AM4: ISMB2006 Tutorial

Title: Chemoinformatics
(Alternative Title: Chemoinformatics, Chemical Genomics, and Drug Discovery)

Topic Area: All the topics below are relevant, roughly equally.

- Database and Data Integration
- Structural Bioinformatics
- Systems Biology (including Pathways and Networks)
- Medical Bioinformatics
- Molecular Simulation and Systems Dynamics
- Machine Learning and Artificial Intelligence

Main Presenter:

- Title: Professor and Director
- Full name: Pierre Baldi
- Affiliation: Institute for Genomics and Bioinformatics and Department of Computer Science
- Mailing Address: School of Information and Computer Sciences, University of California, Irvine Irvine, CA 92697-3435 USA
- Email address: pfbaldi@ics.uci.edu
- Telephone number: +1 949 824-5809 (work) +1 949 735-8221 (cell)
- Fax number: +1 949 824-9813
- Home page URL: www.ics.uci.edu/~pfbaldi

- Extensive teaching experience. Regularly teaches courses in CS and bioinformatics at UCI.
- Extensive tutorial experience, including tutorials at many conferences including ISMB, NIPS, and IJCNN.

Other contributors to the tutorial presentation:
Sanjay J. Swamidass and J. Chen. Both are MD/PhD graduate students in the Baldi laboratory with extensive experience in chemoinformatics and drug screening. They will not present the main tutorial but will contribute to the content, slides, and demonstrations.

50-word abstract: This self-contained tutorial will provide an overview of chemoinformatics, from foundations to state-of-the-art results and challenges. It will cover molecular and reaction data, data structures and the available algorithms for efficiently searching large repositories and annotating or predicting the physical, chemical, and biological properties of compounds and reactions with applications ranging from chemical genomics to drug discovery. The tutorial will leverage analogies and create synergies between bio and chemical informatics.

Tutorial level: The tutorial is introductory in the sense that it is self-contained, but it will also cover advanced topics and present the state-of-the-art, as well as the main open challenges, in chemoinformatics.

Prior knowledge required: No prior knowledge is expected. However basic familiarity with organic chemistry is most desirable. Basic understanding of databases and/or statistics and machine learning is also desirable, but not necessary.

Suitability of this tutorial for ISMB:
This tutorial is timely, interdisciplinary, cutting-edge science, and relevant to several bioinformatics problems. It aims to bring bioinformaticians up to speed with the state-of-the-art in chemoinformatics methods, by exploring similarities and differences between bioinformatics and chemoinformatics. The two main driving forces behind the bioinformatics expansion have been: (1) the development of high throughput methods and the corresponding public availability of large repositories (GenBank, Swissprot, PDB, etc); and (2) the development of search algorithms (BLAST) and related statistical machine learning techniques to analyse the data. Mutatis mutandis, the same is true of chemoinformatics, with the caveats that large repositories of chemical data have started to become available only very recently, over the last two years or so. Many of the basic concepts in bioinformatics (similarity, search, alignments, kernels, data structures, etc) have applications in chemoinformatics. Thus the tutorial is timely and hopes to help bridge the gap and develop synergies between these two disciplines. From a scientific standpoint, chemoinformatics is relevant to bioinformatics since it plays a key role in several applications such as chemical genomics and drug discovery/screening/design applications. Because the tutorial is self-contained, we expect it to benefit and stimulate a broad audience, from students to researchers. The tutorial will also modestly contribute to developing a greater degree of openness in the chemistry and chemoinformatics communities, as well as new synergies between bioinformatics and chemoinformatics.

Profile of Presenter

I have over 20 years of experience in teaching and research (roughly 15 years in bioinformatics, 3 years in chemoinformatics). I have published over 150 scientific articles and four books. *Bioinformatics: the Machine Learning Approach* [MIT Press, Second Edition] has sold over 15,000 copies. I have given many tutorials at ISMB, NIPS, IJCNN and other major conferences. I have recently given invited talks on chemoinformatics at major international conferences, such as IJCNN (International Joint Conference on Neural Networks, Montreal, Canada, August 2005) and GIW (International Conference on Genome Informatics, Yokohama, Japan, December 2005). I regularly teach bioinformatics and statistical machine learning classes at UCI and I am in the process of developing a chemoinformatics curriculum. About half of my group (ca. 5 students and 1 postdoc) are currently working in the area of chemoinformatics and its applications to, for instance, virtual high-throughput screening. Examples of recent publications from my group related to chemoinformatics and drug discovery:

- T. Lin, M. Melgar, S. J. Swamidass, J. Purdon, T. Tseng, G. Gago, D. Kurth, P. Baldi, H. Gramajo, and S. Tsai. Structure-Based Inhibitor Design of AccD5, an Essential acyl-CoA Carboxylase Carboxyltransferase Domain of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences USA*, in press, (2006).
- J. Chen, S. J. Swamidass, Y. Dou, J. Bruand, and P. Baldi. ChemDB: A Public Database of Small Molecules and Related Chemoinformatics Resources. *Bioinformatics*, 21, 4133-4139, (2005).

- S. J. Swamidass, J. Chen, P. Phung, J. Bruand, L. Ralaivola, and P. Baldi. Kernels for Small Molecules and the Prediction of Mutagenicity, Toxicity, and Anti-Cancer Activity. Proceedings of the 2005 Conference on Intelligent Systems for Molecular Biology, ISMB 05. *Bioinformatics*, 21, Supplement 1, i359-368, (2005).

Tutorial Outline:

1 Introduction [20 minutes]
Informatics Differences between Physics, Chemistry, and Biology
Historical Considerations
Focus on Organic Chemistry
Small Molecules
Chemical Space versus Other Spaces
Major Challenges: Data, Search Algorithms, Predictive Algorithms
Quantum Mechanics, Molecular Dynamics, Statistical Machine Learning
Applications: Drug Screening/Design, Systems Biology (Chemical Genomics), Other (Material Science, Polymers, Reaction Discovery, Origin of Life, etc)

2 Data and Databases [25 minutes]
Data Bottleneck: Lack of Public Databases and Datasets
Lack of Public Annotation Projects
Private Databases: ACS, Cambridge, Bielstein
Examples of Small Datasets: solubility, toxicity, NCI cancer inhibition, etc
Public Databases of Compounds: PubChem, ChemBank, ChemDB, Zinc
Public Databases of Chemical Reactions
Specialized Databases (Metabolites, Drugs)
Annotations
Visualization

3 Molecular Representations and Annotations [25 minutes]
1D: SMILES Strings
1D: Fingerprints, Compression
2D: Graph of Bonds
3D: Atomic Coordinates
InChi/IUPAC/common names
Accuracy of Coordinates
Conformers
Isomers
Surfaces
Profiles
Other aspects and models (orbital theory, electron pairs, resonance, homo-lumo, etc)

4 Molecular Similarity and Searches [25 minutes**]**
Graph Isomorphism. "Tree-like" Nature of Small Molecules.
Basic Similarity Measures: Tanimoto, Tversky, MinMax etc
Statistical Distributions of Compounds
Statistical Distributions of Similarity Scores
Hit Significance, Z Scores, Extreme Value Distribution
Fast Search

Profile Search
Substructure Search
Superstructure Search
Filters (Lipinski Rules, etc)

5 Chemical Reactions [20 minutes]
Basic Principles
Models and Representations
Enthalpy and Kinetics
Searching Reactions
Searching Virtual Compounds
Reaction Discovery
Reverse Synthesis
Designing Combinatorial Libraries

6 Machine Learning and Other Predictive Methods [25 minutes]
Recursive Neural Networks, Graphical Models, ILP, Kernels
Spectral Kernels
Classification
Regression
Examples: Prediction of Toxicity, Solubility, LogP, etc
Prediction of 3D Coordinates
Clustering

7 Molecular Docking [20 minutes]
Problems and Approximations
Force Fields and Scoring Functions
Geometrical Searches and Sampling
Optimization
Ranking
Docking Programs (Dock, Glide, ICM, Flex, etc)
Docking Kernels
Differential Docking
Chemical Genomics

8 Applications: Drug Screening/Design [20 minutes]
Docking and Similarity Searches
Additional Filters (e.g. Lipinski's Rules)
Example of Applications

9 Conclusion and Discussion [30 minutes]
Open Challenges
Synergies between Bioinformatics and Chemoinformatics
Discussion