Further development of a curriculum that establishes the educational framework for multidisciplinary graduate degrees and a certificate in chemical informatics will be a natural outcome of the CICC projects.

Personnel

INDIANA UNIVERSITY

Geoffrey C. Fox (PI)	Informatics/Computer Science/Physics/Community Grids Laboratory
Mu-Hyun (Mookie) Baik	Informatics/Chemistry
Randall B. Bramley	Computer Science
Kenneth G. Caulton	Chemistry
Peter T. Cherbas	Biology/Center for Genomics and Bioinformatics
Mehmet (Memo) M. Dalkilic	Informatics
Charles H. Davis	School of Library and Information Science
Richard D. DiMarchi	Chemistry
A. Keith Dunker	Informatics/Medicine/Center for Comp. Biology and Bioinformatics (IUPUI)
P. Andrew Evans	Chemistry
Kelsey Forsythe	Informatics/Chemistry/Computational Molecular Science Facility (IUPUI)
Dennis B. Gannon	Computer Science/Pervasive Technology Laboratories
John C. Huffman	Informatics/Chemistry
Jeffrey N. Johnston	Chemistry
Malika Mahoui	Informatics (IUPUI)
Daniel J. Mindiola	Chemistry
Marlon Pierce	Community Grids Laboratory
Beth A. Plale	Computer Science
Yvonne Rogers	Informatics/School of Library and Information Science
Santiago D. Schnell	Informatics/Biocomplexity Institute
Craig A. Stewart	University Information Technology Services
Gary D. Wiggins	Informatics
David J. Wild	Informatics

EXTERNAL ADVISORS

Dimitris K. Agrafiotis	Johnson & Johnson Pharmaceutical Research & Development
John M. Barnard	Digital Chemistry Ltd.
James M. Caruthers	Purdue University School of Chemical Engineering
Jeremy G. Frey	University of Southampton Department of Chemistry
Val Gillet	University of Sheffield Department of Information Studies
Horst Hemmerle	Lilly Research Laboratories
Stephen J. Lippard	MIT Department of Chemistry
Andrew Martin	Kalypsys Inc.
John McKearn	Kalypsys Inc.
Peter Murray-Rust	Cambridge University Unilever Centre for Molecular Informatics
Martin E. Newcomb	University of Illinois at Chicago Department of Chemistry
Alan D. Palkowitz	Lilly Research Laboratories
John V W Reynders	Lilly Research Laboratories
David Spellmeyer	IBM Almaden Services Research
Peter Willett	University of Sheffield Department of Information Studies

B. Background and Significance

B.0. Introduction

a. Vision. Two distinctively different uses of information and computer technology can revolutionize biomedical research. The utilization of databases to enable rapid access to libraries of experimentally classified molecules is a proven use of information technology. **Virtual Screening** of large libraries is an efficient tool of drug discovery, mostly utilized in commercial research environments. NIH's initiative to make available a public databank with pharmacologically relevant properties of small molecules in PubChem will naturally allow similar research to be carried out in academia and small companies. To facilitate the successful use of PubChem in the targeted research environments, novel tools and application protocols must be developed, which will be a main research thrust in our future work. Computers can also serve as **Virtual Laboratories**, where computer models of biologically relevant molecules that obey quantum or Newtonian mechanical laws allow for a new type of rational discovery. The advent of high-performance computers and efficient methods in theoretical chemistry made it possible to construct remarkably accurate virtual models of complicated chemical processes that provide valuable scientific insights at an unprecedented level of detail and accuracy.

Our vision for the CCR at Indiana University is to combine access to the experimental libraries PubChem, PDB, Reciprocal Net (*vide infra*), and a new molecular modeling database that we propose to develop. (Fig. 1) We will build a novel discovery environment for biomedical research, a **Chemistry Virtual Laboratory (CVL)** that provides an unprecedented level of diversity of potentially interesting small molecules, combined with rigorous modeling results from theoretical chemistry to allow for a high-level analysis of bonding, chemical behavior, and reaction mechanisms. This combined effort of chemistry and informatics will enable a new level of depth to our understanding of reaction mechanisms and form the foundation for truly innovative approaches to discovery and optimization of small molecule drugs and other biologically active agents.

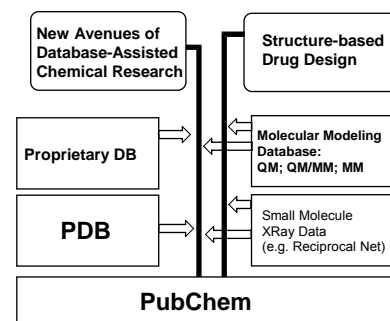


Figure 1. DB Scheme

b. Chemical Informatics as a Tool for New Pathways to Discovery. The foundation of our Exploratory Center for Cheminformatics Research is a highly collaborative network of Chemists, Biologists, Informaticians, and Computer Scientists. This network already exists at Indiana University, and we intend to strengthen it both by adding new connections and by taking the existing collaborative infrastructure to the next level. It is our view that if Chemical Informatics as a new discipline is to succeed the following characteristics are critical:

- (i) The design of Cheminformatics tools must be **conceptually broad**, but facilitate integration with research data on specific chemical problems to enable **immediate relevance** to current chemical research.
- (ii) **Experimental chemists** must be involved at an early stage. They will not be merely the executioners of predictions of virtual screening, but will help to iteratively improve computational tools and theoretical concepts by providing important benchmarks and identifying the most pressing bottlenecks in research.
- (iii) The cyberinfrastructure must be designed at a **high level of technical expertise and competence**, taking full advantage of modern technologies such as GRID-based computing and distributed databases.¹⁻³
- (iv) The architecture of the virtual discovery environment must be **highly compatible**. NIH's PubChem initiative provides a much needed standard that we intend to build on.
- (v) A cohesive administrative framework is needed to facilitate **efficient communication** in the Center. IU provides such framework in the multi-campus School of Informatics, the first informatics school in the US.⁴

B.1. Technology Development

a. Design of a Grid-based distributed data architecture. A data deluge is revolutionizing many fields of science which are also becoming increasingly multidisciplinary and pursued by international distributed teams of researchers⁵ This has motivated the concepts of Cyberinfrastructure in the USA and e-Science or e-Infrastructure in Europe.⁶⁻⁸ Grids are built from web services that are an integrating technology which is growing in importance for the life sciences. The role of grids in this project has several facets. In the context of PubChem, one can build a multi-tiered architecture that allows bio- and cheminformatics applications to run

together in a complex workflow linking data and simulation, while keeping the integration process invisible to end-users.

We interpret the concept of grids broadly and use it to describe the technology that implements what NSF calls Cyberinfrastructure and the UK calls e-Science. In this view grids support virtual organizations enabling distributed science with shared managed computing and data resources. As such, grids are our approach to integrate the different activities in the exploratory center and enable the international collaboration to ensure our work is best practice. In the UK e-Science embraces the view that data is as important as compute power, if not more important. Without data, there is nothing useful on which to do computing. We will integrate compute resources with archived and real-time data. We stress that we are not developing any core grid technology in this proposal but rather developing an architecture that effectively applies to Cheminformatics appropriate grid technologies that are already developed. The highlighted core architecture activity in the proposal is enhanced by an expert on grids working within each Cheminformatics project.

This project does not aim to develop new grid technology but rather to exploit current best-practice. In this sense our proposed grids are not aimed at innovation. Indeed, the successful use of grids by caBIG and BIRN for NIH, CMCS (Chemistry) for DOE, Comb-e-Chem (Chemistry) and myGrid (Bioinformatics) in the UK are excellent examples on which to build.⁹⁻¹⁴ Other large projects demonstrating the importance of Grids include EGEE in Europe and the Open Science Grid in NSF.^{14b-c} NSF's interest in cyberinfrastructure is emphasized by the creation of a new office devoted to this and reporting to the NSF Director. In the digital library area, the next release of science.gov (the inter agency technical information service mentored by CENDI) will be built on grid technology, and a recent National Archives conference "Partnerships in Innovation: Serving a Networked Nation" had a grid focus. The use of grids in Cheminformatics, although building on examples cited above, is quite innovative, but importantly it is a useful idea.

As shown in Figure 2, Cyberinfrastructure links together distributed sensors, instruments, data repositories, computer clusters, high performance systems, and researchers. Grids provide the technology for Cyberinfrastructure by extending existing Web Service Internet systems to provide managed secure sharing of resources. Grids support a variety of research models, data analysis services, high performance simulations and what are called desktop Grids. The latter involve distributed simulation of multiple instances either on clusters (intra-enterprise) or even across the Internet. This model, popularized by SETI@Home, is applied in financial modeling, climate prediction, and in CVL applications. The grid concept has recently been extended to docking and virtual screening.^{15,16} Specifically, Internet-connected computers participated in the large-scale docking of potential drug candidates against a variety of targets. In one Anthrax project, 300,000 potential candidates were found out of 3.57 billion compounds screened against Anthrax Protective Antigen (PA). The project was completed in just 24 days.¹⁶ Another distributed computing project is Folding@Home which studies protein folding, misfolding, aggregation, and related diseases.¹⁷

Many of the core ideas in Grids are still being debated at various levels – distributed objects versus services – WSRF versus WS-I+ and the software infrastructure such as Globus and the OMII is still evolving.¹⁸⁻²² However, the essential ideas are clear, and one builds a system of distributed resources and services communicating by messages. The services have interfaces defined in Web Services Description Language (WSDL), and the message semantics are defined by the SOAP standard.^{23,24}

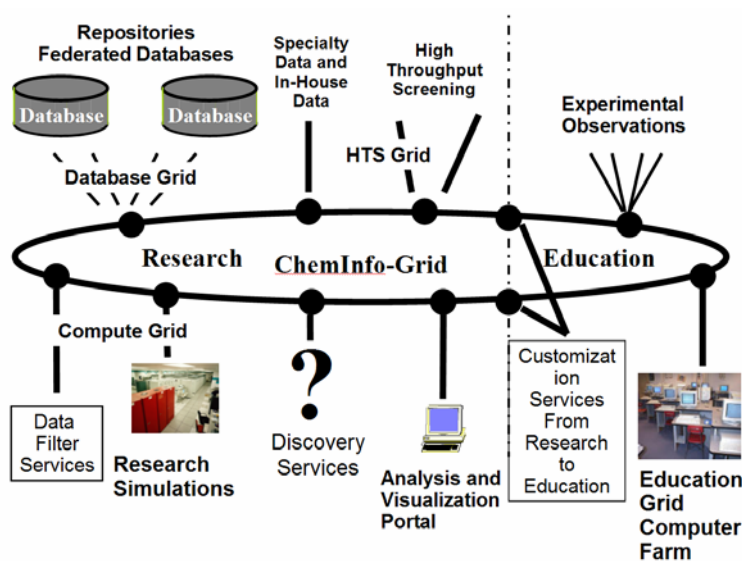


Figure 2. Proposed Cheminformatics Grid

At the lowest level, there is a suite of

specifications commonly called the WS-*, and these include areas like addressing, security, reliable messaging, notification, transactions, workflow, and state. The Global Grid Forum (GGF) is defining higher level specifications covering core data and compute functions with the OGSA-DAI database specification, notable as having already an excellent implementation used successfully, for example, in the cancer Biomedical Informatics Grid caBIG.²⁵⁻²⁷

Indiana University has world leading expertise in Grids with Reciprocal Net exemplifying our Virtual Laboratory for crystallography and the computer science team having broad experience in both developing core Grid technology and applying it to major science areas including earthquake, fusion, and tornado applications. Fox, Gannon, and Plale all are playing major roles in the Global Grid Forum. In particular, we are co-leaders of the Grid portal activities at GGF, and here the portlet and WSRP standards are defining re-usable user interface components which we will use in this project.^{28,29} We believe that a Web Service-based Grid architecture will answer our CCR challenges in high volume data acquisition, processing, and management at the envisioned size and scope. Further, this approach will straightforwardly allow us to integrate our Chemical Informatics Cyberinfrastructure with Grids from other projects and other fields. WS-* and OGSA define the generic services and data standards we need, but this must be supplemented in any e-Science application by domain-specific specifications for both services and data. Here we will build on CML (Chemical Markup Language),³⁰ but this needs to be extended especially in adding the chemistry specific WSDL for services.

b. Development of tools for HTS data analysis and virtual screening. High throughput screening (HTS) is a major technical component in modern drug discovery. Recent *Ultra-high throughput screening (UHTS)* machines permit the screening of 1,000,000 compounds in a single day. The startup and experimental expenses have thus far hampered the impact of HTS on academia and government, an issue which is being boldly tackled by the NIH Molecular Libraries & Imaging Initiative (MLI).³¹ Besides the NCI dataset, recently joined by ZINC,³² there has been very little chemical data freely available to the public until the advent of the MLI and PubChem.³³ In contrast to the relative paucity of public screening results, the explosion of data volume in the pharmaceutical industry has created a “data overload,” which companies have attempted to address with new tools, procedures, and techniques for organizing, navigating, and analyzing HTS results.³⁴⁻³⁷ For example, Bikker et al. introduce a multi-stage HTS analysis process that is evolving at Pfizer, designed to allow chemists to intelligently select a small number of series suitable for medicinal chemistry follow-up.³⁸

Despite the lack of publicly available data, significant research has been carried out in academia, particularly on basic algorithms and methods that can be used for processing and analyzing the screening results. Downs and Barnard chart the evolution of unsupervised cluster analysis methods applied to the organization and analysis of increasingly large (and recently HTS) chemical datasets, including new methods coming out of the data mining community.³⁹ Supervised learning techniques have been used for classifying and discerning structure-activity relationships in HTS data, including recursive partitioning, support vector machines and binary discrimination.⁴⁰⁻⁴² The problem of visualizing HTS datapoints in chemical space has been addressed using nonlinear mapping and Kohonen self-organizing maps.⁴³⁻⁴⁵ Self-organizing maps are used as a visualization tool for the NCI dataset. The filtering of HTS datapoints for drug-likeness has also been investigated.⁴⁶

A small number of companies produce tools specifically aimed at the analysis of HTS data. The main tools available are Spotfire DecisionSite, Tripos SAR Navigator, Bioreason DrugPharmer, Sage Informatics ChemTK, and PowerMV from the National Institute of Statistical Sciences.⁴⁷⁻⁵² With the exception of PowerMV, these are commercial products targeted at the pharmaceutical industry. DecisionSite, SAR Navigator and DrugPharmer were developed through collaborations with pharmaceutical companies (two of which were led on the industry side by one of the investigators for this proposal, Wild). In addition to handling the organization, navigation, and analysis of the screening data, a tool must allow relevant ancillary information to be included in the decision-making processes. Success of a screen in producing suitable drug candidates has been shown to be heavily dependent on the quality of compounds being screened and the information that is taken into consideration to select candidates for follow-up.⁵³ For example, poor solubility, permeability and metabolic stability of compounds are major sources of attrition in early drug development.⁵⁴ These properties must be included in the selection of compounds to minimize the failure rate in the drug discovery process. Further, as more screens are carried out and high-throughput genomic assays become feasible, it is crucial to consider multiple targets simultaneously.

Before any of this processing can be carried out, HTS data needs to be stored in a database with an appropriate architecture and appropriate data. For example, data without an adequate description of the assay protocols will quickly become meaningless; inefficient data retrieval and access performance will at best cause frustration and at worst prevent the use altogether. 2D structure databases pre-clustered in series can vastly speed up processing. An effective system for processing HTS screening results requires a well-thought-out database architecture with a straightforward method of data acquisition, easy access, a flexible set of tools for working with the large volumes of data and good cross-referencing with related data. We are encouraged by recent advances in the development of publicly-accessible screening repositories, collaboration architectures and portals, particularly the PubChem database, the ZINC database, the NCI browser (Frederick/Bethesda), the Collaboratory for Multi-Scale Chemical Science, and the Combechem e-science project.⁵⁵⁻⁵⁹ Our efforts will build on this work and will be particularly focused on the development of efficient tools for organizing, navigating, and analyzing HTS screening results and associated data that can be provided through these architectures.

c. Design of a novel database for quantum mechanical simulation data. To leverage the recent revolution in theoretical chemistry⁶⁰ and enzyme modeling,⁶¹ we propose to implement a library of computer simulation data into the discovery environment. This extension will allow for research at a notably higher degree of physical reason and scientific rigor. We believe that a library of modeling data is a powerful complement to the high volume experimental data that will likely facilitate the acceptance of cheminformatics approaches in academic research and catalyze highly innovative discoveries.

B II. Cheminformatics Education

The Indiana University School of Informatics has dynamic, innovative programs in many areas of science informatics, including bioinformatics and chemical informatics at both the Bloomington and Indianapolis campuses (IUB and IUPUI). In addition, IUPUI has MS graduate programs in laboratory informatics and health informatics and a developing medical informatics program, plus a BS undergraduate program in Health Information Administration. IUB has added several faculty in complex systems (systems biology), and a Ph.D. track is being formulated in that area. The Department of Computer Science has formally joined the School of Informatics at IUB, and twelve of Fox's Ph.D. students are now affiliated with the School of Informatics.

An article on the birth of chemical informatics appeared in *Nature* in 2002. The article notes that just as the genomics boom caused a great need for bioinformaticians, "an explosion in the amount of data generated by combinatorial chemistry and other high-throughput approaches to drug screening and drug design is creating a demand for cheminformaticians."⁶² The Indiana University MS program in Chemical Informatics is one of only three graduate programs in the world that are mentioned in the article, and the only one in the US.^{63,64} Since 2002, the School of Informatics has moved rapidly to recruit additional faculty on both the Bloomington and Indianapolis campuses, to expand its programs to areas that include laboratory informatics and systems biology (complex systems), and to initiate a Ph.D. program in Informatics.⁶⁵ It is essential that this country have at least one major research center in an academic setting where students can pursue advanced degrees in the emerging discipline of cheminformatics, a place where teaching modules can be developed so that others can more easily adopt these techniques. Emerging fields like cheminformatics naturally benefit from distance education as a critical mass of both faculty and students can be assembled by linking distributed sites. Indiana University is a leader in this area, with several installed Access Grid and Polycom sites, a major research project GlobalMMCS led by Fox, and the university-wide Sakai project, blending the e-learning tools of four major universities (IU, Michigan, MIT, Stanford).⁶⁶⁻⁶⁸ Fig. 2 notes that Grid architecture is a powerful way of linking education to research resources.

We expect to see more Ph.D.-level courses related to the Grid and e-Science, such as:

- e-Science, e-Business, e-Government and Their Technologies (Geoffrey Fox)
- E-Science (Beth Plale).

At IUB, the School of Informatics has opportunities for training in the specialized areas of chemical informatics and bioinformatics for both undergraduate and graduate students, as well as for people outside the academy. We offer the BS in Informatics with a cognate in chemistry or biology, the MS in Chemical Informatics, the MS in Bioinformatics, and the Ph.D. degree with tracks in both bio- and chemical informatics. The new Ph.D. program is designed to acquaint students with the human, technical, and subject domain aspects of informatics. Key courses in the science informatics programs include:

- I571 Chemical Information Technology (3 cr. hrs.) (Wild)
- I572 Computational Chemistry and Molecular Modeling (3 cr. hrs.) (Baik)
- I533 Seminar in Chemical Informatics (1-3 cr. hrs.)
- I553 Independent Study in Chemical Informatics (1-3 cr. hrs.)
- L519 Bioinformatics: Theory and Application (3 cr. hrs.) (Dalkilic, Kim)
- L529 Bioinformatics in Molecular Biology and Genetics: Practical Applications (4 cr. hrs.) (Kim)
- I532 Seminar in Bioinformatics (1-3 cr. hrs.)
- I552 Independent Study in Bioinformatics (1-3 cr. hrs.)
- L504 Genome Biology for Physical Scientists (3 cr. hrs.) (S. Jakobsson)
- I510 Data Acquisition and Laboratory Automation (3 cr. hrs.) (Merchant)
- I511 Laboratory Information Management Systems (3 cr. hrs.) (Perry)
- I512 Scientific Data Management and Analysis (3 cr. hrs.) (Merchant)
- I590 Topics in Informatics (offered Spring or Fall 2005)
 - Pervasive Computing (Rogers)
 - The Simplicity of Complexity (Vespignani, Flammini)
 - Biologically Inspired Computing (Rocha)
 - Artificial Life as an Approach to Artificial Intelligence (Yaeger)
 - Structural Bioinformatics (Bolshoy)
 - Programming for Chemical Informatics (Wild)
 - Information Theory (Ortoleva)
 - Information Retrieval from Chemistry and Life Sciences Databases (Wiggins).

B.III. Chemical Prototype Projects

We have begun using a preliminary implementation of the database-assisted discovery network to make progress in a number of projects. We intend to utilize these ongoing projects as prototype applications that will illustrate our chemical informatics approach to biomedical discovery:

a. Discovery of enzymatic reaction mechanisms. In collaboration with the experimental group of Professor Martin Newcomb (UI-Chicago) and Professor Stephen J. Lippard, we will continue to use chemical informatics tools to study the mechanism of radical rearrangements occurring in coenzyme B₁₂-dependent enzymes⁶⁹ and to further understand the mode of action of the bacterial metalloenzyme Methane Monooxygenase.⁷⁰

b. Mechanism-based drug discovery. A second long-term collaboration with the Lippard lab centers on elucidating the details of the mode of action of the potent anti-cancer drug cisplatin.⁷¹⁻⁷³ The utilization of a newly developed chemical informatics tool for the analysis of a few thousand electronic structure calculation results has given new insights to how cisplatin interacts with its primary cellular target, genomic DNA. In a different project we will use the new tools developed in our lab to study the mechanism of hydrogen peroxide formation mediated by Cu- β -Amyloid complexes, suspected to form in neuritic plaques that ultimately give rise to cortical atrophy in Alzheimer's disease (AD)⁷⁴⁻⁷⁶.

c. Application of Chemical Informatics to Natural Product Synthesis. A very new and exciting direction of using chemical informatics tools is the database-assisted design of new transition metal based catalysts that can assemble carbo- and heterocycles with high regio- and stereoselectivity in asymmetric synthesis of natural products. This research will be conducted in collaboration with the organic chemistry groups of Professors Evans, Williams and Johnston, and the experimental organometallic groups of Professors Mindiola and Caulton (all IU faculty members).

Overall, the goals that we have identified for the Chemical Informatics and Cyberinfrastructure Collaboratory are very ambitious and constitute a highly innovative approach to utilizing databases in diverse areas of biomedical research. We seek to find a balance between high-risk endeavors that promise to change the ways chemical research is done today and work that is feasible and immediately beneficial. We believe that the grand challenge of revolutionizing biomedical research can only be met with bold steps of innovation as we propose herein. Nevertheless, we recognize distinctive differences between commercial drug discovery processes and academic biomedical research goals. Whereas research in the pharmaceutical industry is fueled by the desire to discover a marketable product, academic research is generally broader and aims to

discover new paradigms of understanding and treating diseases. Simply making a high volume database available is unlikely to inspire large-scale acceptance from the academic community, as only a few research groups will abandon their traditional ways of doing research and move to a style that closely resembles commercial research and product development. If the integration of cheminformatics into academic research is to succeed, we must design our cheminformatics cyberinfrastructure such that it immediately enhances research productivity without requiring a complete revision of how chemical and biomedical research is done in academic environments.

We have clear ideas for the eventual full NIH Cheminformatics Research Center (CRC), which are focused around developing and implementing useful tools that chemical and life scientists can utilize at the workbench. However, since the integration of bio- and cheminformatics is currently at such an early stage, the details will need to be confirmed by the early research to be carried out in the projects. By convening a meeting of recognized chemical informatics expert advisors early on in the project we seek to identify the key components of a CRC that we would build toward throughout the two years of this project and thus slant our efforts in those directions.

C. Preliminary Studies and Related Developments

Indiana University's School of Informatics faculty have considerable expertise in metadata, data mining, visualization, networks, grid computing, scientific computing, related core technologies, and a bridge to chemistry in order to develop the required tools for dealing with massive data sets and for enhancing scientific collaboration. We can draw on our existing faculty in science informatics (bioinformatics, chemical informatics, systems biology at Bloomington and our Indianapolis Informatics colleagues in bioinformatics, chemical informatics, medical and health informatics) plus human-computer interaction faculty to provide tools for identifying significant patterns in the massive data flowing from the scientific research in the life sciences and chemistry areas.

C.0. Introduction

Grid technologies provide the generic foundation on which we may build distributed computing systems to support scientific enterprises. These systems must couple distributed data sources, computing resources, data archiving systems, and visualization systems. Collaboration, both by sharing data and by audio/video services, is part of the Grid technology substrate. Indiana University has extensive experience applying Grid technology to problems in scientific computing, as described below. The IU Department of Computer Science and the Community Grids Laboratory build these specific Grid application systems on a generic foundation. Proposal team members' experience includes development of such core tools as Web Service-based Grid technologies, topic-based publish/subscribe systems for distributed messaging, sensor Web services, streaming data management, computational Web portals for providing ubiquitous Grid access, audio/video Web Services, and scientific workflow systems.

Such research must take place in the larger Grid computing community, the Global Grid Forum, or GGF. This body provides both research tracks for academic partners and standards-making tracks for those with commercial interests. Fox and Gannon of the proposal team serve on the GGF's steering committee and thus help shape the directions of the Grid community as a whole.

C.I. Technology Development

a. Design of a Grid-based distributed data architecture.

1. Experience Building Reciprocal Net. The Reciprocal Net project at IU has resulted in a distributed, open, extensible digital collection of molecular structures. Hundreds of academic crystallography laboratories in the U.S. and the world each determine hundreds of crystal structures every year. Most of these structure determinations are subsidized or even fully paid by public funds, yet because of the realities of modern scientific publishing practices, many of the crystallographic results are never disseminated beyond the research groups involved. The Reciprocal Net project recognizes the value of these results to the larger scientific community, as well as the propriety of making them available to the taxpaying public. The project tackles the dissemination issue at the root, by establishing a distributed database of molecular structure

information with local nodes at each participating x-ray laboratory. The local face of each node provides a laboratory information management system for the laboratory, so as to facilitate its incorporation into laboratory procedures. The database as a whole provides both local and global search facilities, and exposes an OAI-PMH compliant interface for metadata harvesting (with one major client in the NSF's National Science Digital Library). Software tools for visualizing, interacting with, and rendering printable images of the contents of the files have been developed, as has software for the automated conversion of local database representations into standard formats which can be shared globally. As part of the NSF's National Science Digital Library, tools and components for constructing educational modules based on the collection were created, and a sample collection of common molecules forms the basis of a public repository for educational materials. Reciprocal Net is built primarily from structures contributed by participating crystallography laboratories (nearing 20 participants), and the emphasis to date has been on obtaining structures of general interest and usefulness. Included are Java applets for interactively visualizing the structure (rotating it, scaling it, translating it, and viewing it in multiple drawing modes with one monographic and two stereographic options), analyzing it (measuring interatomic distances and angles), and obtaining high-quality rendered images from supporting servers. A search facility based on a sophisticated SQL database of structure metadata is included. Metadata harvesting allows the server at IU to build the database from the collections of the participating organizations. Bramley was a key participant in the development of the Reciprocal Net.

2. Experience Building LEAD (Linked Environments for Atmospheric Discovery)⁷⁷ Tornadoes, hurricanes and other mesoscale weather events exact a large cost in terms of human life and property damage every year. Our current ability to predict these events is not very great. However, recent advances in networking, large-scale simulation capability and radars are about to change all of this. The NSF "Linked Environments for Atmospheric Discovery" (LEAD) project headed by Kelvin Droegemeier of the University of Oklahoma is building a distributed collaboratory to design and test adaptive workflows that will link real-time observations directly with large supercomputer simulations that focus computational power adaptively on very localized regions where important weather events seem likely to occur. Eventually these simulations will be able to dynamically retarget the radars to provide the most accurate data possible back into the simulations. The goal is to use the resources of the computing and instrument Grid well enough that we can predict the exact location of tornadoes well before they happen. A central component of this is a portal and a specialized database called MyLEAD, which allows the researcher to analyze and catalog and search metadata about the weather and experiments, compose experimental and production workflows, launch simulations, and track the progress of computations. Plale leads the design of the MyLEAD system and Gannon is in charge of the workflow and the web services architecture.

3. Experience Building SERVGrid (Solid Earth Earth Research Virtual Observatory Grid)⁷⁸ The Community Grid Laboratory's Fox and Pierce have experience in applying Grid technology to geophysical problems. They are the designers and developers of the Grid infrastructure for the QuakeSim project, which supports earthquake modeling and forecasting.⁷⁹ This work includes the development of a Web Service-based Grid system for managing distributed data and computing resources. Users access these services through the QuakeSim Portal, a component-based Web browser portal system. Users may create and edit projects that allow jobs to be submitted through their browsers. This portal system relies on several backend services, including a) a Web Service interface to the QuakeTables fault database; b) file management services for data transfer between the user's desktop and backend resources; c) workflow management tools for sequencing interdependent applications; and d) project metadata management tools for storing information about user portal activities. QuakeSim applications include finite element method mesh generators and solvers, time series data analysis codes, and large scale interacting fault model codes.

More recent activities have focused on building standard data grid services using Geospatial Information Systems specifications. We are designing and implementing Web Service-compatible versions of the following Open Geospatial Consortium standards: Web Feature Service software provides access to abstract data models for both natural and man-made Earth surface features; the Web Map Service interacts with Web Feature Services to construct overlay maps; and the Sensor Web family of specifications implements a common interface to time-dependent data series. By adopting Web Service approaches, these services may be directly integrated with the Grid services for code execution and file management. Finally, we have recently completed implementation of a UDDI-based metadata registry system that supports queries against standard GIS service metadata.

SERVOGrid work is built on a reusable foundation: the general purpose Grid services for managing problem archives, moving data, and running applications can be applied to problems in Chemistry Grids and Chemical Informatics. The GIS work, while not directly portable to chemistry problems, serves as an exemplary data grid system: a standard XML data model (the Geographic Markup Language) is accessed and manipulated through a well-defined general set of GIS services. We envision a Chemical Information System that is built on similar principles, with CML serving as the core data model.

4. IU's Computational Environment. Indiana University offers diverse hardware architecture platforms, both in large-scale distributed memory systems as well as large-memory SMP systems. In addition, the AVIDD (Analysis and Visualization of Instrument-Driven Data) Linux clusters include 10 TB of disk storage space and an Itanium-based component plus its large-capacity Pentium IV component. AVIDD comprises a suite of four distributed Linux clusters. These clusters are located 53 miles apart, one in Indianapolis and the other in Bloomington. The Research Database Complex (RDC) is a recent acquisition of midrange Sun systems which serve research database needs, providing a large-memory SMP environment, Oracle software, and large disk storage. It is used by IU's leading database researchers and serves several large biomedical and GIS datasets.

IU's massive data storage system (MDSS) utilizes the High Performance Storage System (HPSS) software to make available to IU researchers a total storage capacity of roughly 1.6 petabytes. A hierarchical storage management (HSM) system by design, HPSS uses a hierarchy of storage media transparently to provide massive, near-line data storage capacity. Users can access data over the network from central research hosts and from personal workstations. Data stored in the MDSS are mirrored between IU Bloomington and IUPUI over Indiana's high performance I-Light network to protect against disasters. Individual researchers are allowed to store up to 500 gigabytes of mirrored data without charge. Usage beyond 500GB is available on a cost-share basis. IU now has redundant, distributed tape storage in Bloomington and Indianapolis connected via the I-Light network. Data written to IU's HPSS system is copied simultaneously to STK tape robots located at both campuses, providing highly reliable disaster protection. This is critical for life sciences data, which is irreplaceable if lost.

Indiana University's Advanced Visualization Lab (AVL) has important development programs in virtual reality technology, particularly in constrained navigation and visualization in the life sciences and the arts. The AVL provides expert consulting, research support, educational opportunities, and hardware and software resources for scientific visualization, virtual reality, high-end computer graphics, and visual tele-collaboration. Resources include high-end graphics workstations and clusters; projection-based virtual reality devices; very high-resolution displays; special-purpose, volume-rendering hardware; haptic display devices; and tools for tele-conferencing and tele-collaboration.

b. Development of tools for HTS data analysis and virtual screening.

1. Rapid dataset organization using cluster analysis techniques. Numerous studies have been done on the use of cluster analysis techniques for the organization of chemical datasets by 2D structure, the consensus being that Ward's clustering produces the most accurate and chemically meaningful groups, at the expense of time (the method is at least $O(N^2)$).⁸⁰⁻⁸² A new method, Divisive K-means, developed by Barnard Chemical Information, has been shown by Wild to be of comparable quality with Ward's, but with drastically reduced runtimes.^{83,84} Further, the Divisive K-means algorithm has been parallelized for Linux clusters. Experiments we have carried out using small parallel Linux clusters indicate that parallelism works very effectively for speeding up clustering. For example, clustering a 120,000 compound dataset on a desktop Linux machine took 7 hours with Ward's, 2.5 hours with Divisive K-means and 44 minutes in parallel on four machines. Using this same Linux cluster we were able to apply cluster analysis to a 1-million compound dataset in 16 hours. We thus believe that the combination of Divisive K-means and larger Linux clusters will allow the clustering of extremely large datasets (tens or hundreds of millions of compounds). We are currently testing this using a 208-node Linux cluster at Indiana University.

2. Accurate virtual screening using high performance computing. The potential for using virtual screening techniques as an *in silico* surrogate for high-throughput screening has been well explored in the literature.⁸⁵ There is a range of computational techniques available to effect virtual screening, depending on the size of the dataset to be screened and the amount of information available about the protein target. For example, if no protein target information is available, but existing compounds are known that exhibit activity, similarity-based screening or pharmacophore-based screening may be used.^{86,87} If a protein structure has

been resolved, and the binding site can be discerned, then docking is generally considered to be the most accurate predictive method.⁸⁸ Virtual screening methods have been parallelized using grid computing.^{89,90} The University of Michigan's Center for Chemical Genomics (CCG), part of their Life Sciences Initiative, consists of a modern Ultra High Throughput Screening laboratory with a compound collection of around 35,000 compounds (and growing). They provide services right through from assay design to after-screen analysis. Dr. David Wild has recently approached personnel at the CCG about the possibility of working with them in the context of the grant. The CCG has shown a willingness to evaluate the tools that we develop and to collaborate on their development, as well as to share datasets with us.

c. Design of a novel database for molecular modeling and computer simulation data

Baik brings significant expertise in molecular modeling, in particular quantum mechanics and combined quantum mechanics and molecular mechanics (QM/MM) methods, to the Center. The research philosophy and infrastructure already in place in the Baik group serve as prototypes of the synergistic and integrated discovery environment that IU's CCR would like to design on a larger scale. Holding formal appointments in both School of Informatics and Department of Chemistry, Baik represents a new generation of scientists who are formally at home both in the fields of traditional chemical research and the new discipline of cheminformatics. In the last few years Baik has built a relational database for quantum chemical electronic structure simulation data (Varuna) and demonstrated its utilization in a real-life chemical research environment in a number of different research projects, some of which form the future prototype projects mentioned below.

Currently, Varuna contains both metadata and results from roughly 25,000 high-level quantum and QM/MM simulations from research areas that the Baik group has been involved in, such as cisplatin-DNA interactions, natural product synthesis and mechanistic studies on cytochrome P450. The main function of Varuna has been to (i) store the raw data from QM and QM/MM simulations at a pre-registered central location and import a selected few important data points to the database. (ii) allow simple and intuitive browsing and access to this data warehouse. (iii) automatically move data from and to the computing sites (Clusters), manage and monitor job execution remotely (iv) generate standard follow-up calculations with minimal human labor and simplify everyday tasks through automation. (v) enable simple browsing, queries and visualization of results. All these functions are currently implemented in Varuna. Thus, Varuna combines the functions of a classical GRID-portal and a data-depository system. Currently, a few portals to quantum molecular simulations in Supercomputing environments exist, but practically all of them are single task applications, where the application can service individual simulations, but does not have an intrinsic "long-term memory" that we propose to implement in form of a Molecular Modeling Database based on the experiences of designing the prototype application Varuna.

As with many of the software packages in use today in academic chemical research, Varuna was initially designed without a vision for large-scale distribution, but mainly because of the desperate need in real-life chemical research for a chemical informatics tool – Baik started the development as a "side-project" in graduate school and continued the implementation as a postdoctoral fellow on a "need-to-be" basis. Therefore, neither the architecture nor the data mining tools are designed in a cohesive and planned manner that would lend itself to a large scale utilization beyond the current use as a customized research platform. But, many lessons have been learned in the past few years that will help to improve the software design. These lessons are particularly relevant, as they were learned in a realistic research environment to solve immediate and real problems that hampered research progress, instead of being problems that a software designer who is not an expert in chemical research assumed to be problems.

C.II. Cheminformatics Education.

a. Introduction. Academic chemistry departments in the US typically offer only limited cheminformatics instruction, if anything at all. Although significant centers of academic excellence in cheminformatics exist in Europe, their activities have in most instances taken place outside the chemistry departments at their institutions.⁹¹⁻⁹³ Consequently, the academic curriculum for cheminformatics is not as well developed as that in other sub-disciplines of the chemical sciences, perhaps because of the diversity and interdisciplinary nature of cheminformatics. Indiana University recognized the need for instruction and research in cheminformatics, and when IU created the School of Informatics just over 5 years ago, it was decided to develop the first cohesive curriculum in chemical informatics in the United States.⁹⁴ In light of that decision, two of the major players in the chemical informatics arena, MDL Information Systems and Daylight Chemical Information

System, have over the past three years each donated \$15,000 fellowships to assist in recruiting talented graduate students. Furthermore, a retired Eli Lilly chemist who is now an adjunct professor at the Bloomington campus gave \$75,000 to jumpstart the program.

Indiana University has proven to be a superb place to build a chemical informatics academic program. It lies in close proximity to and has strong ties with major pharmaceutical companies and chemical informatics organizations. Many innovative, technology-based chemical informatics services are associated with IU, including the Quantum Chemistry Program Exchange (QCPE), the Reciprocal Net project, CHEMINFO, PCMODEL, and BioTech.⁹⁵⁻⁹⁹ In the mid-1990s, we were one of four Midwestern universities to participate in a US Office of Education-sponsored project to design a Web-based portal and training program for the burgeoning bioinformatics Web database resources. The resultant product, BioTech, won many awards, and the Web database developed at Indiana University under that project ultimately formed the basis for Elsevier Science's BioMedNet database of Internet resources.⁹⁹⁻¹⁰⁰

The educational initiatives of the CICC will require the support of significant cyberinfrastructure: computing, data storage, database and visualization resources. Indiana University has a long history of providing computing, visualization and data resources in support of computer science and areas of application of advanced information technology, as well as participation in state and national Grid initiatives. The Indiana Genomics Initiative funding, awarded by the Lilly Endowment in December 2000, resulted in a major expansion of these facilities. In October 2003, both Indiana University and Purdue University were awarded funding from the National Science Foundation to become part of the Extensible TeraScale Facility. This funding is allowing IU and Purdue to jointly contribute resources for the creation of a grid of supercomputing, visualization, and data facilities within the State of Indiana's three largest research campuses (Indiana University-Purdue University Indianapolis, Purdue University West Lafayette, and Indiana University Bloomington). This IP-grid (Indiana Purdue grid) is interconnected via the State of Indiana's I-Light high-speed optical fiber network. The cyberinfrastructure resources described in this proposal will be made available to assist the development and deployment of the ECCR, in keeping with IU's mission of education, research and public service.

b. Existing Program in Cheminformatics Education at Indiana University. The School of Informatics graduate and undergraduate programs in chemical informatics span the two main campuses of Indiana University: Bloomington and Indianapolis. Furthermore, there exists a complementary, well-established graduate chemical information specialist program in the School of Library and Information Science (SLIS) at the Bloomington campus.¹⁰¹ Several of the SLIS faculty members have joint appointments with the School of Informatics. In July 2005, the IUB Department of Computer Science was moved administratively from the College of Arts and Sciences to the School of Informatics, bringing the total faculty in the School at Bloomington to nearly 60. Thus, the chemical informatics academic research and instructional program can draw on faculty in two major academic research centers, including those in the innovative research programs of the IU School of Medicine in Indianapolis.

1. Faculty. Current faculty in the chemical informatics program include Dr. Mu-Hyun Baik, Dr. David J. Wild, Dr. Kelsey Forsythe, and external adjunct faculty members Dr. Dimitris Agrafiotis, Dr. John M. Barnard, Dr. Thompson N. Doman, and Dr. John McKelvey. Dr. Gary Wiggins serves as Director of the Chemical Informatics Program and Interim Director of the Bioinformatics Program at both campuses. Supplementing this group are faculty members who hold School of Informatics adjunct professorships at IU: Dr. David Clemmer, Dr. Gary M. Hieftje, and Dr. Peter Ortoleva from the IUB Department of Chemistry and Dr. Peter Cherbas, Professor of Biology and Director of the Center for Genomics and Bioinformatics at IUB.⁹⁸ The School of Medicine has established a Center for Computational Biology and Bioinformatics, whose director, Dr. A. Keith Dunker, also holds a faculty appointment in the School of Informatics.⁹⁹ Guest speakers have willingly given of their time to participate in the cheminformatics courses, among them such well-known people in the field as Guenter Grethe, George W.A. Milne, Val Gillet, and Sean Ekins. A Science Informatics Advisory Board made up of outstanding representatives from industry and academia will advise us on the curriculum and overall goals of the programs in science informatics.

2. Students. To date, four students have received the MS degrees in chemical informatics, and nine are currently enrolled in the programs. The MS in chemical informatics program and the MS in bioinformatics program exist at both IU campuses. A new Ph.D. program in Informatics is being offered at both locations starting in the fall of 2005, and a unique laboratory informatics track within the chemical informatics MS program is now firmly in place on the Indianapolis campus with 13 MS students.¹⁰⁴ The joint appointment of Baik in Chemistry and Informatics naturally gives rise to a diverse research group consisting of Ph.D. students

both from Informatics and Chemistry. Currently, there are 4 Chem-Ph.D., 1 Info-Ph.D., and 1 Info-M.S. students in the Baik group, while another student has entered the Chem-Ph.D. track after successfully completing a M.S. degree in Bioinformatics at IUB. As the group works in the same laboratory space, there is natural and seamless integration in graduate research between the areas of chemistry, biophysics, informatics, and computer science.

3. Courses. Four graduate chemical informatics courses have been developed to date. They are C571 Chemical Information Technology, C572 Computational Chemistry and Molecular Modeling, and two more specialized courses that are being offered for the first time this semester as I590 topics courses. The School of Informatics undergraduate program requires all BS students to have a cognate area, one option for which is chemistry. Two undergraduate courses, paralleling the subject matter of the graduate cheminformatics courses on a reduced one-semester-hour basis, were created for the cognate in chemistry. Those have been offered for the last two years.

The Chemical Information Technology course, taught by Wild, covers chemical structure and data representation and search systems, as well as chemical information and database systems. Molecular modeling, quantum and molecular mechanics techniques, molecular diversity, combinatorial libraries, and other related topics form the basis of the C572 course. Baik is the instructor for that course.

One of the new topics courses, also taught by Wild, covers programming for chemical informatics. The course is designed to give students a thorough understanding of all aspects of software development for chemical informatics, as well as a broader experience of working in a small scientific computing group. Topics include programming for the web, drawing and depiction of chemical structures in 2D and 3D, chemical informatics toolkits, software APIs, artificial intelligence and machine-learning algorithm development, high performance computing, database management, managing a small software development group, and design and usability of cheminformatics software.

The Information Retrieval from Chemistry and Life Sciences Databases topics course, developed by Wiggins, concentrates on the unique retrieval techniques possible with the major commercial chemistry databases Chemical Abstracts, Beilstein, Gmelin, and the Cambridge Structural Database. Also covered are the STN Messenger search software and the various science databases that are accessible through STN International, including Medline. These are contrasted with the free, public databases in the life sciences area.

4. Software and Databases. A significant amount of software and databases has been amassed for use in this and the other courses in the curriculum (See Table 1).

Company	Products and/or (Target Area)
ArgusLab	(Molecular modeling)
Barnard Chemical Information	Toolkit (Clustering)
Cambridge Crystallographic Data Ctr	Cambridge Structural DB & GOLD
CambridgeSoft	ChemDraw Ultra
Chemical Abstracts Service	SciFinder Scholar
Chemical Computing Group	MOE
Chemaxon	Marvin (and other software)
Daylight Chemical Information Systems	Toolkit
FIZ Karlsruhe	Inorganic Crystal Structure Database
MDL	CrossFire Beilstein and Gmelin
OpenEye	Toolkit (and other software)
Sage Informatics	ChemTK
Serena Software	PCMODEL
Spotfire	DecisionSite
STN International	STN Express with Discover (Anal Ed)
Wavefunction	Spartan

Table 1. Software donated/obtained for use in the chemical informatics program at Indiana University

5. Distributed Education and Web Resources. Since the first year they were taught in 2001, the chemical informatics courses have pioneered the use of videoconferencing and software conferencing

technology to share both computer and human resources. Using such technology, we have expanded our links across the Atlantic Ocean in the past two years. An extensive web site for the Chemical Information Technology class was developed, and we have made heavy use of IU's classroom management program (Oncourse) and the Macromedia Breeze product for interactive instruction.¹⁰⁵ In the fall semester of 2005, a true Distance Education version of the I571 class was offered, with ten students enrolled from California to Connecticut, in addition to the ten on-site students.

Web resources have been an essential component of the instructional activities at IU to date, and all courses provide links to basic resources on the Web. However, a more systematic approach to assisting the novice user, as described in the Web guide project below, needs to be developed.

6. Involvement with Other Cheminformatics Educational Efforts. Both Wild and Wiggins are among several advisors to a chemical informatics instructional project being conducted by Mesa Analytics & Computing's John and Norah MacCuish. That NSF-funded SBIR II project, "Cheminformatics Virtual Classroom," began March 1, 2005 and runs for two years. Wiggins has had preliminary contacts with researchers at the University of Manchester (Steven Liem and Patrick O'Malley) concerning collaboration on instructional modules they are developing.

C.III. Chemical Prototype Projects

Although our concept of conquering new areas of biomedical and chemical research with Cheminformatics tools is generally applicable to many areas, we have chosen to initially concentrate on application areas where we already have the domain knowledge and an extensive network of experimental collaborators available. At the core of our collaborative and multidisciplinary approach is the research laboratory of Baik, who holds formal tenure-track appointments in both the School of Informatics and the Department of Chemistry, thus naturally providing the platform of interdisciplinary communication. The appointment of Baik itself is a reflection on the commitment of Indiana University to developing the field of cheminformatics from a specialty discipline to a widely accessible and applicable vehicle of chemical discovery that is an equal partner to traditional chemical methods. To realize this vision, cheminformatics must engage in biomedical and chemical research that goes beyond what could be considered "classical cheminformatics," which has been lead discovery and optimization. Historically, cheminformatics was born out of the need of organizing and systematically guiding drug discovery processes. Whereas the success of cheminformatics in commercial environments is undeniable, it has also left its mark on how cheminformatics is viewed in the chemical research community. Our approach is bidirectional in the sense that our work outlined above coincides with traditional utilization of cheminformatics, but also embraces new areas. Instead of designing a tool and asking which fields of chemical research we can apply these tools to, we identify chemical problems of global interest, bring on board experimental chemists with a specific research agenda, and design cheminformatics tools that will facilitate the discovery process in that specific domain. Our proposed approach of making low-resolution databases like PubChem powerful catalysts for novel avenues of discovery is complementing them with high-resolution databases, like the Reciprocal Net and computational molecular modeling.

Consequently, the problems that we propose to utilize as prototypes for demonstrating how novel cheminformatics tools could be used, are not only valuable as a proof of concept, but have an immediately relevant research agenda. Of course, the chemical goals of these projects qualify for external funding independently from the developmental aspects of cheminformatics and such funding is already being pursued. So, why did we seek funding for these prototype applications from the ECCR grant? The answer is very simple. As the individual projects center around the chemical interest, we were and would have continued to be unable to properly address architecture and performance issues of the discovery network without separate funding that allows us to dedicate manpower to these technical issues. Below, a number of projects will be profiled with the goal of highlighting the potential use of cheminformatics tools in areas that are not classical areas of cheminformatics.

D. Research Design and Methods

D.0. Introduction

The Indiana University School of Informatics, in partnership with Computer Science, Chemistry, Biology, and University Information Technology Services plus our external advisors, proposes to help develop the next generation of tools to assist scientific research as we move further into the era of e-Science. Through the creation of an Exploratory Center for Cheminformatics Research at Indiana University, we will add significant value to the research teams that have already found collaborators in various disciplines. The informatics partners will help scientists to better discover and manage relevant data and make informed decisions because informatics deals largely with data configurations and their structure, properties, transformations, and applications. With the extensive technical infrastructure and many scientists with interdisciplinary interests on its two main campuses (Bloomington and Indianapolis), Indiana University is in an excellent position to establish a Cheminformatics Research Center in the coming years.

An initial workshop with all advisors and participants will be held in the fall of 2005 to identify a preliminary list of key issues. Out of this will come a chemical informatics topics course in the spring 2006 semester where some of the external advisors and faculty on the proposal will give presentations. A more coherent course will be developed for fall 2006. Prototype projects and the School of Informatics Ph.D. program should be well underway by that time. With preliminary results in hand, we will fund a 2-3 day workshop or conference at IU in the fall of 2006 on emerging themes in chemical informatics. All IU personnel, industry people, and other external advisors will be involved. We will develop cutting-edge solutions for the chemical informatics projects outlined below. In so doing, we will apply those solutions to areas of research and teaching that increasingly must deal with mountains of chemical data. These efforts will ultimately have a significant impact far beyond Indiana University, as chemical informatics finds greater applications to solve questions of the chemical basis of life.

D.I. Technology Development

a. Design of a Grid-based distributed data architecture. (led by Fox and Bramley)

In previous sections, we have surveyed Grid technology and have discussed how it may be applied to domain science problems such as crystallographic data and atmospheric and solid earth simulations. Such applications are characterized by data sources (both archival data bases and streaming sources) and computational methods for both first principles computations and data analysis.

These Grid systems, despite the vast differences in their domains, have many common compatibilities. Typically, problems with data sources include the need to provide programmatic interfaces to remote data sets so that they can be directly integrated with applications. Quite often, applications must be chained into sequences or more complicated workflow graphs, as obtaining final results is a multi-staged process. These various stages may use different computing resources and must generally be managed by notification systems, as the computing steps may take a significant time to complete.

When building the Chemistry Virtual Laboratory (CVL) grid, such low-level Grid technical plumbing must be coupled with higher level tools for the scientific user. Computing Web portals are a common way to create user interfaces to applications. Modern portal systems are built from reusable parts (portlets) that can be used to assemble new portals from pre-built parts. Portlet components are managed by a portal container, which also provides common services such as login and access control to portlet components. It is thus possible to build portal systems that provide multiple views for different users. The LEAD effort is a good example of this: different users may include K-12 students, the general public, project managers, and working scientists. We envision the CVL Grid portal systems as providing a similar collection of differing views.

Computing portals are only one of many possible interfaces to the CVL Grid. By following the principles of Service Oriented Architectures and the related Web Service Architecture, we make a clean separation between user interfaces and their underlying services. We must thus provide more sophisticated, specialized clients to Grid Services. Examples include high-end, interactive visualization client tools and audio/video collaboration clients. Browser clients may be easily integrated with these more specialized tools. The underlying Grid services are the same. For example, jobs may be created and launched through the browser.

User environments such as the computing portal system introduce another class of related services. These are “context and information environment” services that are needed to manage the user’s interaction with the Grid. These are architected and built in exactly the same fashion as Grid services for job and file/data management. Examples include the context services used by SERVOGrid. These allow portal users to organize their interactions with the portal into project folders. Project folders maintain all of the information, or metadata, generated by the user in that session. Typical metadata examples include the codes selected, the parameters and input files used, and the hosts where the jobs were run. Such systems allow users to recover old sessions and modify work. More importantly, they act as a very precise scientific notebook that allows the user to reproduce his or her results.

As shown by such projects as MyGrid and CMCS, metadata services can be quite interesting and sophisticated. The common problem here is that results of calculations must be shared between researchers by putting these results into online digital libraries. The results in these libraries may be used to produce other results in the library, creating long dependency chains. It is inevitable that the entries in these archives must be verified to isolate and remove errors. The metadata associated with the entries, such as the author or creator, the method of creation (simulation technique, experimental method), and the direct ancestor data entries, all constitute the data entry’s provenance, or pedigree. Expressing this pedigree is best done with the techniques of the Semantic Web.¹⁰⁶ The Scientific Annotation Middleware project of Pacific Northwest National Laboratory’s chemistry group is an exemplary effort in the chemistry metadata management field.¹⁰⁷

Much software developed for other application Grids is general in nature and can be directly reused by the CVL. Typically, we have found that the data grid components are the least reusable, since they are specialized to the data models of specific scientific domains. When building the data management services components of the CVL, we will adopt best practices (if not software) from related efforts. For example, experience with other data Grid service systems such as Geographical Information Systems has shown the power of combining common abstract data models with a common core suite of services for managing instances of these data models. In GIS systems, Web Feature and Web Map services are used to interact with and render images of the core Geometry Markup Language. Similar sets of services would benefit the CVL data grid components. The Chemical Markup Language, for instance, is already available to serve as the common data model.

b. Development of tools for HTS data analysis and virtual screening. (led by Wild, Gannon, and Pierce)

A compound database in a typical company would likely hold the following information:

1. Basic information about a compound
 - Company’s unique ID for the compound
 - 2D structure (SMILES)
 - Pre-calculated properties (LogP, MW, etc)
 - Flags (rule of five violation, toxicity, etc)
 - Source (i.e., who made it, where it came from)
 - Barcode or other stockroom lookup
 - Amount available
 - Notebook reference for making the compound.
2. HTS screening data (same idea for non-HTS)
 - For each compound tested:
 - Percent inhibition
 - IC50 if known
 - Possibly raw values (fluorescence, etc), but may be in different database
 - Active / inactive designation
 - Plate number
 - Well number
 - Active / Inactive cutoff
 - Control data

- Experiment name
- Experiment protocol description
- Date of experiment
- Notebook reference

We anticipate that many of the types of data above may eventually end up in the PubChem database. The following projects are planned to help people make optimum use of such data.

1. Rapid dataset organization using cluster analysis techniques. This project aims to develop and evaluate procedures for rapidly organizing the chemical structures in a repository into structurally-related groups, using 2D structure fingerprints and fast cluster analysis. Cluster analysis techniques have traditionally been used in chemical informatics for one of three purposes: to group together compounds that may exhibit similar biological activity, to group compounds into series or structurally related groups, and for picking small sets of “representative compounds” from large datasets.¹⁰⁸⁻¹⁰⁹ Each of these would be of relevance to a public data repository and highly useful for data analysis, but the clustering method has to be fast enough to organize millions of compounds in reasonable time, and to include new data points as added. We will evaluate a number of fast clustering algorithms, including the use of parallel Linux cluster versions, and develop tools for the automatic clustering of the repository by 2D structural fingerprints. These tools would be structured as Grid Services to allow plug & play substitution of algorithms and convenient composition by Grid workflow.

2. Design of intuitive interfaces for navigating and analyzing large chemical datasets. This project seeks to create powerful, intuitive, web-based tools for analyzing the data points in a publicly-accessible HTS data repository. Each High Throughput Screening experiment produces data points for thousands, tens of thousands or even hundreds of thousands of chemical structures. Tools are needed to organize the compounds and data (for example, into chemically-related groups), to enable scientists to navigate and explore the results, and to aid in their understanding and analysis. We are currently researching a number of approaches to this, including the development of ultra-fast cluster analysis algorithms, the use of visual tools for interacting with the data points, and the development of new interfaces for visualizing the data (including how large numbers of chemical structures are depicted). This project will involve specifically developing and testing these approaches in the context of a publicly-accessible High Throughput Screening structure and data repository. Contextual Design and Usability techniques will be used to evaluate a variety of different interfaces. We shall seek to understand how chemists can most effectively and intuitively interact with the chemical structures and biological data in the repository, and develop and publish recommendations for the design and implementation of software used for handling large chemical datasets. We will develop and make publicly available web-based tools that are formed out of this research, and which can be directly applied to the existing PubChem database. The user interfaces will be hosted as Portlets.

3. Accurate virtual screening using high performance computing. This project will enable the use of computationally demanding virtual screening techniques on an HTS repository through the use of parallel computing. *Virtual screening* refers to the process of computationally analyzing the chemical compounds in a large chemical database (represented by the compounds’ 2D or 3D structures) in order to extract the compounds most likely to show biological activity against a particular protein target. It is often used as an *in silico* surrogate for high-throughput screening. The most accurate methods (namely docking and pharmacophore-based screening) are also the most computationally demanding, and are thus not traditionally considered suitable for large datasets. However, the advent of Linux clusters and Grid computing offer the possibility of vastly speeding up such virtual screens by doing many calculations in parallel. In this project, we will use Indiana University’s extensive parallel processing resources (including large Linux clusters and access to the Teragrid grid computing framework) to test the feasibility of applying accurate calculations to very large datasets. If successful, this would enable “virtual screens” to be carried out on publicly accessible HTS dataset repositories. Of particular interest would be whether these techniques could be used to “fill in gaps” in a screening repository. For example, if actual screening results are submitted for a portion of the compounds present in the repository, they could be used as a basis for calculation of “virtual” results for the rest of the compounds in the repository. We will exploit the Grid interfaces to the local and national resources to make it straightforward to run on multiple back-ends from the same user (portlet) interface.

4. A standard format for HTS data exchange. Working with appropriate bodies in the standards setting area, we shall strive to develop standard formats for the exchange of High Throughput Screening data. A high

throughput screening experiment will generate multiple levels of data, for example, raw fluorescence data, percent inhibitions, IC-50's, control data, plate layouts, experimental designs and protocols, error margins and annotations. For successful subsequent chemical analysis and follow-up, this data must be made available in a form meaningful to a chemist, along with 2D chemical structure information, calculated properties, and related experimental data. We will work with high-throughput-screening biologists as well as chemists and informatics experts to develop a standard format or formats, based on XML, for the communication of the data and semantics of HTS data, along with tools deployed as Grid services for converting existing free-form data formats (such as comma-separated files) into the standard format or formats.

5. Data mining techniques for detecting structure-activity relationships across multiple screens. We shall attempt to develop new, rapid methods that can be used to discern structure-activity relationships between compounds and multiple biological data points, and effective techniques for communicating these relationships in ways meaningful to chemists. Traditionally, structure-activity relationship studies (both manual and computational) are applied to small groups of related compounds to discern the relationships between structural variations and biological activity. Since the advent of HTS, new algorithms for discovering these relationships in much larger, more diverse datasets have been developed. In particular recursive partitioning and support vector machines have proved effective.^{108,109} However, these are generally only used in the context of discovering relationships with a single target. In this project, we will try to demonstrate that these and other methods can be used to proactively data-mine structure-activity relationships across different screens, and thus be able to quantify factors such as selectivity and promiscuity as well as suggest potential links between related screens.

6. Systems Biology Approaches. In bioinformatics and chemical informatics, the functions of uncharacterized proteins have usually been inferred on the basis of quantum reaction mechanisms, sequence similarities, common structural motifs, gene order, or similarities in protein structure and gene expression. Recently mathematical and computational methods developed by systems biologists predict function based on the role of genes in complex biochemical networks.¹¹⁰ These methods allow us to predict functions for proteins independent of homologies in structure or sequence and provide a way to characterize proteins that have not yet been studied, using published biological data from high-throughput technologies. High throughput experimental assays play a major role in the current shift from reductionist to "systems" approaches to biological and biomedical problems. The data sets these experiments generate promise to identify the components and interactions of regulatory biochemical networks.

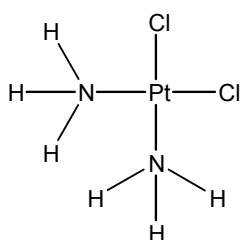
At the Indiana University School of Informatics, systems biologists are focusing on the development of algorithms and approaches to determining the function and kinetics of proteins on complex reaction networks from protein interaction maps, and nuclear magnetic resonance/mass spectrometry time series data of complex biochemical pathways.¹¹¹ This work is being architected as Grid services extending CellML and will be consistent with the CIC approach. Our aim is to investigate and develop techniques that allow the function and kinetics of proteins in biochemical network to be inferred from high throughput experimental data, with little prior information about the pathways involved. Mechanistic studies of biochemical networks are important for several reasons: (i) an improved understanding of the functional role of different molecules can be achieved only with the knowledge of the mechanism of specific reactions and the nature of key intermediates in a complex network context; (ii) the control (or regulation) of different biochemical pathways can best be understood if some hypothesis for the complex biochemical pathway is available; and (iii) kinetic modeling, which forms the basis for understanding reaction kinetics, is based on comprehensive information about the reaction mechanism. Kinetic models allow simulation of complicated pathways, and even whole-cell dynamics, which is proving to be an increasingly important predictive tool in the post-genomic era.¹¹² The methodologies to be investigated will apply to any organism, including humans, where only three to five percent of genes have identified functions. Understanding the function of genes and proteins in a network context, will improve our ability to predict and control their responses to internal and external perturbations. We need to start building a systematic library annotated against complex systems, and with multiple compounds, begin to define profiles that impact the key biological signaling points. In other words, we need to build links from the chemistry to where the biology leads us. Data should be capable of being interrogated in the context of more complex systems. Licensing of the CACTVS cheminformatics toolkit, with the full ASN.1 parser for CACTVS, which understands the full data specification for structures and assay data, may be an appropriate first step to interrogating the data in PubChem and other NCBI systems.

In all of the projects described in this section, care will be taken to work closely with chemists to make sure that the data being considered are chemically meaningful, that they include the appropriate values or features, and that the results include compounds which would be practical to synthesize.

c. Augmenting the database network. (led by Baik and Plale)

1. Resolution of data in different databases. The network of resources and knowledge domains that we propose to connect into a cohesive discovery environment is currently very heterogeneous in scope and resolution. For example, a simple PubChem query on cisplatin, which is one of the prototype anticancer drugs that we propose to study, shows the entry shown in Figure 3. On a low-resolution scale, this entry is sufficiently accurate, as some of the most important characteristics of this small molecule drug are captured. However, for utilizing this information on a high-resolution scale of academic small-molecule research it is important to augment it with a specialty database that will provide higher resolution information. Figure 4 shows an entry of cisplatin in our current Molecular Modeling database. Without proper adjustment for the different scale and resolution, the PubChem entry of cisplatin cannot be utilized effectively: The structure shown in Figure 3 is actually trans-diamido-Platinum(II) dichloride, which does not exist. The molecular formula is also inaccurate on an atomic scale, as the ammine ligands have been replaced by amido groups. In addition, the PubChem entry closely resembles the isomers trans-Platinum(II)diammine dichloride, which is inactive as an anticancer agent.

Given the size of PubChem and its purpose of providing experimental



Name: cis-Platinum(II)diammine dichloride
Molecular Formula: $\text{Cl}_2\text{H}_6\text{N}_2\text{Pt}$

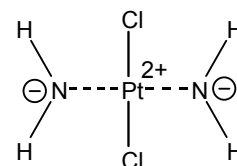
Figure 4. Specialty DB entry for cisplatin

assay data and housing the results of massive HTS, where the actual structure of the molecule being sampled may be unknown, inconsistencies and certain level

of uncertainty are unavoidable. Therefore, for facilitating an effective use of PubChem in a high-resolution academic environment, we must augment PubChem with a specialty database that will house the expertise and experience of the specific research group. We propose to examine and test possible implementation strategies for maximizing the efficiency of combining general low-resolution databases with specialty databases with a much higher resolution. The test platforms for this effort will be the distributed XRay-crystal structure database Reciprocal Net and the high-resolution molecular modeling database Varuna. One solution to merging data of different resolution is

introducing a measure for the fidelity and reliability of the entry as a part of the metadata. Data mining tools will be designed that perform proper queries to extract and combine entries in different database to match the specified level of resolution. We will use emerging Grid standards including OGSA-DAI to integrate the databases.

2. Chemical Markup Language. A key challenge that needs to be addressed is finding a common language platform that will allow for the members of the heterogeneous database network to efficiently communicate, exchange data and allow for assembling query results from different databases into one cohesive format amenable for further processing. We will further develop Chemical Markup Language (CML) as a variant of XML for this purpose. Encoding 2D and 3D structural information and other chemical information in a has now become common and PubChem already has XML reporting capabilities. We will expand the existing CML with molecular modeling and Xray-crystal structure specific features. A non-trivial problem that needs attention is how to encode equilibrium structures, transition states and excited state molecules that were computed using high-level molecular modeling techniques and annotate them appropriately to allow for the construction of reaction energy profiles. Ideally, we would like to be able to precisely describe the reaction trajectories, visually be able to identify reactants, intermediates and products, and indicate where experimental data (e.g., structural information about trapped intermediates) are available. A potentially powerful tool would be to start from the reaction energy profile and query the database network for structures that resemble the putative intermediates or the transition states, The CML must be flexible enough to both report these types of new queries. We will start with the CML as currently implemented in PubChem and devise a number of different strategies for extending it to make the format more flexible.



SID: 63885
CID: 129129
Name: cis-Platinum(II)diammine dichloride
Molecular Formula: $\text{Cl}_2\text{H}_4\text{N}_2\text{Pt}$

Figure 3. PubChem entry for cisplatin

3. Design of a novel molecular modeling database. Molecular modeling has become an important tool in chemical research. A superficial survey of recent publications in all peer reviewed journals shows convincingly that modern research in chemistry today increasingly incorporates quantum mechanical (QM), molecular mechanical (MM) and combined QM/MM simulations to either compliment experimental findings or to give access to plausible model chemistry where experimental data are difficult or impossible to obtain. QM/MM methods allow for combining the efficiency of a molecular mechanics model that treats the atoms in a molecule as Newtonian particles with the accuracy of quantum mechanics at the reactive centers of enzymes where bond-breaking and bond-formation events occur. If the technical modeling framework is chosen carefully, very credible and crisp concepts for biochemical processes on a molecular level can be obtained. Equilibrium structures, transition states, and thus activation barriers of reactions, redox potentials and almost all spectroscopic observables, vibrational frequencies (infrared spectroscopy), NMR chemical shifts, UV/VIS absorption spectra and many other chemically meaningful properties can be reliably computed at a high level of accuracy. A number of characteristics make molecular modeling uniquely well suited to complement the planned federation of databases:

(i) As the actual data is machine generated, they are intrinsically machine-friendly or can be made machine-friendly without much effort.

(ii) By connecting a molecular modeling database with an automated portal to the computing GRID, it is possible to implement an easy protocol for expanding the database knowledge without or with minimal human interaction. If we detect an incomplete data set, e.g. a crucial transition state of a reaction that has been suspected to be most relevant for a study, it is relatively easy to fill the void by requesting the simulation to be carried out on the computing GRID.

(iii) The resolution of data is orders of magnitude higher than what is usually available in experimental data. If we carry out a large-scale QM/MM calculation on a metalloenzyme active site, we can not only obtain the structure of the protein and the conformation of the active site, but trace every electron by examining the computed wavefunctions.

(iv) Transition state (TS) structures and their energies are available – experimentally, it is practically impossible to get direct structural information from transition states.

Of course, since all results are computer simulations, there is the need to collate and compare to experimental benchmark data to assess whether or not the simulation is a realistic model of the chemical reality. Therefore, the proposed federation of simulation data with experimental databases offers the unique opportunity to also systematically compare the simulations with experimental data. For this vision to become reality, a number of demanding challenges have to be overcome:

(i) Implementing a database for the metadata (computational method, basis set, simulation type, etc.) is simple and has already been done in our preliminary prototype database VARUNA. Similarly straightforward, although technically somewhat challenging is the import of a set of the key results, such as computed energies, HOMO and LUMO energies or the computed 3D structures in Cartesian coordinates. However, it is not clear whether importing the whole wavefunction of a quantum simulation into a database is helpful or not. Just as the actual protein structures are not imported into the PDB, it might be more efficient to simply introduce a pointer to a file that contains the computed wave functions. In approximately 2 years, the Baik group has collected data from roughly 25,000 high-level QM and QM/MM simulations, each with a wave function coefficient matrix of roughly 1000x1000 on average. Efficiency issues have to be explored and we will experiment with different scenarios of using the wavefunction analysis.

(ii) As the only input parameters in quantum chemical simulations are usually xyz coordinates of a molecule, connectivity data is usually not available at the onset of a computer simulation. Simply using Cartesian coordinates for recognizing structural similarities and clustering of molecules based on substructures is completely out of the question. Thus, we must both implement an explicit 3D structural descriptor that is amenable to rapid 3D structure mapping and also introduce a bitstring representation of the functional groups present in the structure to allow for fast substructure searches. We will begin by implementing the BCI fragment dictionary based fingerprint system from Barnard Chemical Information Ltd. – Dr. John Barnard, who is the founder of BCI and holds an adjunct Professor position in our School will be an associate of the Center and be available for consultation and assistance.

D.II. Cheminformatics Education (led by Wiggins and Fox)

After four years of providing professional masters education in cheminformatics, the Indiana University School of Informatics is moving to the next level of research and teaching by offering a PhD in Informatics starting fall 2005. The field of cheminformatics education has benefited from several recently published textbooks and reference works, but a range of curricular materials is needed to make it easy to export cheminformatics instruction to other environments.¹¹³⁻¹¹⁷ Only then will the many researchers and students who could benefit from cheminformatics gain the necessary knowledge to effectively select and use cheminformatics techniques. We intend to further develop these offerings and make our training modules accessible on a much wider basis in the coming years.

1. Advanced seminar in chemical informatics. Working with the School of Informatics Science Informatics Advisory Board (see Appendix) and the advisors who will participate in the exploratory center project, we will identify the key topics for an advanced seminar in chemical informatics. The course will be offered on a preliminary basis during the spring 2006 semester and refined during the second year of the award period. The topics defined for that course should identify the subject matter most critically needed to effectively integrate cheminformatics into the research practices of academic and other research groups. Modules on interfacing with the Grid and targeted programming are among the key topics likely to be included.

2. Training modules for cheminformatics instruction on the Web. In partnership with others who are working on key educational components for cheminformatics, we will develop and place on the Web sample instructional modules that cover the basic tenets of cheminformatics, suitable for undergraduate and beginning graduate students. These will treat traditional aspects of chemical informatics, as well as molecular modeling and computational chemistry. A Web clearinghouse that provides a one-stop shopping center for free or low-cost instructional materials, aimed at those who desire to develop short courses on appropriate topics, will also be established. Exercises that facilitate practice-based instruction will be a key feature of the clearinghouse. A mechanism to establish and maintain easy communication among a virtual community of cheminformatics scholar-instructors will also be implemented. Wiggins has years of experience facilitating communication among a diverse group of over 1500 people interested in chemical information sources through the listserve CHMINF-L, which has been in existence since 1991.¹¹⁸ Furthermore, he has successfully directed the Clearinghouse for Chemical Information Instructional Materials (CCIIM) for the last ten years.¹¹⁹ The CCIIM is sponsored by the respective chemical information divisions of the American Chemical Society and the Special Libraries Association (SLA), while CHMINF-L is sponsored by ACS, the Royal Society of Chemistry, the American Society for Information Science and Technology, and SLA.

3. Web guide for essential cheminformatics resources A Web guide that leads to the main sources of cheminformatics will be created to facilitate finding relevant cheminformatics information resources and experts via the Web. This will include RSS feeds so that users can automatically become aware of new free and commercial sources worldwide as they are added to the resource.¹²⁰ The guide will be designed for the novice to intermediate user and will have paths to sites with more in-depth coverage as fluency in cheminformatics is attained. Modern practices of Web design and database access methods will be applied in the guide.

4. Graduate research. It is extremely important for cheminformatics graduate students to find research partners among the active chemistry and biosciences research groups at IU who will take part in these projects. We will place students in groups where the interactions will produce feedback that will influence the design of both our educational and research modules. This is in keeping with the basic philosophy of the School of Informatics to educate our students in the human, technical, and domain aspects of informatics.

D.III. Chemical prototype projects (led by Baik)

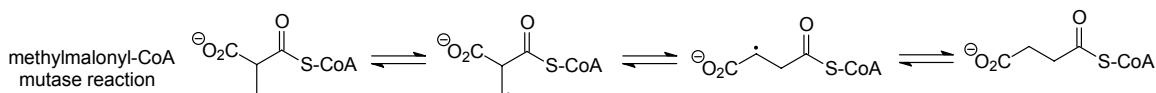
For Cheminformatics to truly have a high impact on academic research, the tools must be designed to perfectly fit into currently ongoing research. When incorporating information technology into classical natural research areas that had little to no use of modern computer resources historically, as is the case for most of the fundamental biomedical research ongoing in academic environments and natural product synthesis research that we are targeting specifically with the proposed center, it is crucial to design a system that is aware of the chemical logic and semantics common to the subfield. Chemical discovery tools that are either ignorant of the current state of the research field or do not convincingly convey the immediate benefit will likely not be embraced by the academic research community at large. Thus, we propose a bottom-up approach, where we get actively involved in specific areas of chemical research, customize our Cheminformatics tools to

the very specific needs of these prototype projects and demonstrate the power of the new technologies specifically. We have assembled a solid foundation of experimental collaborators and have already demonstrated using preliminary and partial implementations of our Cheminformatics tools how the research efficiency can be increased while being able to address questions that are too complex to address without the aid of data mining and data visualization tools.

We stress that conducting the actual chemical research of the projects outlined below is only a secondary goal and thus plays only a minor role for this proposal. Instead, we propose to use the funds of the ECCR grant to **customize and redesign the tools** that we need to accomplish the chemical goals. The ECCR funds will help us to spend effort and time to design tools that are reusable and fit into the grand scheme of developing a discovery network. Currently, all of the tools that we developed for these specific application projects are "hacks" that allow us to "get the job done," but are by no means broadly applicable. As with all software development projects, the implementation of error handling, generalization and optimization for scalability and robustness is tedious and time consuming. Without the external funding specifically for software design, we are unable to address these issues properly. The tools will be structured as Grid services to ensure maximal interoperability.

All projects outlined below represent a very **radical new approach to future biomedical research**: It is our view that whereas combinatorial medicinal chemistry and other systematic, more or less random searches for new biologically and biomedically active agents need to be further developed at a high pace, they alone do not provide the solution to the ultimate challenges of improving human health and finding treatments for currently untreatable diseases. We propose to combine high-volume chemical data with extremely high-resolution data from rigorous molecular modeling work, introducing an unprecedented level of chemical reasoning. By seamlessly and aggressively merging experimental and theoretical chemistry, our discovery strategies and hypotheses will be iteratively improved. We believe this network of diverse domain expertise is a powerful answer to the challenges that lie ahead.

1. Enzymatic reaction mechanism discovery: B₁₂-dependent enzymes. In collaboration with Professor Martin Newcomb (UI-Chicago), we have started studying radical reactions occurring in the coenzyme B₁₂-dependent enzymes, in particular *Methylmalonyl-CoA Mutase (MCM)*. This enzyme is responsible for the catabolism of several amino acids, certain lipids and cholesterol. Microscopically, MCM promotes a one-atom migration of a group in the substrate radical to give the product radical¹²¹.



Although coenzyme B₁₂ has been studied for years, the mechanisms of reactions in the coenzyme B₁₂-dependent enzymes including MCM are not known. Advances in mechanistic understanding of the enzyme-catalyzed reactions can be expected to lead to the development of methods to control the enzymes in nature. Being able to control enzymes based on a rationally constructed strategy is perhaps the most rational and most satisfying approach to development of new therapeutics. The contemporary mechanistic question involves the nature of the catalysis. The currently accepted mechanism implicates the involvement of acids that promote the radical rearrangement reaction in the enzyme is acid catalyzed¹²²⁻¹³¹. This suggestion was presented in 1973 by computational chemists in the context of catalysis of the diol dehydratase reaction¹³². Although no acid in an enzyme will be strong enough to protonate the substrate radicals, the computations indicated that hydrogen-bonding to a weak acid, so-called "partial protonation", will result in large reductions in the barriers for the associative reactions. The putative acid catalyst is protonated histidine or possibly a carboxylic acid.

With new methods, kinetic studies in the Newcomb lab confirmed the high barriers for the neutral radical reactions, but the new results are in conflict with computational predictions for polarized systems. For cyclizations of models for the α -methylene-glutarate radical, no catalysis was found with the strong acid CF₃CO₂H¹³³. In models for the methylmalonyl-CoA radical, the cyclization reactions were slightly accelerated when the radicals were complexed with the strong acid CF₃CO₂H but unaffected when the radicals were dissolved in acetic acid¹³⁴. One conclusion from the combination of the small model computational studies and the new experimental results is that the simple associative and dissociative mechanisms are not acceptable for the radical reactions in the enzymes, nor is "partial protonation" a viable explanation in some cases. The reactions must be more complex, and we hypothesized that nucleophilic catalysis of the radical reactions is a

key feature. Another important conclusion is that we have effectively reached the limit of what can be gained from small model work in the computational realm; larger models are needed for further progress. In addition, we have to sample a much larger number of potential reaction mechanisms involving explicit substrate-protein interactions within the active site. We have already generated a few dozens of QM/MM potential reaction trajectories where both technical parameters, such as the QM model size, inclusion of different residues of the active site as quantum mechanical particles, as well as the conceptual chemistry have been probed. We anticipate a much larger number of computational models that will allow us to construct a better mechanism. A crucial tools for elucidating the role of the active site and understanding the enzymatic catalysis is to first model the reaction with the proposed new mechanism (nucleophilic, Michael addition type of a reaction) and then to virtually mutate the residues that have contact to the substrate and remodel the reaction. We will develop tools that will allow to conduct such virtual mutagenesis calculation series systematically and conveniently. In addition, we must implement an analysis tool that will go through each of the simulated reactions and collect for example the electrostatic potential changes during the reaction as a function of mutations in the active site. Once the most promising mutation for the demonstrating the effect of the new mechanism is identified, they will be probed experimentally in he Newcomb group.

2. Methane Monooxygenase. The bacterial enzyme *Methane monooxygenase (MMO)* converts methane into methanol. The extensively catalytic cycle of MMO⁷⁰ is shown below in fig. 5, along with proposed structures of the catalytic core for each intermediate in the cycle. Baik's postdoctoral work established the currently accepted mechanism for the hydroxylation of methane performed by MMO, providing an atomic level description of the structures and energies of all of the stationary points (minima and transition states) implied by Fig.1, using high-level quantum mechanical simulations. One of the key accomplishments of Baik previous research was the identification of an intuitively comprehensible connection between the structural distortion of the diiron core and the *in situ* generation of a radicaloid oxygen, which ultimately performs the oxidation chemistry. This mechanistic proposal has received much attention in part because of its potential relevance for the mode of action of a structurally related human enzyme Ribonucleotide Reductase¹³⁵ and the functionally similar enzyme cytochrome P450¹³⁶. For elucidating the electronic structure of MMO, Baik used a novel electronic structure analysis tool that was implemented to mine the data out of a few dozens of quantum mechanically computed wave functions, each of which consisted of a square-matrix with roughly 1.2 million

Figure 5. Catalytic cycle of the methane hydroxylation reaction promoted by MMO.

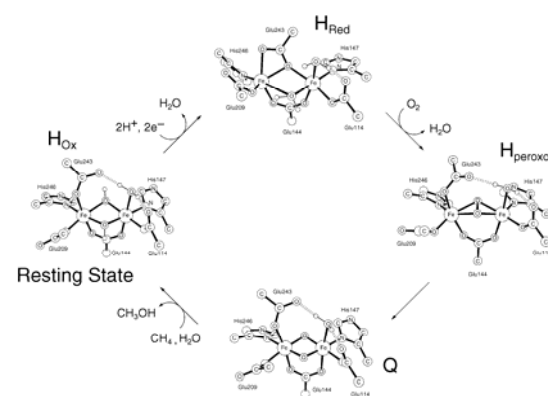
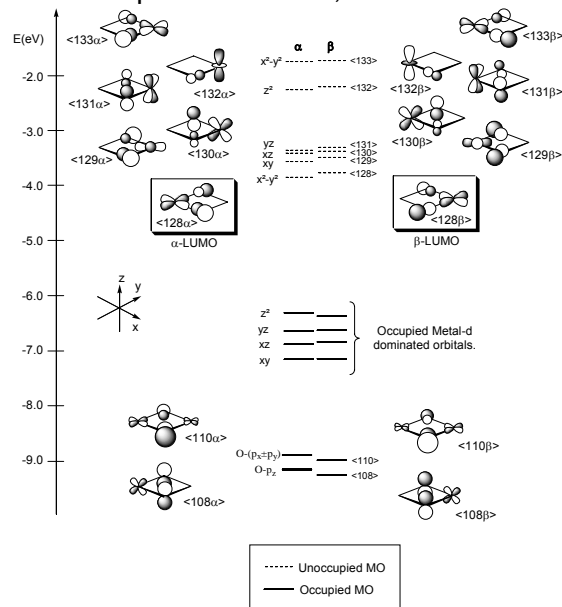


Figure 6. Molecular orbital diagram of the reactive species of MMO, Q.



of this machine-assisted analysis, which was "one of the most complex electronic structure studies in the recent past", as was pointed out by a reviewer of the publication.¹³⁷ The current collaboration with the experimental collaborator, Professor Stephen J. Lippard (MIT), seeks to take the knowledge obtained from studying the enzyme a step further and design a small molecule that will mimic the enzymatic behavior. Predictions of structure-reactivity relationships that came out of our electronic structure analysis are currently being pursued in the Lippard lab. For example, a semi-combinatorial virtual analysis of the diiron core reactivity using the preliminary implementation of our data mining tools identified that the orientation of the N-donor ligands (histidine residues in the enzyme) are crucial. To date, the biomimetics that were prepared in the Lippard lab had all been the anti-isomers, as that is the commonly preferred geometry if the protein is not present to enforce the syn-configuration. Future plans include further explorations of synthetic models of this enzyme which will require a highly detailed electronic

structure analysis. This work showcases a very different Cheminformatics tool that we would like to implement in a robust and general framework. Given a very complex electronic structure of a molecule and its catalytic reaction profile, we require a tool that can analyze and recognize the similarities of molecular orbitals so that they can be correlated and traced easily along the reaction profile. Currently, no such tool exists, and our currently available analysis tool in Varuna, which requires much manual intervention and source-code changes of the tool is to be used for a different system, far too specialized. We would like to explore a much broader approach to automated molecular orbital analysis by transferring the large wave function coefficient matrix into a database and utilizing carefully designed queries for pattern discovery.

3. Platin-based Anticancer Drugs. A second area of intensive collaboration with the Lippard lab is concerned with understanding the interaction of anticancer drug cisplatin with its primary cellular target, DNA. We have addressed a number of important questions in the past⁷¹⁻⁷³. These studies provided for the first time an intuitive and detailed concept of how cisplatin interacts with purine bases and explained the dominating preference of guanine over adenine as the primary binding site. Since Baik's arrival at Indiana University, the last step towards a complete molecular level understanding of the platination event has been completed using a combinatorial computational approach coupled to a novel molecular orbital descriptor.¹³⁸ In the last 30 years of intensive research towards finding cisplatin alternatives, only two Pt-based drugs, carboplatin and oxaliplatin have been FDA-approved. With the novel concepts for cisplatin binding that we discovered in the last few years at hand, we will examine in detail how these alternative Pt-drugs differ in their reactivity towards DNA and which features are identical in all three anticancer drugs. In addition, we will generate a library of Pt-drug candidates that have already been examined in the Lippard lab and identified to be "failure cases". For training our discovery tools, such "failure cases" represent strongly inferential data. By computationally reexamining these cases, we will generalize the concepts that we have established before. We anticipate profiling a few hundred Pt-drug candidates at the least, which will generate a broad database of Pt-compounds that will allow for a clustering analysis. The tools needed for this project must provide means of enumerating a relatively large number of computed and experimentally probed Pt-molecules to discover possible trends of inactivity. A number of possible reasons for failure will have to be examined automatically, including, decreased binding affinities to the cellular target or potential side reactions that give rise to a sharp decrease of the drug's reactivity.

4. Alzheimer's Disease. Despite substantial efforts from many research groups, the pathophysiological basis of Alzheimer's disease (AD) is currently not well understood. There is desperate and increasing need for a treatment of this disease, as higher standards of living allowed the life-expectancy of the average American to increase notably, which also increased the group of potential AD patients at the same time. Currently, there is no treatment available publicly and as there is no fundamental understanding of the disease, rational searches for a drug or possible therapies are difficult and vague at best.

There is general agreement that the formation of neuritic plaques composed of β -amyloid complexes of Cu play a major role.¹³⁹ But there is no widely accepted explanation for why and how the formation of the metal-protein complex gives rise to neurodegeneration. The most plausible proposal from a chemist's viewpoint is the hypothesis that the Cu center surrounded by certain residues becomes an efficient catalyst for the generation of hydrogen peroxide and other reactive oxygen species consuming biological reducing agent, such as cholesterol, vitamin C or dopamine.⁷⁶ It is simple to understand that the massive release of these highly reactive species would ultimately lead to cell death and apoptosis. A straightforward prevention of the plaque formation would be to use molecules that will more tightly bind to Cu and effectively remove it before Cu/ β -amyloid complexes can be formed. The success of this method is unclear on the onset, as there is no way of foreseeing the effects of substantial removal of Cu from the cellular space. In addition, such a "drug" would only be useful as a preventive measure, because it will not be able to intervene once plaques are formed.

A more promising and targeted effort towards developing therapeutics requires a molecular level understanding of the catalytic mechanism that generates the reactive oxygen species. Ideally, we would like to identify which portion of the catalysis is most vulnerable to disruption by small molecules with a set of designed properties. Obtaining such knowledge and detailed view of a chemical reaction from traditional research methods is nearly impossible. We are using computer simulations to gain insights into the structural and electronic features that govern the chemical reaction at the Cu center. As the exact structure of the Cu-peptide complex is not known, we must sample a large number of possible structures, compute experimentally

observable spectroscopic signatures, which will all be stored in the new database and compared against experimentally available data to guide the construction of a realistic model. We will also carry out virtual combinatorial screening studies to broaden the foundation for possible experimental searches for agents that might disrupt the catalytic cycle.

5. Catalysts for natural product syntheses. The pursuit for new drugs often involves studying bacteria and plants, which serve as rich sources for complicated organic molecules that often display promising therapeutic activities. Isolating and identifying natural products with desirable biological functions are difficult. At least equally daunting is research that is needed to find a synthetic path to these natural products. The ability to synthesize a biologically active molecule in a laboratory setting and later in an industrial production scale is key for making a new drug publicly accessible. Being able to synthesize the molecule in a laboratory also allows for variations of the original molecule in pursuit of an optimal efficacy, overcoming limits of having to rely on expensive extractions from natural sources. One such natural product is Taxol¹⁴⁰, a complex polyoxygenated diterpene molecule isolated from the Pacific Yew, *Taxus brevifolia*, which was discovered in the late 60s and has been FDA-approved for breast cancer treatment in 1994. Plant growth alone is unable to supply the required amounts of Taxol, and industrial synthesis is essential. One of the main challenges in natural product synthesis is the fact that natural products are often highly decorated with chiral stereocenters. Two molecules that are identical in their composition, but are mirror images from each other are called enantiomers. Usually, the desirable biological function of natural products is directly associated with one enantiomer and the wrong enantiomer is often a potent toxin leading to catastrophic consequences if given to patients. In the case of Taxol, the challenge lies in assembling an eight-membered ring structure¹⁴¹ in a precise manner. Decades of dedicated research from leading organic synthesis laboratories¹⁴²⁻¹⁵³ led to a number of synthetic pathways that are all tantalizingly complicated and low yielding. Recently, transition metal catalysts have begun to revolutionize the field asymmetric synthesis (the selective preparation of one enantiomer), as they allow for the assembly of structural motifs in regio- and stereoselective fashion with an unprecedented ease and versatility. Transition metal catalysts already play a pivotal role in organic synthesis and will likely become more important in the future. One of the main problems holding back progress is the difficulty of deriving detailed mechanistic knowledge from experimental data. In addition, the transition metal catalysts in use today in organic laboratories are primitive in the sense that they are usually not rationally designed, but discovered through a tedious trial-and-error screening and/or by accidents. Often the details of how they work are not known and it is usually very difficult to find rational catalyst optimization strategies. Catalyst discovery for important organic reactions is an exciting research field where future combination of rigorous quantum chemical simulation backed by modern knowledge mining and management infrastructure promises rapid progress. For such a progress to be made, we must rely on seamless teamwork and alliance between the Cheminformatics research group, experimental organic chemists who can incorporate the novel rational strategies in new experiments and experimental inorganic chemists who can provide the next generation of better designed catalysts. We have assembled such a three-way collaborative network between the PI's of this proposal, the organic laboratory of Professors P. Andrew Evans, David R. Williams and Jeffrey Johnston and the organometallic/inorganic laboratories of Profs. Daniel J. Mindiola and Kenneth G. Caulton (all IU).

(i) We have demonstrated the synergistic benefit of collaborative work between Cheminformatics and synthetic chemistry in a very productive relationship between us and the group of Profs. Daniel J. Mindiola^{154,155} and Kenneth G. Caulton.¹⁵⁶ Over the course of the last 15 months, we have successfully completed a number of important studies on transition metal complexes that represent a systematic exploration of the next generation of catalysts¹⁵⁵⁻¹⁵⁷.

(ii) We have started to examine the utilization of Rh catalysts for the stereoselective assembly of eight-membered rings in collaboration with the Evans' lab¹⁵⁸.

(iii) In addition to the above mentioned collaborative work, we continue to study fundamental mechanistic aspects of important chemical transformations and address electronic structure questions of high relevance to catalysis, which resulted in a number of publications^{157,159-163}.

In these studies, we explored important organic and inorganic reaction mechanisms and catalysts with the goal of establishing firm and diverse bases for our informatics approach to reactive chemistry discovery. We

will continue to work with our experimental collaborators towards the goal of devising new robust catalysts for stereoselective organic synthesis and increase our efforts in examining transition metal complexes with unique electronic properties that serve to broaden the applicability of our data mining tools while deriving bonding concepts that will immediately guide the experimental transition metal complex community in discovering new catalysts.

SUMMARY

The eventual Cheminformatics Research Center at Indiana University needs to be robust enough to allow for new collaborative research among scientists of all stripes and access from non-traditional chemistry areas, such as forensic chemistry. One of the goals of the CRC would be to develop easy-to-use software for researchers to access the databases and associated search technologies from their desktops. This is in line with our long-term vision to provide freely available cheminformatics tools and techniques in an open source cheminformatics tool repository. The various projects share a common theme of making the cheminformatics arsenal available to advance scientific research in the chemical and biomedical areas.

We will have excellent communications links to tie together the various participants in the projects and to form a synergistic basis for the creation of our Chemistry Virtual Laboratory. Regular updates using blogs, Wikis, and other modern communication techniques will keep all participants, including external advisors, aware of the progress of the projects. Furthermore, we will utilize the outstanding teleconference and Access Grid facilities at IU to maintain close contact with all pertinent people. Frequent meetings of all major players will ensure that all projects are on target and that we are incrementally formulating a plan for transitioning to a P50 center. The initial workshop will give an opportunity for all participants to get to know each other, and the follow-up meeting in the second year will ensure that we are moving in a coherent and cohesive direction on all fronts. It is essential to have face-to-face meetings in the early stages of the project, so that all participants can bond.

Indiana University firmly believes in the concept of the Cheminformatics Research Center, and we have obtained a promise from the Dean of the School of Informatics that he will return to the project \$70,000 in overhead funds during the two-year period of the grant. This will be used to bring the tuition subsidy for the graduate students to 100% and for other expenses associated with the project.

Substantial expertise exists within and outside the School of Informatics on distributed computing and grid capabilities, metadata tagging for more effective and widespread collaborative use of data, and laboratory informatics, computational chemistry and modeling. We have been careful to build links to industry and to relevant academic disciplines, as well as to include on our team a renowned expert in human computer interaction design, Dr. Yvonne Rogers. On the traditional chemical informatics side, the suggested projects concerning High Throughput Screening are very relevant, and we firmly believe that we can provide excellent results in these and other areas covered by the prototype projects of the CICC. Data will be of maximal benefit to public health if they are treated as a community resource and made publicly available, so we are actively pursuing ideas such as institutional repositories for the easy sharing of data and electronic services.

Cheminformatics is very strongly tied to the pharmaceutical industry, in large part because that is where the bulk of the research money is located to take advantage of the tools and techniques that cheminformatics offers. Our vision should ultimately be broader than that, particularly stressing the desire to introduce cheminformatics into the academic world and (eventually) to popularize the field among other segments of the chemical industry beyond the pharmaceutical sector. Thus, if we were to develop a mission statement for a full Cheminformatics Research Center at Indiana University, it might be:

MISSION: to build on the major advances in chemical informatics in the past decades by expanding the knowledge base of the field and championing the use of chemical informatics tools and techniques in both industrial and academic life science and chemistry research and teaching through appropriate use of Grid computing technology.

The National Institutes of Health, through its Molecular Libraries and Imaging Initiative, is laying down a firm foundation for the infrastructure and data input that will allow substantial repositories of chemical data to be available at no charge. We have identified areas where we can develop research and instructional tools that will make the use, analysis, and interpretation of those data easier to incorporate into life sciences and chemistry research and teaching projects. We plan to build on the knowledge and expertise of the key programs and interests that already exist at IU, both at the Bloomington campus and at the Indianapolis campus.

The subprojects described in our proposal are quite varied. The work will involve feasibility testing and evaluation of approaches, but working models should be created in all cases during the ECCR phase. We expect that fully functioning tools and populated databases would emerge in a CRC stage. In the long term, cheminformatics education will be needed from high school upward, so the bulk of the educational activities will occur in the full Center phase, when a coordinated program can be developed to satisfy a broader base of students.

Among the criteria for the establishment of an NIH research center are the study of important problems; the need for core resources, where sharing of resources provides economies of scale; attraction of a sufficient number of investigators; identification of strategic focus; the need for interdisciplinary interaction; the need to provide distinctive training environments; the assembly of multi-investigator teams; and the cross-disciplinary or translational research training of students, postdoctoral fellows, and others.¹⁶⁴ Most of those features were incorporated in our plans for the CICC, and they will be elaborated on in a CCR proposal.

E. Literature Cited.

- (1) Foster, I.; Kesselman *The Grid: Blueprint for a New Computing Infrastructure.*; 2nd ed.; Elsevier: Amsterdam, 2004.
- (2) Berman, F.; Fox, G.; Hey, T. *Grid Computing: Making the Global Infrastructure a Reality*; Wiley: Chichester, England, New York, 2003.
- (3) Parashar, M.; Lee, C. A., Special Issue on Grid Computing., *Proceedings of the IEEE* **2005**, 93, passim.
- (4) Indiana University School of Informatics Bloomington: <http://www.informatics.indiana.edu> and Indianapolis: <http://www.informatics.iupui.edu>.
- (5) Hey, T.; Trefethen, A., The data deluge: an e-Science perspective, In *Grid Computing: Making the Global Infrastructure a Reality.*; Berman, F., Fox, G., Hey, A. J. G., Eds.; Wiley: New York, 2003, p Ch. 36 <http://www.grid2002.org/>.
- (6) Atkins, D. E. e. a. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.*; National Science Foundation CISE: Arlington, VA, January 2003. <http://www.nsf.gov/cise/sci/reports/toc.jsp>.
- (7) National e-Science Centre <http://www.nesc.ac.uk/>.
- (8) Fox, G.; Walker, D. *e-Science Gap Analysis*; Report UKeS-2003-01, http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/index.html, June 30 2003.
- (9) GridChem <http://www.ncsa.uiuc.edu/Projects/AllProjects/Projects62.html>.
- (10) Combechem. <http://www.combechem.org/>.
- (11) Collaboratory for Multi-Scale Chemical Science (CMCS) <http://cmcs.ca.sandia.gov/>.
- (12) caBIG <http://cabig.nci.nih.gov/>.
- (13) BIRN; Biomedical Informatics Research Network <http://www.nbirn.net/>.
- (14) myGRID <http://www.mygrid.org.uk/>.
- (14b) EGEE; Enabling Grids for E-Science <http://public.eu-egee.org/>
- (14c) OSG; Open Science Grid http://osg.grid.iu.edu/OSG/index.php?option=com_frontpage&elMenu=Home
- (15) Richards, W. G., Virtual screening using grid computing: The screensaver project, *Nature Reviews Drug Discovery* **2002**, 1, 551-555.
- (16) Richards, W. G.; Grant, G. H.; Harrison, K. N., Combating bioterrorism with personal computers, *Journal of Molecular Graphics and Modeling* **2004**, 22, 473-478.

- (17) Folding@Home <http://folding.stanford.edu/>.
- (18) The UK OGSA Evaluation Project <http://sse.cs.ucl.ac.uk/UK-OGSA/>.
- (19) The WS-Resource Framework <http://www.globus.org/wsrfl/>.
- (20) Atkinson, M.; DeRoure, D.; Dunlop, A.; Fox, G.; Henderson, P.; Hey, T.; Paton, N.; Newhouse, S.; Parastatidis, S.; Trefethen, A.; Watson, P., Web Service Grids: An Evolutionary Approach, *UK e-Science Technical Report. July 13 2004 Special Issue on Grid Architecture of Concurrency&Computation: Practice and Experience* **2005**, *17*, 377-389.
- (21) The Globus Alliance <http://www.globus.org/>.
- (22) Open Middleware Infrastructure Institute <http://www.omii.ac.uk/>.
- (23) Web Services Description Language (WSDL) <http://www.w3.org/TR/wsdl>.
- (24) Simple Object Access Protocol (SOAP) <http://www.w3.org/TR/2001/WD-soap12-20010709/>.
- (25) Global Grid Forum (GGF) <http://www.gridforum.org/>.
- (26) Open Grid Services Architecture Data Access and Integration (OGSA-DAI) <http://www.ogsadai.org.uk/>.
- (27) caBIG cancer Biomedical Informatics Grid <http://cabig.nci.nih.gov/>.
- (28) JSR 168: Portlet Specification <http://www.jcp.org/en/jsr/detail?id=168>.
- (29) OASIS Web Services for Remote Portlets 1.0 Primer. Committee Draft 1.00 03 December 2004. <http://www.oasis-open.org/committees/download.php/10539/wsrp-primer-1.0.html>.
- (30) Murray-Rust, P.; Rzepa, H. S., Chemical Markup, XML, and the World Wide Web. 4. CML schema, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 757-772.
- (31) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S., NIH Molecular Libraries Initiative, *Science*, *306*, 1138-1139.
- (32) Irwin, J. J. S., Brian K, ZINC-A free database of commercially available compounds for virtual screening, *Journal of Chemical Information and Modeling* **2005**, *45*, 177-182.
- (33) PubChem <http://pubchem.ncbi.nlm.nih.gov/>.
- (34) Mullin, R., Dealing with data overload, *Chemical & Engineering News* **March 22, 2004**, *82*, 19-24.
- (35) Wild, D. J.; Blankley, C. J., VisualiSAR: a web-based application for clustering, structure browsing, and structure-activity relationship study, *Journal of Molecular Graphics and Modelling* **1999**, *17*, 85-89, 120-125.
- (36) Bikker, J. A.; Dunbar, J. B., Jr.; Bornemeier, D.; Wild, D. J.; Calvet, A.; Humblet, C. In *Abstracts of Papers, 222nd American Chemical Society National Meeting*: Chicago, IL, 2001.
- (37) Ertl, P.; Selzer, P.; Muehlbacher, J., Web-based cheminformatics tools deployed via corporate intranets, *Drug Discovery Today: BIOSILICO* **2004**, *2*, 201-207.
- (38) Bikker, op. cit.
- (39) Downs, G. M.; Barnard, J. M., Clustering methods and their uses in computational chemistry, *Reviews in Computational Chemistry* **2002**, *18*, 1-40.
- (40) Rusinko, I., Andrew; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S., Analysis of a large structure/biological activity data set using recursive partitioning, *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 1017-1026.
- (41) Calvet, A. In *2nd Joint Sheffield Conference on Chemoinformatics*: Sheffield, UK., 2001.
- (42) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. L., Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination, *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1295-1300.
- (43) Agrafiotis, D. K.; Lobanov, V. S., Nonlinear mapping networks, *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 1356-1362.
- (44) Bienfat, B., Applications of High-Resolution Self-Organizing Maps to Retrosynthetic and QSAR Analysis, *Journal of Chemical Information and Computer Sciences* **1994**, *34*, 890-898.
- (45) Self-Organizing Map of NCI Antiviral Screen Compounds. http://cactus.nci.nih.gov/services/som_qsar/.
- (46) Teckentrup, A.; Briem, H.; Gasteiger, J., Mining High-Throughput Screening data of combinatorial libraries: Development of a filter to distinguish hits from nonhits, *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 626-634.
- (47) Ahlberg, C., Visual exploration of HTS databases: bridging the gap between chemistry and biology, *Drug Discovery Today* **1999**, *4*, 370-376.
- (48) Spotfire <http://www.spotfire.com>.
- (49) Tripos <http://www.tripos.com>.

- (50) Bioreason www.bioreason.com.
- (51) ChemTK (Sage Informatics) <http://www.sageinformatics.com>.
- (52) Liu, K.; Feng, J.; Young, S. S., PowerMV: A software environment for molecular viewing, descriptor generation, data analysis and hit evaluation, *Journal of Chemical Information and Modeling* **2005**, in press.
- (53) Gribbon, P.; Sewing, A., High-throughput drug discovery: what can we expect from HTS?, *Drug Discovery Today* **2005**, *10*, 17-22.
- (54) Darvas, F.; Keseru, G.; Papp, A.; Dorman, G.; Urge, L.; Krajcsi, P., In Silico and ex silico ADME approaches for drug discovery, *Current Topics in Medicinal Chemistry* **2002**, *7*, 402-408.
- (55) Austin, op. cit.
- (56) Irwin, op. cit.
- (57) Frederick/Bethesda Data and Online Services <http://cactus.nci.nih.gov>.
- (58) Collaboratory for Multi-Scale Chemical Science <http://cmcs.ca.sandia.gov/>.
- (59) Combechem <http://www.combechem.org/>.
- (60) Kohn, W.; Becke, A. D.; Parr, R. G., Density functional theory of electronic structure, *Journal of Physical Chemistry* **1996**, *100*, 12974-12980.
- (61) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G., How enzymes work: Analysis by modern rate theory and computer simulations, *Science* **2004**, *303*, 186-195.
- (62) Russo, E., Chemistry Plans a Structural Overhaul, *Nature* **September 12, 2002**, (6903), (naturejobs, 4-7).
- (63) Wiggins, G. D. In *Abstracts of Papers, 228th ACS National Meeting*, August 22-26, 2004, pp CINF-039.
- (64) Schofield, H.; Wiggins, G.; Willett, P., Recent developments in chemoinformatics education, *Drug Discovery Today* **September 15, 2001**, *6*, 931-934.
- (65) Ph.D. in Informatics at the Indiana University School of Informatics - <http://www.informatics.indiana.edu/academics/phd.asp>.
- (66) Access Grid <http://www.accessgrid.org/>.
- (67) Global MMCS (Global Multimedia Collaboration System) <http://www.globalmmcs.org/>.
- (68) The Sakai Project <http://www.sakaiproject.org/>.
- (69) Banerjee, R., Radical carbon skeleton rearrangements: Catalysis by coenzyme B-12-dependent mutases, *Chemical Reviews* **2003**, *103*, 2083-2094.
- (70) Baik, M. H.; Newcomb, M.; Friesner, R. A.; Lippard, S. J., Mechanistic studies on the hydroxylation of methane by methane monooxygenase, *Chemical Reviews* **2003**, *103*, 2385-2419.
- (71) Baik, M. H.; Friesner, R. A.; Lippard, S. J., Theoretical Study on the Stability of N-Glycosyl Bonds: Why Does N7-Platination Not Promote Depurination?, *Journal of the American Chemical Society* **2002**, *124*, 4495-4503.
- (72) Baik, M. H.; Friesner, R. A.; Lippard, S. J., Theoretical Study of Cisplatin Binding to Purine Bases: Why Does Cisplatin Prefer Guanine Over Adenine as Substrate?, *Journal of the American Chemical Society* **2003**, *125*, 14082-14092.
- (73) Baik, M. H.; Friesner, R. A.; Lippard, S. J., cis-{Pt(NH₃)₂(L)}^{2+/*} (L = Cl, H₂O, NH₃) Binding to Purines and CO: Does π -Backdonation Play a Role?, *Inorganic Chemistry* **2003**, *42*, 8615-8617.
- (74) Barnham, K. J.; Haeffner, F.; Ciccotosto, G. D.; Curtain, C. C.; Tew, D.; Masters, C. L.; Chemy, R. A.; Cappai, R.; Bush, A. I., The mechanism by which A beta produces hydrogen peroxide and its role in neurotoxicity, *Neurobiology of Aging* **2004**, *25*, S544-S544.
- (75) Barnham, K. J.; Haeffner, F.; Ciccotosto, G. D.; Curtain, C. C.; Tew, D.; Mavros, C.; Beyreuther, K.; Carrington, D.; Masters, C. L.; Cherny, R. A.; Cappai, R.; Bush, A. I., Tyrosine gated electron transfer is key to the toxic mechanism of Alzheimer's disease beta-amyloid, *Faseb Journal* **2004**, *18*.
- (76) Opazo, C.; Huang, X. D.; Cherny, R. A.; Moir, R. D.; Roher, A. E.; White, A. R.; Cappai, R.; Masters, C. L.; Tanzi, R. E.; Inestrosa, N. C.; Bush, A. I., Metalloenzyme-like activity of Alzheimer's disease beta-amyloid - Cu-dependent catalytic conversion of dopamine, cholesterol, and biological reducing agents to neurotoxic H₂O₂, *Journal of Biological Chemistry* **2002**, *277*, 40302-40308.
- (77) LEAD Project <http://lead.ou.edu/>.
- (78) SERVO Grid (Solid Earth Research Virtual Observatory Grid) <http://www.servogrid.org/>.
- (79) QuakeSim Project <http://quakesim.jpl.nasa.gov/>.

- (80) Brown, R. D.; Martin, Y. C., Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 572-584.
- (81) Wild, D. J.; Blankley, C. J., Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering, *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 155-162.
- (82) Stahura, F. L.; Bajorath, J., Virtual screening methods that complement HTS., *Combinatorial Chemistry and High Throughput Screening* **2004**, *7*, 259-269.
- (83) Barnard Chemical Information Ltd. <http://www.bci.gb.com>.
- (84) Wild, D. J., New techniques for clustering and analyzing large volumes of chemical information, *ACS 36th Central Regional Meeting June 2, 2004*.
- (85) Shoichet, B., Virtual screening of chemical libraries, *Nature* **2004**, *432*(7019), 862-865.
- (86) Willett, P., Similarity-based approaches to virtual screening, *Biochemical Society Transactions* **2003**, *31*, 603-606.
- (87) Fang, X.; Wang, S., A Web-based 3D-database pharmacophore searching tool for drug discovery, *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 192-198.
- (88) Mestres, J.; Knegt, R. M. A., Similarity versus docking in 3D virtual screening, *Perspectives in Drug Discovery and Design* **2000**, *20*, 191-207.
- (89) Richards, *op. cit.* **2002**.
- (90) Richards, *op. cit.* **2004**.
- (91) Cheminformatics at Sheffield <http://cisrg.shef.ac.uk/>.
- (92) Cheminformatics MSc at the University of Manchester <http://www.manchester.ac.uk/degreeprogrammes/postgraduate/taught/1107.htm>.
- (93) Computer Chemistry Center, University of Erlangen-Nuremberg <http://www.ccc.uni-erlangen.de/>.
- (94) Indiana University School of Informatics <http://www.informatics.indiana.edu>.
- (95) QCPE <http://qcpe.chem.indiana.edu/>.
- (96) Reciprocal Net <http://www.reciprocalnet.org/>.
- (97) CHEMINFO <http://www.indiana.edu/~cheminfo/>.
- (98) Serena Software's PCMODEL <http://www.serenasoft.com/pcm8.html>.
- (99) Biotech <http://biotech.icmb.utexas.edu/> (NB: The project moved to the University of Texas after all grant and external support was exhausted. It has not been further developed since that time.).
- (100) Jourdan, D.; Ellington, A. D.; Wiggins, G. D. In *Book of Abstracts, 210th ACS National Meeting*: Chicago, IL, August 20-24, 1995, pp COMP-044.
- (101) IU SLIS's Chemical Information Specialization and Master of Information Science Program http://www.slis.indiana.edu/degrees/joint/cheminfo_mis.html.
- (102) Center for Genomics and Bioinformatics (Bloomington) <http://cgb.indiana.edu/>.
- (103) Center for Computational Biology and Bioinformatics (Indianapolis) <http://compbio.iupui.edu/>.
- (104) School of Informatics at IUPUI Laboratory Informatics Program <http://informatics.iupui.edu/i/36>.
- (105) C571 Chemical Information Technology <http://www.indiana.edu/~cheminfo/C571/571home.html> The ONCOURSE site was used for Fall 2004: <http://oncourse.iu.edu>.
- (106) Semantic Web <http://www.w3.org/2001/sw/>.
- (107) Scientific Annotation Middleware (SAM) <http://collaboratory.emsl.pnl.gov/docs/collab/sam/>.
- (108) Wild; Blankley, *Op. cit.*
- (109) Stahura; Bajorath, *Op. cit.*
- (110) Crampin, E. J.; Schnell, S.; McSharry, P. E., The mathematical and computational techniques to deduce complex biochemical reaction mechanisms, *Progress in Biophysics and Molecular Biology* **2004**, *77*-112.
- (111) Crampin, E. J.; P.E., M.; Schnell, S., Extracting biochemical reaction kinetics from time series data, *Lecture Notes in Artificial Intelligence* **2004**, *3214*, 329-336.
- (112) Noble, D., The rise of computational biology, *Nature Reviews Molecular and Cellular Biology* **2002**, *3*, 459-463.
- (113) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer: Dordrecht, 2003.
- (114) Gasteiger, J.; Engel, T., Eds. *Chemoinformatics; A Textbook.*; Wiley-VCH: Weinheim, 2003.

- (115) Gasteiger, J., Ed. *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*; Wiley-VCH: Weinheim, 2003.
- (116) Bajorath, J., Ed. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, Humana Press: Totowa, NJ, 2004.
- (117) Noordik, J. H., Ed. *Cheminformatics Developments: History, Reviews and Current Research*; IOS Press B.V.: Amsterdam, 2004.
- (118) Wiggins, G., CHMINF-L: the chemical information sources discussion list, *Journal of the American Society for Information Science* **1995**, *46*, 614-617.
- (119) Clearinghouse for Chemical Information Instructional Materials
<http://www.indiana.edu/~cheminfo/cciiimnro.html>.
- (120) Murray-Rust, P.; Rzepa, H.; Williamson, M. J.; Willighagen, E. L., Chemical Markup, XML, and the World Wide Web 5. Applications of chemical metadata in RSS aggregators, *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 462-469.
- (121) Banerjee, R., Radical carbon skeleton rearrangements: Catalysis by coenzyme B-12-dependent mutases, *Chem. Rev.* **2003**, *103*, 2083-2094.
- (122) Smith, D. M.; Golding, B. T.; Radom, L., Understanding the mechanism of B-12-dependent methylmalonyl-CoA mutase: Partial proton transfer in action, *J. Am. Chem. Soc.* **1999**, *121*, 9388-9399.
- (123) Smith, D. M.; Golding, B. T.; Radom, L., Facilitation of enzyme-catalyzed reactions by partial proton transfer: Application to coenzyme-B-12-dependent methylmalonyl-CoA mutase, *J. Am. Chem. Soc.* **1999**, *121*, 1383-1384.
- (124) Smith, D. M.; Golding, B. T.; Radom, L., On the mechanism of action of vitamin B-12: Theoretical studies of the 2-methyleneglutarate mutase catalyzed rearrangement, *J. Am. Chem. Soc.* **1999**, *121*, 1037-1044.
- (125) Smith, D. M.; Golding, B. T.; Radom, L., Toward a consistent mechanism for diol dehydratase catalyzed reactions: An application of the partial-proton-transfer concept, *J. Am. Chem. Soc.* **1999**, *121*, 5700-5704.
- (126) Wetmore, S. D.; Smith, D. M.; Radom, L., How B-6 helps B-12: The roles of B-6, B-12, and the enzymes in aminomutase-catalyzed reactions, *J. Am. Chem. Soc.* **2000**, *122*, 10208-10209.
- (127) Smith, D. M.; Golding, B. T.; Radom, L., Understanding the mechanism of B-12-dependent diol dehydratase: A synergistic retro-push-pull proposal, *J. Am. Chem. Soc.* **2001**, *123*, 1664-1675.
- (128) Wetmore, S. D.; Smith, D. M.; Golding, B. T.; Radom, L., Interconversion of (S)-glutamate and (2S,3S)-3-methylaspartate: A distinctive B-12-dependent carbon-skeleton rearrangement, *J. Am. Chem. Soc.* **2001**, *123*, 7963-7972.
- (129) Wetmore, S. D.; Smith, D. M.; Radom, L., Enzyme catalysis of 1,2-amino shifts: The cooperative action of B-6, B-12, and aminomutases, *J. Am. Chem. Soc.* **2001**, *123*, 8678-8689.
- (130) Wetmore, S. D.; Smith, D. M.; Radom, L., Catalysis by mutants of methylmalonyl-CoA mutase: A theoretical rationalization for a change in the rate-determining step, *Chembiochem* **2001**, *2*, 919-922.
- (131) Wetmore, S. D.; Smith, D. M.; Bennett, J. T.; Radom, L., Understanding the mechanism of action of beta(12)-dependent ethanolamine ammonia-lyase: Synergistic interactions at play, *J. Am. Chem. Soc.* **2002**, *124*, 14054-14065.
- (132) Golding, B. T.; Radom, L., Facilitation of intramolecular 1,2-shifts in radicals by protonation, and the mechanism of reactions catalyzed by 5'-deoxyadenosylcobalamin, *J. Chem. Soc., Chem. Commun.* **1973**, 939-941.
- (133) Newcomb, M.; Miranda, N., Kinetic results implicating a polar radical reaction pathway in the rearrangement catalyzed by alpha-methyleneglutarate mutase, *J. Am. Chem. Soc.* **2003**, *125*, 4080-4086.
- (134) Daublain, P.; Horner, J. H.; Kuznetsov, A.; Newcomb, M., Solvent polarity effects and limited acid catalysis in rearrangements of model radicals for the methylmalonyl-CoA mutase- and isobutyryl-CoA mutase-catalyzed isomerization reactions, *J. Am. Chem. Soc.* **2004**, *126*, 5368-5369.
- (135) Lovell, T.; Li, J.; Noodleman, L., Density functional and electrostatics study of oxidized and reduced ribonucleotide reductase; comparisons with methane monooxygenase, *Journal of Biological Inorganic Chemistry* **2002**, *7*, 799-809.

- (136) Guallar, V.; Baik, M. H.; Lippard, S. J.; Friesner, R. A., Peripheral heme substituents control the hydrogen-atom abstraction chemistry in cytochromes P450, *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 6998-7002.
- (137) Baik, M.-H.; Gherman, B. F.; Friesner, R. A.; Lippard, S. J., Hydroxylation of Methane by Non-Heme Diiron Enzymes: Molecular Orbital Analysis of C–H Bond Activation by Reactive Intermediate Q, *Journal of the American Chemical Society* **2002**, *124*, 14608-14615.
- (138) Sandvoss, L.; Baik, M.-H., A combinatorial DFT study of how cisplatin binds to purine bases, *Abstracts of Papers, 227th ACS National Meeting, Anaheim, CA, United States, March 28-April 1, 2004* **2004**, CINF-053.
- (139) Sipe, J. D., Amyloidosis, *Annual Review of Biochemistry* **1992**, *61*, 947-975.
- (140) Nicolaou, K. C.; Dai, W. M.; Guy, R. K., Chemistry and Biology of Taxol, *Angewandte Chemie-International Edition* **1994**, *33*, 15-44.
- (141) Mehta, G.; Singh, V., Progress in the construction of cyclooctanoid systems: New approaches and applications to natural product syntheses, *Chemical Reviews* **1999**, *99*, 881-930.
- (142) Holton, R. A.; Somoza, C.; Kim, H. B.; Liang, F.; Biediger, R. J.; Boatman, P. D.; Shindo, M.; Smith, C. C.; Kim, S. C.; Nadizadeh, H.; Suzuki, Y.; Tao, C. L.; Vu, P.; Tang, S. H.; Zhang, P. S.; Murthi, K. K.; Gentile, L. N.; Liu, J. H., First Total Synthesis of Taxol .1. Functionalization of the B-Ring, *Journal of the American Chemical Society* **1994**, *116*, 1597-1598.
- (143) Holton, R. A.; Kim, H. B.; Somoza, C.; Liang, F.; Biediger, R. J.; Boatman, P. D.; Shindo, M.; Smith, C. C.; Kim, S. C.; Nadizadeh, H.; Suzuki, Y.; Tao, C. L.; Vu, P.; Tang, S. H.; Zhang, P. S.; Murthi, K. K.; Gentile, L. N.; Liu, J. H., First Total Synthesis of Taxol .2. Completion of the C-Ring and D-Ring, *Journal of the American Chemical Society* **1994**, *116*, 1599-1600.
- (144) Nicolaou, K. C.; Yang, Z.; Liu, J. J.; Ueno, H.; Nantermet, P. G.; Guy, R. K.; Claiborne, C. F.; Renaud, J.; Couladouros, E. A.; Paulvannan, K.; Sorensen, E. J., Total Synthesis of Taxol, *Nature* **1994**, *367*, 630-634.
- (145) Masters, J. J.; Link, J. T.; Snyder, L. B.; Young, W. B.; Danishefsky, S. J., A Total Synthesis of Taxol, *Angewandte Chemie-International Edition in English* **1995**, *34*, 1723-1726.
- (146) Nicolaou, K. C.; Nantermet, P. G.; Ueno, H.; Guy, R. K.; Couladouros, E. A.; Sorensen, E. J., Total Synthesis of Taxol .1. Retrosynthesis, Degradation, and Reconstitution, *Journal of the American Chemical Society* **1995**, *117*, 624-633.
- (147) Nicolaou, K. C.; Liu, J. J.; Yang, Z.; Ueno, H.; Sorensen, E. J.; Claiborne, C. F.; Guy, R. K.; Hwang, C. K.; Nakada, M.; Nantermet, P. G., Total Synthesis of Taxol .2. Construction of a-Ring and C-Ring Intermediates and Initial Attempts to Construct the Abc Ring-System, *Journal of the American Chemical Society* **1995**, *117*, 634-644.
- (148) Nicolaou, K. C.; Yang, Z.; Liu, J. J.; Nantermet, P. G.; Claiborne, C. F.; Renaud, J.; Guy, R. K.; Shibayama, K., Total Synthesis of Taxol .3. Formation of Taxols Abc Ring Skeleton, *Journal of the American Chemical Society* **1995**, *117*, 645-652.
- (149) Danishefsky, S. J.; Masters, J. J.; Young, W. B.; Link, J. T.; Snyder, L. B.; Magee, T. V.; Jung, D. K.; Isaacs, R. C. A.; Bornmann, W. G.; Alaimo, C. A.; Coburn, C. A.; DiGrandi, M. J., Total synthesis of baccatin III and taxol, *Journal of the American Chemical Society* **1996**, *118*, 2843-2859.
- (150) Shiina, I.; Iwadare, H.; Sakoh, H.; Tani, Y.; Hasegawa, M.; Saitoh, K.; Mukaiyama, T., Stereocontrolled synthesis of the BC ring system of taxol, *Chemistry Letters* **1997**, 1139-1140.
- (151) Wender, P. A.; Badham, N. F.; Conway, S. P.; Floreancig, P. E.; Glass, T. E.; Granicher, C.; Houze, J. B.; Janichen, J.; Lee, D. S.; Marquess, D. G.; McGrane, P. L.; Meng, W.; Mucciario, T. P.; Muhlebach, M.; Natchus, M. G.; Paulsen, H.; Rawlins, D. B.; Satkofsky, J.; Shuker, A. J.; Sutton, J. C.; Taylor, R. E.; Tomooka, K., The pinene path to taxanes .5. Stereocontrolled synthesis of a versatile taxane precursor, *Journal of the American Chemical Society* **1997**, *119*, 2755-2756.
- (152) Wender, P. A.; Badham, N. F.; Conway, S. P.; Floreancig, P. E.; Glass, T. E.; Houze, J. B.; Krauss, N. E.; Lee, D. S.; Marquess, D. G.; McGrane, P. L.; Meng, W.; Natchus, M. G.; Shuker, A. J.; Sutton, J. C.; Taylor, R. E., The pinene path to taxanes .6. A concise stereocontrolled synthesis of taxol, *Journal of the American Chemical Society* **1997**, *119*, 2757-2758.
- (153) Shiina, I.; Iwadare, H.; Sakoh, H.; Hasegawa, M.; Tani, Y.; Mukaiyama, T., A new method for the synthesis of baccatin III, *Chemistry Letters* **1998**, 1-2.

- (154) Basuli, F.; Bailey, B. C.; Brown, D.; Tomaszewski, J.; Huffman, J. C.; Baik, M.-H.; Mendiola, D. J., Terminal Vanadium-Neopentylidyne Complexes and Intramolecular Cross-Metathesis Reactions to Generate Azametacyclohexatrienes, *Journal of the American Chemical Society* **2004**, *126*, 10506-10507.
- (155) Basuli, F.; Bailey, B. C.; Huffman, J. C.; Baik, M. H.; Mendiola, D. J., Terminal and four-coordinate vanadium(IV) phosphinidene complexes. A pseudo Jahn-Teller effect of second order stabilizing the V-P multiple bond, *Journal of the American Chemical Society* **2004**, *126*, 1924-1925.
- (156) Ozerov, O. V.; Watson, L. A.; Pink, M.; Baik, M.-H.; Caulton, K. G., Terminal Acetylenes React to Increase Unsaturation in [(tBu₂PCH₂SiMe₂)₂N]Re(H)₄, *Organometallics* **2004**, *23*, 4934-4943.
- (157) Burland, M. C.; Meyer, T. Y.; Baik, M.-H., Proton as the Simplest of All Catalysts for [2 + 2] Cycloadditions: DFT Study of Acid-Catalyzed Imine Metathesis, *Journal of Organic Chemistry* **2004**, *69*, 6173-6184.
- (158) Baik, M. H.; Baum, E. W.; Burland, M. C.; Evans, P. A., Diastereoselective Intermolecular Rhodium-Catalyzed [4+2+2] Carbocyclization Reactions: Computational and Experimental Evidence for the Intermediacy of a New Metallocycle Intermediate, *Journal of the American Chemical Society* **2005**, *127*, 1602-1603.
- (159) Bhattacharyya, S.; Pink, M.; Baik, M. H.; Zaleski, J. M., A Unique Approach to Metal-Induced Bergman Cyclization: Long Range Eneidyne Activation by Ligand-to-Metal Charge Transfer, *Angewandte Chemie* **2005**, in press.
- (160) Yang, X.; Baik, M. H., Electronic Structure of Water-Oxidation Catalyst [(bpy)₂(OHx)RuORu(OHy)(bpy)₂]²⁺: Weak Coupling Between the Metal-Centers is Preferable Over Strong Coupling, *Journal of the American Chemical Society* **2004**, *126*, 13222-13223.
- (161) Baik, M.-H.; Friesner, R. A.; Parkin, G., Theoretical investigation of the metal-metal interaction in dimolybdenum complexes with bridging hydride and methyl ligands, *Polyhedron* **2004**, *23*, 2879-2900.
- (162) Herbert, B. J.; Baik, M. H.; Green, J. C., Hydrogen transfer between ligands: A density functional study of the rearrangement of M(eta(6)-C₇H₈)(₂) into M(eta(7)-C₇H₇)(eta(5)-C₇H₉) M = Mo, Mo+, Zr, *Organometallics* **2004**, *23*, 2658-2669.
- (163) Gherman, B. F.; Baik, M. H.; Lippard, S. J.; Friesner, R. A., Dioxygen activation in methane monooxygenase: A theoretical study, *Journal of the American Chemical Society* **2004**, *126*, 2978-2990.
- (164) *NIH Extramural Center Programs: Criteria for Initiation and Evaluation*. Committee for Assessment of NIH Centers of Excellence Programs Board on Health Sciences Policy. Frederick J. Manning, Michael McGeary, Ronald Estabrook, Editors. Washington, D.C.: The National Academies Press, 2004.
- (165) Toner, Bernadette. "At BMS, Web Services integrate tiered informatics." *Genome Technology* **2005 (July/August)**, (54), 37.
- (166) Gardner, Stephen P. "Ontologies and semantic data integration." *Drug Discovery Today* **2005 (July)**, 10(14), 1001-1007.

Appendices

- I. Education Data and Programs
- II. Science Informatics Advisory Board
- III. Timeline to Move Toward a Cheminformatics Research Center at Indiana University

Appendix I: Education Data and Programs

Enrollment in Graduate Science Informatics Programs at Indiana University
(Chemical Informatics, Laboratory Informatics, Bioinformatics, Health Informatics)
as of 8/13/05

	MS: Chem	MS: Lab	MS: Bio	MS: Health	PhD: Chem	PhD: Bio	Phd: Health	TOTAL
IUB	3	0	38	0	1	3	0	45
IUPUI	6	15	34	36	0	5	3	99
TOTAL	9	15	72	36	1	8	3	144

A. Undergraduate Students

The Bachelor of Science in Informatics requires a cognate in a subject discipline. For example, with a chemical informatics cognate, a student takes the equivalent of an undergraduate minor in chemistry (15 semester credit hours) and two courses that are undergraduate versions of I571 and I572 (see below) for a total of 18 semester credit hours in the cognate. A minimum of 34 credits of Informatics courses is also required.

B. Graduate Students

1. MS in Chemical Informatics/MS in Bioinformatics

The MS in Chemical Informatics Program admitted its first students at IUB in 2002, with the Bioinformatics program having started a year prior to that. These are 36-hour degrees, with some core informatics classes and core bioinformatics or 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):
 - L519 Bioinformatics: Theory and Application (3 cr. hrs)
 - L529 Bioinformatics in Molecular Biology and Genetics: Practical Applications (4 cr. hrs.).

2. 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 students entered the program in August 2005. The twelve students admitted at IUB include 1 Ph.D. student on the chemical informatics track (the recipient of the Elsevier MDL Excellence in Informatics Fellowship), 3 bioinformatics students, 3 in Human Computer Interaction Design, and the remainder in areas ranging from complex systems to cybersecurity to social informatics.

C. Continuing Non-Degree Students

We have just begun to offer our first graduate certificate program. The Certificate in Chemical Informatics will be awarded to those who complete the following courses:

- I571 Chemical Information Technology (3 cr. hrs.)
- I572 Computational Chemistry and Molecular Modeling (3 cr. hrs.)
- I590 Programming for Science Informatics (3 cr. hrs.)
- I553 Independent Study in Chemical Informatics (3 cr. hrs.)

The certificate is available to on-site and distance education (DE) students. The fall 2005 semester marks the first time that we have offered a course via DE other than point-to-point teleconferencing between Bloomington and Indianapolis. Among the out-of-state students who enrolled for the I571 class are a patent specialist with a pharmaceutical company, a chemist in another pharmaceutical company, and a researcher at a major bio/cheminformatics software company.

Appendix II. Science Informatics Advisory Board (SIAB)

Malorye A. Branca, Senior Informatics Editor, Bio-IT World

Jeremy Frey, Dept. of Chemistry, University of South Hampton

Vance Kershner, President & CEO, LabWare, Inc.

Caroline A. Kovac, General Manager, IBM Healthcare & Life Sciences

Rudy Potenzzone, CambridgeSoft

Lura Powell, President & CEO, Advanced Imaging Technologies

John Reynders, Information Officer, Discovery & Development Informatics, Eli Lilly & Co.

Rick Roberts, Executive Director, Worldwide Head of Informatics Strategy Development, Pfizer, Inc.

Ray Salemme, CEO, Linguagen

Mick Savage, Consultant, Former President & CEO, Molecular Simulations (Now Accelrys)

MaryJo Zaborowski, Senior Vice President and Global Head of Research Informatics, Roche Pharmaceuticals

Appendix III

