

— Keynote Address —

Chemoinformatics, Drug Design, and Systems Biology

Pierre Baldi

pfbaldi@ics.uci.edu

Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California, Irvine, Irvine, CA, USA 92697-3425

Keywords: small molecules, docking, kernels, similarity

1 Introduction

In spite of its central role between physics and biology, chemistry has remained in a backward state of informatics development compared to its two close relatives. Computers, public databases, and large collaborative projects have become the pervasive hallmark of research in physics and biology. The Human Genome Project, for instance, required collaboration among dozens if not hundreds of scientists across the world. And the resulting human DNA sequence, as well as a wealth of other biological information, are available for anyone to download from public repositories on the Web such as GenBank, Swissprot, the PDB, and PubMed. Virtually every biologist today uses publicly available tools, such as BLAST, to search sequence databases and analyze high-throughput data. Similar observations can be made in physics with large collaborative efforts in, for instance astronomy or high-energy physics. The Web itself was born at CERN, a European consortium with over half a century of history, and the world largest particle physics laboratory. In stark contrast, large collaborative efforts and public databases and software are comparatively absent from chemical research.

1.1 Small Molecules and Chemical Space

To address the backward state of chemoinformatics together with other fundamental problems in bioinformatics, systems biology, and drug design, here we shall focus exclusively on small molecules in organic chemistry. Small molecules with at most a few dozen atoms play a fundamental role in organic chemistry and biology. They can be used as combinatorial building blocks for chemical synthesis [1, 15, 16], as molecular probes for perturbing and analyzing biological systems in chemical genomics and systems biology [6, 16, 18], and for the screening, design, and discovery of useful compounds. These include of course new drugs [9, 11], the majority of which are small molecules. Furthermore, huge arrays of new small molecules can be produced in a relatively short period of time [7, 15].

It is worth comparing small-molecule chemical space to our own astronomical space. Astronomical space contains on the order of 10^{22} stars, roughly 10^{11} galaxies, each containing 10^{11} stars. The number of known small molecules, encountered so far in nature, or synthesized by man, is on the order of 10^7 (the ACS database, the largest chemical database, currently contains 26 million compounds). However, estimates in the literature of the size of the virtual space of small molecules that could be created vary between 10^{18} and 10^{200} , with 10^{60} being currently one of the more [4]. Thus by any of these estimates, chemical space remains rather unexplored and uncharted. A second essential difference between chemical and astronomical space is that chemical space is comparatively easier to travel, both virtually and in reality. Small molecules can be recursively enumerated in silico and

synthesized in vitro from known building blocks and known reactions. Of course we do not mean to imply that chemical synthesis is a trivial matter—it is not. But general guiding principles and tools are available. And would you rather have to synthesize a new small molecule or travel to a new galaxy thousands of light years away from Earth? In short, with 10^{60} enumerable and synthesizable compounds remaining to be explored, it is hard to see how the computer could avoid becoming the chemoscope—e.g. the central tool of future chemical astronomers.

Table 1: Small molecule chemical space versus astronomical space.

	Stars	Compounds
Visited	0-1	10^7
Existing	10^{22}	10^7
Virtual	10^{22} (?)	10^{60}
Travel	Difficult	Easy

1.2 Chemoinformatics Challenges

The key challenge for computational methods then is not traveling through chemical space per se, but rather to be able to focus traveling expeditions in a vast chemical space towards interesting regions, and to be able to recognize interesting stars and galaxies when they are encountered. The notion of what is interesting may vary of course with the task (e.g. drug discovery, reaction discovery, polymer discovery). But at the most fundamental level what is needed are tools to predict the physical, chemical, and biological properties of small molecules and reactions in order to focus searches and filter search results.

Computational methods in chemistry can be organized along a spectrum ranging from Schrodinger equation, to molecular dynamics, to statistical machine learning methods. Quantum mechanical methods, or even molecular dynamics methods, are computationally intensive and do not scale well to large datasets. These methods are best applied to specific questions on focused small datasets. Statistical and machine learning methods are more likely to yield successful approaches for rapidly sifting through large datasets of chemical information.

Because in the absence of large public database and datasets, chemoinformatics is in a state reminiscent of bioinformatics two or three decades ago, it may be productive to adapt the lessons learnt from bioinformatics to chemoinformatics, while maintaining also a perspective on the fundamental differences between these two relatively young interdisciplinary sciences. If this analogy is correct, two key ingredients were essential for unlocking the large-scale development of bioinformatics and the application of modern statistical machine learning methods to biological data [2], data and similarity measures. In bioinformatics, such as Genbank, Swissprot, and the PDB while alignment algorithms have provided robust similarity measures with their fast BLAST implementation becoming the workhorse of the field. Mutatis mutandis, the same is likely to be true in chemoinformatics.

2 Method and Results

2.1 Data, Data Sets, and Annotations

Limited catalogs of small molecules are available in digital format from many vendors across the world, as well as a number of public Web sites. As datasets of small molecules become increasingly available, it is important to develop computational methods to both organize these data in rapidly searchable databases and to extract or predict useful information for each molecule, including its physical, chemical, and biological properties. Conversely, large and well-annotated datasets are essential for developing statistical machine learning methods in chemoinformatics, whether supervised

or unsupervised, including predictive classification, regression, and clustering of small molecules and their properties [13, 14]. Aggregation and organization of datasets of chemical information allows for massive *in silico* processing that would be impractical or even impossible in a traditional experimental setting.

Several parallel efforts have emerged recently to start to address the data bottleneck, including PubChem (<http://pubchem.ncbi.nlm.nih.gov>), the Harvard Chembank [19], UCSF's ZINC [8], and the UCI ChemDB [5]. The UCI ChemDB is a public database containing over 4M compounds as well as a repository of annotated datasets that can be used to develop statistical machine learning methods. Together, these datasets already pose important challenges for both supervised and unsupervised machine learning methods, from clustering to kernel methods [14, 20].

In the longer run, a critical challenge is going to be the annotation of these databases. Can annotation be carried by chemistry laboratories in a concerted way across the world and deposited in a central public repository? Can data gathered over the years by large pharmaceutical companies become public? How much annotation can be derived more or less automatically from the literature using automated information retrieval methods? Can new annotation models be implemented (e.g. using organic chemistry classes to produce annotations)? This area is likely also to produce new challenges for database technology due to the sheer size of chemical space.

Another very important related set of challenges has to do with the creation of a public repository of chemical reactions, currently in progress at UCI. Such a repository is essential for a variety of tasks ranging from reaction discovery, to automatics determination and optimization of synthetic pathways to travel chemical space.

2.2 Similarity Measures, Kernels, and Prediction

Good similarity metrics between compounds, or between reactions, are essential to rapidly search large databases of compounds or reactions. Consider, for instance, a classical drug discovery problem where the starting point is a protein of known structure and perhaps a corresponding ligand. With a good database of small molecules, the discovery process can proceed from both ends. Starting from the protein, one can dock millions of small molecules to the protein *in silico* [17]. In fact, with sufficient computing power, one ought to be able to dock all known small molecules to all proteins with known structure contained in the PDB [3]. Producing such a matrix ought to be a significant goal for systems biology and pharmacology in the coming years. On the other hand, starting from the ligand, one can search the database of small molecules for compounds that are "similar" to the known ligand(s), where similarity can be defined in different ways. In both approaches, additional filters can be used to eliminate molecules that are, for instance, poorly soluble, too flexible, or toxic [5]. Furthermore, *in silico* chemical reactions applied to the molecules in the database can further expand the space of interesting molecules being screened or designed. Thus similarity to a single compound, or a set of compounds, must be defined precisely in ways that can be computed efficiently together with a statistical theory for assessing the significance of the hits (c.f. "the e-values of BLAST").

But similarity is also essential to develop predictive machine learning methods to predict the physical, chemical, and biological properties of compounds. This is not too surprising since, given an annotated training set of molecules (e.g. toxic/non-toxic), the properties of a new molecule ought to be inferred from its similarities to the molecules in the training set. This is precisely the basic idea behind kernels methods, one of the leading methods in machine learning. To our advantage, compounds (and reactions) can be represented in many ways, including 1D SMILES strings, 2D graphs of bonds, and 3D structures. Good kernels can be derived for each one of these representations. Spectral kernels in particular, counting the number of occurrences of each possible substructure, lead to efficient molecular "fingerprints" and similarity measures that are useful both in database searches and statistical machine learning applications [14, 20].

Additional similarity measures and kernels can be derived based on molecular surfaces (2.5D),

pharmacophores (3D), and even beyond (4D) using conformers, isomers, or dynamic evolution. Challenges remain in developing and testing these similarity measures, their complementary values and usages, as well as their statistical properties and extreme value distributions.

2.3 Docking and Drug Screening

Finally, small molecules can be used for docking and drug screening/discovery. Small molecules, as well as their synthetic derivatives, can be docked to a protein target and computationally filtered (e.g. by solubility) to produce a ranked list of candidates that can then be tested in the laboratory. Known ligands can also be used in similarity searches, or as scaffold for further molecular engineering. We will present several recent drug discovery efforts that leverage ChemDB and the computational tools described above. In particular, we will describe the discovery of several compounds that can bind to the Carboxyltransferase domain of Acyl-CoA Carboxylase, AccD5 from *Mycobacterium tuberculosis*, a new TB therapeutic target [10].

Acknowledgments

Work supported in part by grants from the NIH, NSF, and a Laurel Wilkening Faculty Innovation Award.

References

- [1] Agrafiotis, D.K., Lobanov, V.S., and Salemme, F.R., Combinatorial informatics in the post-genomics era, *Nature Reviews Drug Discovery*, 1:337–346, 2002.
- [2] Baldi, P. and Brunak, S., *Bioinformatics: The Machine Learning Approach*, Second edition, MIT Press, 2001.
- [3] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., The protein data bank, *Nucl. Acids Res.*, 28:235–242, 2000.
- [4] Bohacek, R.S., McMartin, C., and Guida, W.C., The art and practice of structure-based drug design: A molecular modelling perspective, *Medicinal Research Reviews*, 16(1):3–50, 1996.
- [5] Chen, J., Swamidass, S.J., Dou, Y., Bru, J., and Baldi, P., ChemDB: A public database of small molecules and related chemoinformatics resources, *Bioinformatics*, 2005, In Press.
- [6] Dobson, C.M., Chemical space and biology, *Nature*, 432:824–828, 2004.
- [7] Houghten, R.A., Parallel array and mixture-based synthetic combinatorial chemistry: tools for the next millenium. *Annual Review of Pharmacology and Toxicology*, 40:273–282, 2000.
- [8] Irwin, J.J. and Shoichet, B.K., ZINC—A free database of commercially available compounds for virtual screening, *Journal of Chemical Information and Computer Sciences*, 45:177–182, 2005.
- [9] Jonsdottir, S.O., Jorgensen, F.S., and Brunak, S., Prediction methods and databases within chemoinformatics: Emphasis on drugs and drug candidates, *Bioinformatics*, 21:2145–2160, 2005.
- [10] Lin, T., Melgar, M., Swamidass, S. J., Purdon, J., Tseng, T., Gago, G., Kurth, D., Baldi, P., Gramajo, H., and Tsai, S., Crystal structure of the carboxyltransferase domain of Acyl-CoA carboxylase, AccD5 from mycobacterium tuberculosis: New TB therapeutic target, *PNAS*, 2005, submitted.

- [11] Lipinski, C. and Hopkins, A., Navigating chemical space for biology and medicine, *Nature*, 432:855–861, 2004.
- [12] Marris, E., Chemistry society goes head to head with NIH in fight over public database, *Nature*, 435(7043):718–719, 2005.
- [13] Micheli, A., Sperduti, A., Starita, A., and Biancucci, A.M., A novel approach to QSPR/QSAR based on neural networks for structures, *Soft Computing Approaches in Chemistry*, Cartwright, H. and Sztandera, L.M. (ed.), *Springer Verlag* 265–296, 2003.
- [14] Ralaivola, L., Swamidass, S.J., Saigo, H., and Baldi, P., Graph kernels for chemical informatics. Neural Networks, 2005. Special issue on Neural Networks and Kernel Methods for Structured Domains, 2005, In press.
- [15] Schreiber, S.L., Target-oriented and diversity-oriented organic synthesis in drug discovery, *Science*, 287:1964–1969, 2000.
- [16] Schreiber, S.L., The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology, *Chemical and Engineering News*, 81:51–61, 2003.
- [17] Shoichet, B.K., Virtual screening of chemical libraries, *Nature*, 432:862–865, 2004.
- [18] Stockwell, B.R., Exploring biology with small organic molecules, *Nature*, 432:846–854, 2004.
- [19] Strauseberg, R.L. and Schreiber, S.L., From knowing to controlling: a path from genomics to drugs using small molecule probes, *Science*, 300(5617):294–295, 2003.
- [20] Swamidass, S.J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., and Baldi, P., Kernels for small molecules and the prediction of mutagenicity, toxicity, and anti-cancer activity, Proceedings of the 2005 ISMB Conference, *Bioinformatics*, 21(Supplement 1):i359–368, 2005.