

Cheminformatics – An Important Scientific Discipline

Johann GASTEIGER^{a*} and Kimito FUNATSU^b

^aComputer-Chemie-Centrum and Institute of Organic Chemistry, University Erlangen-Nuremberg[§]
91052 Erlangen, Germany

^bDepartment of Chemical System Engineering, School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

(Received: October 27, 2005; Accepted for publication: February 10, 2006; Published on Web: March 31, 2006)

Cheminformatics is the application of informatics methods to solve chemical problems. Although this term was introduced only a few years ago, this field has a long history with its roots going back more than 40 years. These different origins have now merged into a discipline of its own that is full of activities. All areas of chemistry from analytical chemistry to drug design can benefit from cheminformatics methods. And there are still many challenging chemical problems waiting for solutions through the further development of cheminformatics.

1 Introduction

The term “Cheminformatics” appeared a few years ago and rapidly gained widespread use. Workshops and symposia are organized that are exclusively devoted to cheminformatics, and many job advertisements can be found in journals. The first mention of cheminformatics may be attributed to Frank Brown. [1]

The use of information technology and management has become a critical part of the drug discovery process.

Cheminformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.

Whereas we see here cheminformatics focused on drug design, Greg Paris came up with a much broader definition [2]:

Cheminformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.

Clearly, the transformation of data into information, and of information into knowledge is an endeavor needed in any branch of chemistry not only in drug design. We therefore share the view that cheminformatics methods are needed in all areas of chemistry and adhere to a much broader definition:

Cheminformatics is the application of informatics

[§]<http://www2.chemie.uni-erlangen.de>

methods to solve chemical problems.

Why do we have to use at all informatics methods in chemistry?

First of all, chemistry has produced an enormous amount of data and this data avalanche is rapidly increasing. More than 45 million chemical compounds are known and this number is increasing by several millions each year. Novel techniques such as combinatorial chemistry and high-throughput screening generate huge amounts of data. All this data and information can only be managed and made accessible by storing them in databases.

On the other hand, for many problems the necessary information is not available. We know the 3D structure, determined by X ray crystallography for about 300,000 organic compounds. Or, as another point, the largest database of infrared spectra contains about 200,000 spectra. Although these numbers may seem large, they are small in comparison to the number of known compounds: We know from less than 1% of all compounds their 3D structure or have their infrared spectra. The question is then, can we gain enough knowledge from the known data to make predictions for those cases where the required information is not available?

There is another reason why we need informatics methods in chemistry: Many problems in chemistry are too complex to be solved by methods based on first principles through theoretical calculations. This is true, for the relationships between the structure of a compound

and its biological activity, or for the influence of reaction conditions on chemical reactivity.

All these problems in chemistry require novel approaches for managing large amounts of chemical structures and data, for knowledge extraction from data, and for modeling complex relationships. This is where chemoinformatics methods can come in.

2 History of Chemoinformatics

With all these problems at hand in chemistry, complex relationships, profusion of data, lack of necessary data, quite early on the need was felt in many areas of chemistry to have resort to informatics methods. These various roots of chemoinformatics often go back more than 40 years into the 1960s.

2.1 Chemical Structure Representatio

In the early sixties, various forms of machine readable chemical structure representations were explored as a basis for building databases of chemical structures and reactions. [3] Eventually, connection tables that represent molecules by lists of the atoms and of the bonds in a molecule gained universal acceptance. Connection tables were also used for the Chemical Abstracts Registry System which appeared in the second half of the sixties. [4]

2.2 Structure Searching

A connection table is essentially a representation of the molecular graph. Therefore, for storing a unique representation of a molecule and for allowing its retrieval, the graph isomorphism problem had to be solved to define from a set of potential representations of a molecule a single one as the unique one. The first solution was the Morgan algorithm for numbering the atoms of a molecule in a unique and unambiguous manner. [5] This provided the basis for full structure searching. Then, methods were developed for substructure searching, for similarity searching, and for 3D structure searching.

2.3 Quantitative Structure Activity / Property Relationship (QSAR/QSPR)

Building on work by Hammett and Taft in the fifties, Hansch and Fujita showed in 1964 that the influence of substituents on biological activity data can be quantified. [6]

In the last 40 years, an enormous amount of work on relating descriptors derived from molecular structures with a variety of physical, chemical, or biological data has appeared. These studies have established Quantitative Structure–Activity Relationships (QSAR) and Quantitative Structure–Property Relationships (QSPR) as fields

of their own, with their own journals, societies, [7] and conferences.

2.4 Chemometrics

Initially, the quantitative analysis of chemical data relied exclusively on multilinear regression analysis. However, it was soon recognized in the late sixties that the diversity and complexity of chemical data need a wide range of different and more powerful data analysis methods. Pattern recognition methods were introduced in the seventies to analyze chemical data. In the nineties, artificial neural networks gained prominence for analyzing chemical data.[8] The growing of this area led to the establishment of chemometrics as a discipline of its own with its own society, [9] journals, and scientific meetings.

2.5 Molecular Modeling

In the late sixties, R. Langridge and coworkers developed methods for visualizing 3D molecular models on the screens of Cathode Ray Tubes. At the same time, G. Marshall started visualizing protein structure on graphic screens. The progress in hardware and software technology, particularly as concerns graphics screens and graphics cards, has led to highly sophisticated systems for the visualization of complex molecular structures in great detail. Programs for 3D structure generation, for protein modeling, and for molecular dynamics calculations have made molecular modeling a widely used technique.

2.6 Computer-Assisted Structure Elucidation (CASE)

The elucidation of the structure of a chemical compound, be it a reaction product or a compound isolated as a natural product, is one of the fundamental tasks of a chemist. Structure elucidation has to consider a wide variety of different types of information mostly from various spectroscopic methods, and has to consider many structure alternatives. Thus, it is an ambitious and demanding task. It is therefore not surprising that chemists and computer scientists had taken up the challenge and had started in the 1960 's to develop systems for computer-assisted structure elucidation (CASE) as a field of exercise for artificial intelligence techniques. The DENDRAL project, initiated in 1964 at Stanford University gained widespread interest. [10]

Other approaches to computer-assisted structure elucidation were initiated in the late sixties by Sasaki at Toyohashi University of Technology [11] and by Munk at the University of Arizona. [12]

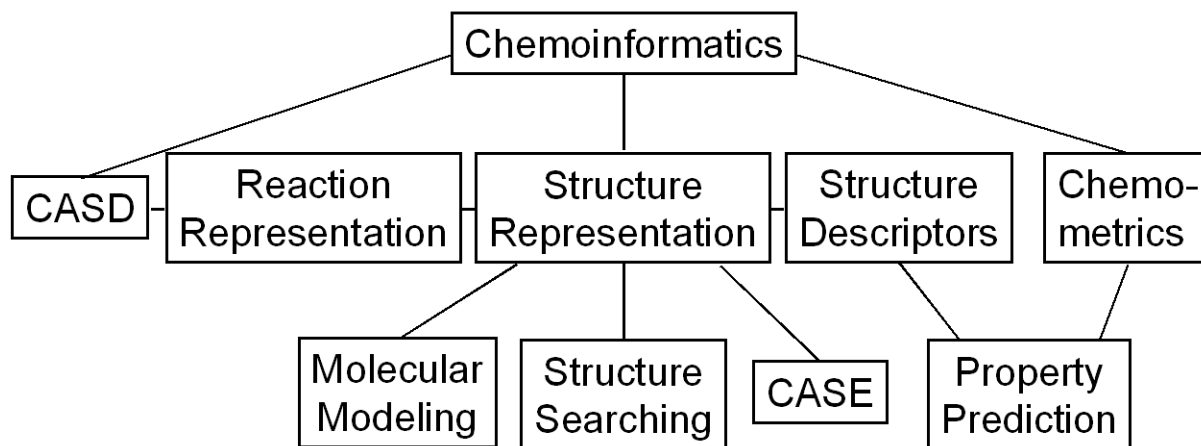


Figure 1. The various areas of activities in chemoinformatics

2.7 Computer-Assisted Synthesis Design (CASD)

The design of a synthesis for an organic compound needs a lot of knowledge about chemical reactions and on chemical reactivity. Many decisions have to be made between various alternatives as to how to assemble the building blocks of a molecule and which reactions to choose. Therefore, computer-assisted synthesis design (CASD) was seen as a highly interesting challenge and as a field for applying artificial intelligence techniques. In 1969 Corey and Wipke presented their seminal work on the first steps in the development of a synthesis design system. [13] Nearly simultaneously several other groups such as Ugi and coworkers, [14] Hendrickson [15], and Gelernter [16] reported on their work on CASD systems. Later also at Toyohashi work on a CASD system was initiated. [17]

3 Chemoinformatics – A Mature Discipline of Its Own

3.1 A New Discipline

The various fields outlined in the previous section have grown from humble beginnings 40 years ago to areas of intensive activities. On top of that it has been realized that these areas share a large number of common problems, rely on highly related data, and work with similar methods. Thus, these different areas have merged to a discipline of its own: Chemoinformatics. (Figure 1)

The extent of this field has recently been documented by a “Handbook of Chemoinformatics”, covering 73 contributions by 65 scientists on 1850 pages in four volumes. [18]

It is also felt that chemoinformatics has to be taught in courses of their own to prepare experts in this field for academia and industry. Furthermore, the major topics of chemoinformatics have to be integrated into chemistry curricula to produce a new generation of chemists aware and knowledgeable in the essentials of chemoinformatics to make chemical research and development more efficient. In order to assist in this endeavor, a “Textbook on Chemoinformatics” has been produced. [19]

3.2 Overview of Topics

The following gives an overview of chemoinformatics, emphasizing the problems and solutions – common to the various more specialized subfields. These topics also constitute the various chapters of the Textbook and of the Handbook of Chemoinformatics. [18, 19]

1. Representation of Chemical Compounds

A whole range of methods for the computer representation of chemical compounds and structures has been developed: linear codes, connection tables, matrices. Special methods had to be devised to uniquely represent a chemical structure, to perceive features such as rings and aromaticity, and to treat stereochemistry, 3D structures, or molecular surfaces.

2. Representation of Chemical Reactions

Chemical reactions are represented by the starting materials and products as well as by the reaction conditions. On top of that, one also has to indicate the reaction site, the bonds broken and made in a chemical reaction. Furthermore, the stereochemistry of reactions has to be handled.

3. Data in Chemistry

Much of our chemical knowledge has been derived from data. Chemistry offer a rich range of data on physical, chemical, and biological properties: binary data for

classification, real data for modeling, and spectral data having a high information density. These data have to be brought into a form amenable to easy exchange of information and to data analysis.

4. Datasources and Databases

The enormous amount of data in chemistry has led quite early on to the development of databases to store and disseminate these data in electronic form. Databases have been developed for chemical literature, for chemical compounds, for 3D structures, for reactions, for spectra, etc. The internet is increasingly used to distribute data and information in chemistry.

5. Structure Search Methods

In order to retrieve data and information from databases, access has to be provided to chemical structure information. Methods have been developed for full structure, for substructure, and for similarity searching.

6. Methods for Calculating Physical and Chemical Data

A variety of physical and chemical data of compounds can directly be calculated by a range of methods. Foremost are quantum mechanical calculations of various degrees of sophistication. However, simple methods such as additivity schemes can also be used to estimate a variety of data with reasonable accuracy.

7. Calculation of Structure Descriptors

In most cases, however, physical, chemical, or biological properties cannot be directly calculated from the structure of a compound. In this situation, an indirect approach has to be taken by, first, representing the structure of the compound by structure descriptors, and, then, to establish a relationship between the structure descriptors and the property by analyzing a series of pairs of structure descriptors and associated properties by inductive learning methods. A variety of structure descriptors has been developed encoding 1D, 2D, or 3D structure information or molecular surface properties.

8. Data Analysis Methods

A variety of methods for learning from data, of inductive learning methods is being used in chemistry: statistics, pattern recognition methods, artificial neural networks, genetic algorithms. These methods can be classified into unsupervised and supervised learning methods and are used for classification or quantitative modeling.

4 Applications of Chemoinformatics

4.1 Fields of Chemistry

The range of applications of chemoinformatics is rich indeed; any field of chemistry can profit from its methods. The following lists different areas of chemistry and indicates some typical applications of chemoinformatics. It has to be emphasized that this list of applications is by far not complete!

1. Chemical Information

- storage and retrieval of chemical structures and associated data to manage the flood of data
- dissemination of data on the internet
- cross-linking of data to information

2. All fields of chemistry

- prediction of the physical, chemical, or biological properties of compounds

3. Analytical Chemistry

- analysis of data from analytical chemistry to make predictions on the quality, origin, and age of the investigated objects
- elucidation of the structure of a compound based on spectroscopic data

4. Organic Chemistry

- prediction of the course and products of organic reactions
- design of organic syntheses

5. Drug Design

- identification of new lead structures
- optimization of lead structures
- establishment of quantitative structure-activity relationships
- comparison of chemical libraries
- definition and analysis of structural diversity
- planning of chemical libraries
- analysis of high-throughput data
- docking of a ligand into a receptor
- *de novo* design of ligands
- modeling of ADME-Tox properties
- prediction of the metabolism of xenobiotics
- analysis of biochemical pathways

Varied as these areas are and diversified as these applications are, the field of chemoinformatics is by far not fully developed. There are many areas and problems that can still benefit from the application of chemoinformatics methods. There is much space for innovation in seeking for new applications and for developing new methods.

4.2 Teaching Chemoinformatics

Chemists have to become more efficient in planning their experiments, have to extract more knowledge from their data. Chemoinformatics can help in this endeavor. Furthermore, it is important that a certain amount of chemoinformatics is integrated into chemistry curricula in order that chemists realize where chemoinformatics could help them, where they best ask chemoinformatics experts. In addition, a few universities have to offer training for chemoinformatics specialists. The first steps have already been made at a variety of universities around the globe. More has to come in order that more experts on chemoinformatics are trained that society so urgently needs.

5 Conclusions

Chemoinformatics has developed over the last 40 years to a mature discipline that has applications in any area of chemistry. The field has gained so much in importance that the major topics of chemoinformatics have to be integrated into chemistry curricula, a few universities have to offer full chemoinformatics curricula to satisfy the urgent need for chemoinformation specialists. There are still many problems that await a solution and therefore we still will see many new developments in chemoinformatics.

References

- [1] F. K. Brown, *Ann. Reports Med. Chem.*, **33**, 375-384 (1998).
- [2] G. Paris (August 1999 Meeting of the American Chemical Society), quoted by W. Warr at <http://www.warr.com/warrzone.htm>
- [3] F. A. Tate, *Ann. Rev. Inf. Sci. Technol.*, **2**, 285-309 (1967).
- [4] G. M. Dyson, M. F. Lynch, H. L. Morgan, *Inf. Storage Retrieval*, **4**, 27-83 (1968).
- [5] H. L. Morgan, *J. Chem. Docum.*, **5**, 107-113 (1965).
- [6] C. Hansch, T. Fujita, *J. Am. Chem. Soc.*, **86**, 856-864 (1964).
C. Hansch, T. Fujita, *J. Am. Chem. Soc.*, **86**, 1616-1626 (1964).
- [7] QSAR and Modelling Society: <http://www.qsar.org>
Molecular Graphics and Modeling Society: <http://mgms.org>
- [8] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, 2nd Edition, Wiley-VCH, Weinheim (1999).
- [9] International Chemometrics Society:
<http://www.mamics.nysaes.cornell.edu/chem-society.html>
- [10] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Applications of Artificial Intelligence for Organic Chemistry; the Dendral Project*, McGraw-Hill, New York (1980).
- [11] S. I. Sasaki, H. Abe, T. Ouki, M. Sakamoto, S. Ochiai, *Anal. Chem.*, **40**, 2220-2223 (1968).
- [12] C. A. Shelley, T. R. Hays, M. E. Munk, H. V. Roman, *Anal. Chim. Acta*, **103**, 121-132 (1978).
- [13] E. J. Corey, W. T. Wipke, *Science*, **166**, 178-193 (1969).
- [14] J. Blair, J. Gasteiger, C. Gillespie, P. D. Gillespie, I. Ugi, *Tetrahedron*, **30**, 1845-1859 (1974).
- [15] J. B. Hendrickson, *J. Am. Chem. Soc.*, **93**, 6847-6854 (1971).
- [16] H. L. Gelernter, N. S. Sridharan, A. J. Hart, S.-C. Yen, *Top. Curr. Chem.*, **41**, 113-150 (1973).
- [17] K. Funatsu, S. Sasaki, *Tetrahedron: Comput. Methodol.*, **1988**, 127-137.
K Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **39**, 316-325 (1999).
- [18] J. Gasteiger, Editor, *Handbook of Chemoinformatics – From Data to Knowledge*, Wiley-VCH, Weinheim (2003).
- [19] Chemoinformatics – A Textbook, J. Gasteiger, T. Engel, Editors, Wiley-VCH, Weinheim (2003).

