Foreword

Mathematically speaking, two chicken halves give one chicken. This is true with respect to mass. Such a chicken, however, is not a living one. Moreover, the two halves must be mirror images to form the shape of a chicken. Obviously, numbers are only half of the story, information comes in disparate guises. Chemistry, the science of materials and their transformations, exhibits a broad diversity of information, which by now encompasses an enormous body of knowledge about chemical structures, properties, and reactions. However, despite all the achievements during the last two centuries, which changed early chemical craftsmanship into a sophisticated natural science, chemistry is still devoid of the all-round theory, having, for examples the potential to predict precise structure-activity or structure-function relationships. It would explain, for instance, why palytoxin $C_{129}H_{223}N_3O_{54}$, being isolated from a Hawaiian coral in 1979, is one of the most poisonous natural substances. With its 64 chiral carbon atoms and six olefinic bonds, offering more than 10^{21} possible isomers, the total synthesis of palytoxin carboxylic acid may be compared with the first climbing of Mount Everest (Kishi, et al. at Harvard University, 1989).

Comparable efforts are needed to master the flood of information and accumulated knowledge in chemistry today. While until 1960 the number of natural and laboratory-produced compounds had almost linearly increased to roughly one million in about 150 years, its growth expanded exponentially from then on, reaching 18 million in 2000. This is just one aspect of the revolution in chemistry brought about by the rapid advancement of computer technology since 1965. Methods of physics, mathematics and information science entered chemistry to an unprecedented extent, which furnished laboratories with powerful new instrumental techniques. Also, a broad variety of model-based or quantum-mechanical computations became feasible, which were thought impossible a few decades ago. For example, the computer modeling of water transport through membranes mediated by aquaporins yields the time dependence of the spatial position of typically 10^5 atoms on a picosecond scale up to 10 ns (Grubmüller et al., MPI Göttingen). Such huge data arrays can be searched for and accessed via computer networks and then evaluated in a different context („data mining“). Furthermore, new chemical techniques, such as combinatorial synthesis, have high data output. Overall it can be stated that, particularly for the chemical and pharmaceutical industries, researchers now spend more time in digesting data than in generating them, whereas the reverse was true a few years ago.

In the 1970's, chemists increasingly encountered varying aspects of the triumvirate "chemistry-information-computer" (CIC) while conducting their research. Common to all was the use of computers and information technologies for the generation of data, the mixing of data sources, the transformation of data into information and then information into knowledge for the ultimate purpose of solving chemical problems, e.g. organic synthesis planning, drug design, and structure elucidation. These activities led to a new field of chemical expertise which had distinctly different features compared with the traditional archiving approach of chemical information, which has been established about 200 years ago and comprises primary journals, secondary literature, and retrieval systems like Chemical Abstracts.

In the 1980s, computer networks evolved and opened a new era for fast data flow over almost any distance. Their importance was not generally recognized in the chemical community at the beginning. The situation may be characterized with words from the late Karl Valentin: "A computer network is something that one does not want to be in the need to have, nevertheless simply must want to have, because one always might be in need to use it. (Ein Datennetz ist etwas was man eigentlich nie brauchen müssen möchte, aber doch einfach wollen muß, weil man es immer brauchen tun könnte.)" In 1986 Johann Gasteiger, also in Munich, coinitiated the Task Force CIC of the German Chemical Society (GDCh). In the same year, he started the CIC Workshops on Software Development in Chemistry which found overwhelming acceptance. Until today these annual meetings have served as a forum for the presentation and dissemination of recent results in the various CIC fields, including chemical information systems. The Task Force CIC later merged with GDCh Division Chemical Information under the name "Chemie-Information-Computer (CIC)".

The 1990s saw the advent of the Internet, which boosted the use of computer networks in chemistry. In its sequel, at the turn of the century, the work of the forerunners at the intersection of chemistry and computer science eventually received recognition in its entirety as a new interdisciplinary science: "Chemoinformatics" had come of age. It encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information (G. Paris). Its cousin, Bioinformatics, which was developed somewhat earlier, generally focuses on genes and proteins, while chemoinformatics centers on small molecules. Yet the distinction is fuzzy, e.g. when the binding of small molecules to proteins is addressed. From the viewpoint of the life sciences, the borderline may blur completely.

A young scientific discipline grows with its students. For students, in turn, the efforts pay dividends. The demand from industrial employers increases steadily for chemoinformaticians, and so the field is expected to become big business. The question therefore arises how and where chemoinformatics can be learned. One good place to go is Erlangen. Johann Gasteiger and his group were practicing chemoinformatics for 25 years without even knowing its name. In 1991 he received the Gmelin Beilstein medal of the German Chemical Society (GDCh) and in 1997, he received the “Herman Skolnik Award“ of the Division of Chemical Information of the American Chemical Society in recognition for his many achievements in the CIC field. With the present comprehensive textbook on chemoinformatics for undergraduate and graduate students, Gasteiger and his group lay a solid foundation-stone for many more “Erlangens“ in the world. My warm recommendation goes with this book, in particular to my academic colleagues. It seems to be the right time for universities to begin to teach chemoinformatics on a broad scale. Subject to local reality this may be conducted as part of a diploma course of study in chemistry, as a master’s study after a BS in chemistry, or as a full course of study in chemoinformatics, like that available in bioinformatics. In all cases, this book and its supplementary material would provide the adequate basis for teaching as well as for self-paced learning.

Dieter Ziessow
Chairperson
“Chemie-Information-Computer (CIC)”
of the German Chemical Society (GDCh)

Preface

Computers have penetrated nearly every aspect of daily life. Clearly, scientists and engineers took the lead in this process by applying computers for solving problems. In chemistry, it was realized quite early on that the huge amount of information available can only be handled by electronic means, by storing this information in databases. Only in this way can the huge number (35 million) of chemical compounds known at present be handled. Thus, already in the 1960s, work on chemical databases was initiated. Furthermore, many relationships between the structures of compounds and their physical, chemical, or biological properties are highly complex, asking either for highly sophisticated computations or for analysis of a host of related data to make predictions on such properties. Chemical societies in many countries have recognized the importance of computers in their field and have founded divisions that focus on the use of computers in chemistry.

From the very beginning, however, it could be observed that there was a split between theoretical chemists using computers for quantum mechanical calculations and chemists using computers for information processing and data analysis. The American Chemical Society has two divisions, the Division of Computers in Chemistry and the Division of Chemical Information. In Germany there is the Theoretische Chemie group associated with the Deutsche Bunsen-Gesellschaft für Physikalische Chemie and the Division Chemie-Information of the Gesellschaft Deutscher Chemiker. In fact, in 1989 this Division changed its name to Chemie-Information-Computer (CIC) to recognize the growing importance of using computers for processing chemical information. A small group of scientists within this division was quite active in spreading the message of using computers in chemistry. Two workshops were initiated in 1987, one on *Software Development in Chemistry* and one on *Molecular Modeling*, ever since these workshops have been held on a yearly basis.

On another level, the German Federal Minister of Research and Technology (BMFT; later renamed BMBF) initiated programs in the 1980s to found so-called Fachinformationszentren (FIZ) and, in addition, to build databases. Chemists can consider themselves fortunate that the experts and politicians recognized the importance of databases in chemistry. Thus, some of the internationally most highly recognized databases were initiated: the Beilstein Database for organic compounds, the Gmelin Database for inorganic and organometallic compounds, the ChemInform RX reaction database, and the SpecInfo database for spectroscopic information.

In spite of all these activities it must nevertheless be observed that chemists have only gradually accepted the computer as a much needed tool in their daily work. But they gradually – or grudgingly? – did accept it: databases are used routinely for retrieving information, and quantum chemical or molecular mechanics programs are used – mostly a *posteriori* – to further an understanding of chemical observations. Furthermore, since the advent of combinatorial chemistry and high-throughput screening it has become increasingly clear that the flood of information produced by these techniques can only be handled by computer methods.

Thus, computers will continue to penetrate every aspect of chemistry and we have to prepare the next generation of chemists for this process. In fact, we will see that the various types of computer applications in chemistry will increasingly be used in concert to solve chemical problems. Therefore, a unified view of the entire field is needed; the various approaches to using computers in chemistry have to be ordered into a common framework, into a discipline of its own: Chemoinformatics.

With this textbook we present the first comprehensive overview of chemoinformatics, current material that can be integrated into chemistry curricula or can serve on its own as a basis for an entire course on chemoinformatics.

This textbook can build on 25 years of research and development in my group. First of all, I have to thank all my co-workers, past and present, that have ventured with me into this exciting new field. In fact, this textbook was written nearly completely by members of my research group. This allowed us to go through many text versions in order to adjust the individual chapters to give a balanced and homogeneous presentation of the entire field. Nevertheless, the individual style of presentation of each author was not completely lost in this process and we hope that this might make reading and working through this book a lively experience. Writing these contributions on top of their daily work was sometimes an arduous task. I have to thank them for embarking with me on this journey.

We also want to thank the Federal Minister of Education and Research (BMBF) for funding a project "Networked Education in Chemistry", administered by FIZ CHEMIE, Berlin. Within this project we are developing eLearning tools for chemoinformatics.

In addition, we thank Dr. Gudrun Walter of Wiley-VCH, for encouraging us to embark on this project and Dr. Romy Kirsten for the smooth collaboration in processing our manuscripts.

We just hope that this Textbook will generate interest in chemoinformatics for a wider audience, and will make them excited about this field much in the same way as we are excited about it.

Erlangen, May 2003 *Johann Gasteiger*