Chemoinformatics: Past, Present, and Future[†]

William Lingran Chen*

Elsevier MDL, 2440 Camino Ramon, Suite 300, San Ramon, California 94583

Received January 16, 2006

The history of chemoinformatics is reviewed in a decade-by-decade manner from the 1940s to the present. The focus is placed on four traditional research areas: chemical database systems, computer-assisted structure elucidation systems, computer-assisted synthesis design systems, and 3D structure builders. Considering the fact that computer technology has been one of the major driving forces of the development of chemoinformatics, each section will start from a brief description of the new advances in computer technology of each decade. The summary and future prospects are given in the last section.

INTRODUCTION

In the middle of the 1990s, there were some arguments about the best name for the newly emerging discipline of computer applications in chemistry. At that time, there were already societies of chemometrics.[1] Two journals had the word chemometrics in their names: *Chemometrics and Intelligent Laboratory Systems*[2] and *Journal of Chemometrics*.[3] On the other hand, although there were neither societies of computer chemistry nor journals with the names containing the term “computer chemistry” (The *Journal of Computer Chemistry* was launched in 2002 in Japan[4]), this term was quite widely used by chemists whose main research was focused on the development of chemical database retrieval systems and chemical expert systems. There were several institutes with the term “computer chemistry” in their names, such as Computer-Chemistry-Center,[5] Erlangen, Germany. Two books were titled *Computer Chemistry*.[6,7] Therefore, some chemists liked chemometrics, while others preferred computer chemistry for the name of the new discipline.

Surprisingly, neither chemometrics nor computer chemistry won the game; instead, both lost to a newcomer—chemoinformatics.

In 1976, the word bioinformatics (or in Dutch “Bioinformatica”) first appeared in the name of a research group called “Bioinformatics Group” at University of Utrecht, The Netherlands. The first paper using the term bioinformatics was published 2 years later by Hogeweg from that group.[8] They defined bioinformatics as “the study of informatic processes in biotic systems”. With the explosion of publicly available genomic information, such as that resulting from the Human Genome Project,[9] in the middle of the 1990s bioinformatics has become very popular not only in the scientific community but also in the general audience as well. This has led to the coining of the counterpart of bioinformatics in chemistry: chemoinformatics. It is generally accepted that the term chemoinformatics was first introduced by Dr. Frank Brown in the *Annual Reports of Medicinal Chemistry* in 1998.[10] With the publication of Professor Johann Gasteiger’s *Handbook of Chemoinformatics* in 2003,[11] chemoinformatics has now become the *de facto* standard for applications of computer and informatics technology in chemistry. The only thing we should be aware of is that a shorter form of it, cheminformatics, has also been widely used.

WHAT IS CHEMOINFORMATICS?

In 1998, Dr. Brown gave his original definition of chemoinformatics as follows: “The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics 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.”[10] The scope of chemoinformatics has now greatly exceeded the drug discovery area, as evidenced by the four volumes of the *Handbook of Chemoinformatics*, in which the editor, Professor Gasteiger, introduced a much broader definition: “Chemoinformatics is the use of informatics methods to solve chemical problems.”[12]

Although the term chemoinformatics is only 8 years old, the roots of the field date back much, much earlier. For example, the first and most important journal covering this field, the *Journal of Chemical Documentation* (JCD), was established in 1961.[13] To clarify its role and to attract papers from the ACS Division of Computers in Chemistry, JCD was renamed in 1975 as the *Journal of Chemical Information and Computer Sciences* (JCICS). To better reflect the contents of the journal, JCICS was again renamed as the *Journal of Chemical Information and Modeling* (JCIM) in 2005.[14] Although JCICS was not renamed as the *Journal of Chemoinformatics*, as some chemists had expected,[15] its new title was indeed a well chosen one, and it also indicates that the current definition of chemoinformatics may still be too narrow: chemoinformatics should also encompass chemical modeling. In fact, several chapters of Gasteiger’s *Handbook of Chemoinformatics* were also dedicated to molecular modeling.

In this contribution we will present an overview of the major historical developments of chemoinformatics in a decade-by-decade manner from the 1940s to the present. However, it must be pointed out that since this article is

[†] Dedicated to Professor Johann Gasteiger.

* Corresponding author phone: (925)543-7541; e-mail: L.Chen@mdl.com. This author was previously known as Lingran Chen.

10.1021/ci060016u CCC: $33.50
Published on Web 06/03/2006

written for a special JCIM issue dedicated to Professor Gasteiger, our intention is to emphasize his contributions, and therefore this will not be a comprehensive review of all areas of chemoinformatics or all researchers in these areas, though efforts have been made to balance the narration. The article will focus mainly on four "traditional" areas of chemoinformatics: chemical database systems, computer-assisted structure elucidation (CASE) expert systems, computer-assisted synthesis design (CASD) expert systems, and 3D structure builders. We have chosen these "traditional" areas mainly for two reasons. First, they have been the major research areas in the history of chemoinformatics. Second, many of the fundamental algorithms and methodologies (such as chemical structure representation, substructure searching, ring perception, stereochemistry representation and manipulation, and so on) used in many newer chemoinformatics systems (such as combinatorial chemistry, high-throughput screening (HTS), docking) were originally invented during the development of these "traditional" chemoinformatics systems.

Each section starts with an early application of electronic computers to each related area. Therefore even earlier, precomputer-era work of historic importance will not be covered. Furthermore, other very important topics, such as quantitative structure−activity relationship (QSAR), docking, combinatorial chemistry, HTS, and so on, will not be covered here. It should also be mentioned that, unlike traditional review articles, this paper will focus mainly on the development of chemoinformatics systems, and thus, many algorithms and methodologies of historical importance will not be covered. The readers who are interested in those topics may refer to Gasteiger's comprehensive *Handbook of Chemoinformatics*.[11] In addition, considering the fact that computer technology has been one of the major driving forces of the development of chemoinformatics, each section will start with a brief description of the new advances in computer technology of each decade.

1940S: COMPUTER TECHNOLOGY

The 1940s witnessed one of the greatest inventions in the human history: the birth of the electronic computer, which brought the revolution in the last century. In 1943, an electronic computer called Colossus and a more general purpose electromechanical programmable computer named Harvard Mark I were built in Britain and U.S.A., respectively.[16] In 1946, the first general-purpose computer, ENIAC (Electronic Numerical Integrator and Computer), was invented, which weighed 30 tons, contained 18 000 electronic valves, consumed around 25 KW of electrical power, and carried out 100 000 calculations per second.[16] These "first generation" computers were built using punched cards and vacuum tubes. All programs were implemented in machine code.

1940S: BIRTH OF CHEMOINFORMATICS

In 1946 King et al.[17] published an article illustrating the use of IBM's business accounting machines in carrying out the construction of the rotational spectra of asymmetric rotors by the evaluation of mathematical equations for line position and line intensity. This may be the earliest work of application of computer technology in chemistry, and thus, the year 1946 may be regarded as the birth year of chemoinformatics.

1950S: COMPUTER TECHNOLOGY

In the 1950s, the "second generation" computers, based on transistors and printed circuits, were built. In 1955, IBM made the first delivery of its landmark computer—IBM 650 to customers. In 1957, IBM made the first high level computer language—FORTRAN (FORmula TRANslation) available to customers. FORTRAN became the most widely used computer language for technical work.[18] In this decade, the electronic computers first became available for general use by scientists.[19]

1950S: DATABASE SYSTEMS

The earliest major pioneering work in the chemoinformatics field was mainly focused on the conversion of the printed collections of chemical data such as mass spectra and chemical literature into electronic formats and the developments of the corresponding database search systems. The earliest method in mass spectral search systems using punched cards for encoding spectral data was described by Zemany in 1950.[20] The punched card systems for storing IR spectra and retrieving spectral matches using sorters or collators were reported by Kuentzel in 1951.[21]

The Chemical Abstract Service (CAS) of the American Chemical Society, a major provider of chemical information, took advantage of computer support in its operations from the very beginning of the computer age. In 1955, CAS formed a Research and Development Department and laid the groundwork for producing computer-based chemical information databases.[22]

One of the most important algorithms in chemoinformatics is the substructure matching algorithm. In 1957, Ray and Kirsch described the first substructure searching algorithm based on a backtracking algorithm and a connection table for representation of chemical structures.[23] Their algorithm is still widely used today and often called the atom-by-atom matching.

In 1959, Opler and Baird described probably the first graphical display of chemical structures as computer output on the face of a cathode-ray tube.[24]

1960S: COMPUTER TECHNOLOGY

In the 1960s, computer technology was improved rapidly. The "third generation" computers based on the first integrated circuits were built. In 1964, IBM launched the first series of compatible mainframe computers, IBM 360.[18] In the same year, the first minicomputer, DEC PDP-8 Mini Computer, was built by Digital Equipment Corporation (DEC),[25] which was acquired by Compaq in 1998, and now is a part of Hewlett-Packard.[26] These computers, which supported the FORTRAN language, were much more powerful than the "second generation" computers, and became available at major research laboratories, leading to an explosion in the use of computers.

It is interesting to note that in 1965, Gordon E. Moore, a cofounder of Intel described the following empirical observation in *Electronics*, an American trade journal, on April 19, 1965: "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year ... Certainly over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years. That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000. I believe that such a large circuit can be built on a single wafer."[27] Later, this observation was slightly modified and widely called the Moore's law. The most popular formulation of the Moore's law is of the doubling of the number of transistors on integrated circuits (a rough measure of computer processing power) every 18 months. Although the Moore's law was initially made in the form of an observation and prediction, it has since become an important industry driver.

The 1960s saw the dramatic expansion of the research areas of chemoinformatics, from database retrieval systems, to more intelligent expert systems, to quantitative structure–activity relationships (QSAR), and to molecular modeling.

1960S: DATABASE SYSTEMS

In the 1960s, significant progress began to be made in the development of chemical database retrieval systems. In 1964, several reports were published on the use of computer techniques to search the ASTM database which contained tens of thousands of IR spectra.[28] In 1965, the British Government's Atomic Weapons Research Establishment (AWRE) launched a project to create a worldwide database of mass spectra, leading to the establishment of the Mass Spectrometry Data Centre (MSDC) at Aldermaston. A few years later, the U.S. National Institutes of Health (NIH) Laboratory of Chemistry started to develop a computer-based library retrieval system based on MSDC's mass spectral databases and one from Professor Biemann at MIT.[29]

Another noticeable event was that a new database collecting crystal structures, Cambridge Structural Database (CSD), was begun to be built in Cambridge, U.K., in 1965.[30] CSD has since become the earliest important resource for experimental 3D structure data. For example, most 3D structural builders, such as CONCORD and CORINA, have derived their templates from this database. It has played a more and more important role in solving difficult problems in structural chemistry and drug design. In 1966, Dubois et al. described the DARC (Documentation et d'automatisation des recherches de corrélations) system for documentation.[31] In 1969, Dubois described the principles of the DARC topological system.[32]

The most important new developments in the database area during this time came from CAS. In 1961, CAS produced *Chemical Titles*, the world's first periodical to be organized, indexed, and composed almost totally by computer, employing the Keyword-In-Context indexing technique. In 1965, CAS installed its Chemical Registry System. In 1968, CAS created *CA Condensates*, the first computer-readable file to cover the full range of abstracted documents.[22]

In 1965, Gluck[33] reported a chemical structure storage and search system developed at Du Pont, which was briefly described in *Chemical & Engineering News* in December 9, 1963. In this system Gluck developed the first algorithm for generating a canonical form of connection table (through atom-by-atom and bond-by-bond description) for each compound, which could be used to identify compounds. Much had been made of the uniqueness quality of manually derived chemical notations, which was prone to human error, and Gluck's step was a very important issue for connection tables. However, Gluck's algorithm was demonstrated to be erroneous. Morgan modified the derivation of Gluck's algorithm to increase its breadth of application and published his results in the same year.[34] This improved algorithm was used by CAS to generate a unique machine description of chemical structures for the CAS Chemical Compound Registry and has since been widely used and named the Morgan Algorithm, which would be better named the Gluck–Morgan Algorithm.

It should also be mentioned that some other important pioneering work was also carried out in this decade. In 1963, Vleduts suggested that the reaction site, which holds a key role in the development of reaction database systems, could be detected by comparison of the reactants with products.[35] Based on Vleduts' idea, in 1967 Armitage and Lynch described a technique for the automatic determination of structural similarities among pairs of chemical structures and how this technique could be applied to the computer-based manipulation of chemical information.[36]

In 1965, Sussenguth described the first set-reduction substructure searching algorithm,[37] which showed significant improvements in efficiency over the Ray-Kirsch atom-by-atom matching algorithm.[23]

In the early days, the computer hardware and software available could not handle the two-dimensional chemical structure diagrams. And thus, the earliest methods for chemical structure handling were mainly based on chemical nomenclature and linear notation.[38] A variety of linear notations has been suggested, but only a few gained significance. The Dyson notation, which was suggested by Dyson[39] in 1946 and adopted by the International Union for Pure and Applied Chemistry (IUPAC), was used in some of the earliest substructure searching experiments at CAS in the 1960s.[39] However, the Dyson notation never won many users because it was too radically different from familiar linear notation. In contrast, the Wiswesser Line notation (WLN),[40] which was first proposed by Wiswesser[41] in 1949, was widely used during the 1960s and early 1970s, becoming the *de facto* standard in the chemical industry during this period of time. The Index Chemicus Registry System (ICRS) from the Institute for Scientific Information and the Commercially Available Organic Chemicals Index (CAOCI) were two major WLN based databases.[38] In 1967, the ICI Pharmaceutical group described a system even with the word Wiswesser incorporated into the system's name: Computer Retrieval of Organic Structures Based on Wiswesser (CROSSBOW).[42] As a matter of fact, CROSSBOW, which had a large number of users worldwide, used not only WLN but also fragment codes and connection tables derived from WLN as well, as described by Warr.[38]

It is interesting to mention that in 1963, Ledley described probably the first direct input of chemical structural diagrams drawn freehand.[43]

1960S: CASE EXPERT SYSTEMS

One of the most exciting pioneering research frontiers of chemoinformatics in the 1960s was the exploration of artificial intelligence (AI) and its application in chemistry—chemical expert systems. An expert system is an AI application program for doing a task that requires expertise. It attempts to mimic human experts to solve narrow domain problems. The first expert system in chemistry was developed through the DENDRAL project begun in 1965. This project originated in Lederberg's studies at Stanford University on the scope of structural isomerism and of methods for representing chemical structures in terms of mathematical models.[44] The purpose of the DENDRAL project was to develop an expert system to automatically determine the structure of an unknown compound from the corresponding mass spectrum. The system consisted of three parts: spectral interpreter, structure generator, and candidate evaluator. First, the structural constraints were derived from the mass spectrum of the unknown compound. Then, all possible structural isomers were generated which met the requirements of the known molecular formula and the structural constraints. Finally, the mass spectral features of each structural candidate were predicted and compared with the experimental mass spectrum of the molecule, leading to the final set of most probable structural candidates for the new or unknown chemical compound.[45] The DENDRAL project is famous not only in chemoinformatics but also in computer science as well. In many textbooks of Artificial Intelligence, the DENDRAL system is cited as one of the earliest successful expert systems in the world. It should also be mentioned that the DENDRAL project is a good example of the collaboration among chemists, geneticists, and computer scientists, revealing the interdisciplinary nature of chemoinformatics.

DENDRAL-like expert systems are called computer-assisted structure elucidation systems. In the late 1960s, several other research groups also began developing similar expert systems. In 1967, Munk at Arizona State University started the development of his CASE program.[46] In 1968, Sasaki's group in Japan also began their research on the development of their own computer-assisted structure elucidation system.[47]

It should also be mentioned that other pioneering work was also carried out in this decade. For example, in 1964, Grant and Paul described an additivity model for the calculation of ^{13}C NMR chemical shift data for the alkanes, which has since become widely used.[48]

1960S: CASD EXPERT SYSTEMS

Another problem an organic chemist often encounters is how to find out the best route to synthesize a known compound. The traditional methods the chemists use to design the possible synthesis routes are fully knowledge/experience based. The second earliest chemical expert system of historical interest is OCSS (Organic Chemical Simulation of Syntheses) developed by Corey and Wipke at Harvard University and published in 1969.[49] OCSS was designed to be able to mimic the process of organic synthesis planning that a chemist uses—the retrosynthetic analysis approach. OCSS was based on the state-space search strategy, where the states were the original target structure and the new targets generated (intermediates) and the transform operators were reactions. OCSS worked strictly at a single level of granularity based on the individual reactions.

The introduction of these early expert systems in chemistry was welcomed with great enthusiasm and high expectations. Several other research groups also began developing this kind of expert system, called a computer-assisted synthesis design system, in the late 1960s.

It is worth mentioning that several other important new research areas were explored in the 1960s. In 1961, Hendrickson published a pivotal paper describing the first use of the electronic computer to calculate molecular geometry.[50] In 1964, Marsh and Hermann initiated a research project at Eli Lilly and Company to apply the semiempirical method to the study of drugs. At about the same time, some other pharmaceutical companies launched research projects of studying quantitative structure—activity relationships.[51]

In addition, in 1963, Merrifield published his seminal paper on solid-phase peptide synthesis, which is generally regarded as the origin of combinatorial chemistry.[52]

1970S: COMPUTER TECHNOLOGY

In the 1970s, the "fourth generation" computers based on Large Scale Integration (LSI) of circuits were built, which led to the wide availability of minicomputers, such as VAX (Virtual Address eXtension) 11/780 (in 1977).[53] This was also supported by the new operating systems, such as UNIX (in 1970) and VMS (in 1977), and a new programming language, C, which was developed by Dennis Ritchie of Bell Laboratories (in 1972).[54] It should also be mentioned that Vint Cerf and Robert Kahn developed the concept of connecting networks of computers into an "internet" and also developed the Transmission Control Protocol (TCP) in 1974.

Bill Gates and Paul Allen founded Microsoft under the name of "Micro-soft" in Albuquerque, NM in 1975. Gates first used the name Microsoft, without the hyphen in his letter to Allen on November 29, 1975, and in November 26, 1976 this name became a registered trademark.[55] Microsoft Corporation has now become the world's largest software company. Steve Wozniak and Steve Jobs started Apple Computers, Inc.[56] and released Apple I, the first single circuit board computer, on April Fool's Day, 1976.[57] In 1977, they released the Apple II, which was generally credited with creating the home computer market.[58] Lawrence J. Ellison (Larry Ellison) founded Oracle in 1977 under the name Software Development Laboratories, which was renamed Relational Software, Inc. in 1979. In 1983, the company was renamed Oracle Corporation[59] to more closely align itself with its flagship product Oracle database.[60]

1970S: DATABASE SYSTEMS

There were a number of interesting advances in the database area in the 1970s. Crowe et al.[61] developed fragment-based screening systems, significantly improving the performance of searching larger structure databases, as described by Graf et al.[62]

In 1971, Dr. Walter Hamilton established the Protein Data Bank (PDB) at Brookhaven National Laboratory. The PDB is a database which contains experimentally determined crystallographic data (3D structures) of biological macromolecules, such as proteins and nucleic acids.[63] In 1974, Gund et al. described a system for 3D structure searching,[64] laying the foundation for the further development of 3D structure searching systems.

One of the early collections of ^{13}C NMR spectra utilizing atom-centered fragments to organize substructure/shift correlations is the Johnson and Jankowski's work described in 1972.[65] They used the alphanumeric codes to represent the alpha and beta environment of a resonating atom. In 1975, Jezl and Dalrymple described a program based on a fixed length atom-centered code which represents the alpha and beta shells of a resonating carbon atom together with summary information about the gamma shell.[66] A serious disadvantage of their coding method is that it has to be performed manually. In 1975, Bremser et al. developed an atom-centered fragment code named HOSE-code (Hierarchically-Ordered Spherical Environments) for the resonating nucleus.[67] HOSE-codes explicitly define successive shell-levels up to four-bond layers and can be automatically generated for each atom in a molecule. This coding scheme was subsequently revised to avoid ambiguity.[68]

In 1972, Erni and Clerc reported the first system provided for searching combined spectral databases of ^{1}H NMR, IR, and mass spectra.[69] In 1973, Kwok et al. described the STIRS system for retrieving substructures from mass spectral searches.[70] In 1979, Dubois and Bonnet described the DARC Pluridata system for the ^{13}C NMR database.[71]

From about 1970 to 1984, the U.S. Government developed the Chemical Information System (CIS), which included a subsystem called the Mass Spectral Search System (MSSS). MSSS was at that time the only large time-sharing system. It is interesting to note that the original MSSS was sponsored by the U.K. Government, and later it was taken over by the U.S. Government.[29] With the increase of the sizes of mass spectral databases, the quality of the spectra contained in the files became an obvious concern of the scientific community. And thus, in 1974, the U.S.−U.K. group decided to remove redundant or multiple copies of spectra from the file, which also reduced the storage space usage and computer search time. The CAS Registry number was used as the unique identifier, and the CAS nomenclature was accepted for the primary names.[29] In 1978, Speck et al.[72] described a mass spectral quality index for selecting the best of any duplicate set of spectra of a given compound or identifying any highly dubious data. This quality index was used to check the data in the NIH-EPA-MSDC mass spectral collection and duplicate, lower quality spectra were discarded, as described by Heller et al.[73]

In 1973, Adamson and Bush explored the possibility of using the number of common substructural fragments of a pair of chemical structures as a measure of their similarity.[74] Later, their method was extended by Carhart et al. for QSAR studies in 1985[75] and by Willett et al. for database searching in 1986.[76]

In 1976, Ullmann described an efficient subgraph isomorphism algorithm.[77] The Ullmann algorithm has since been widely used in many fields, including chemoinformatics.

In 1978, Lynch and Willett reported a method for automatic detection of chemical reaction sites.[78] In the same year, Evans et al. reported the structural search codes for online compound registration.[79]

It is interesting to note that the first chemical software company was born in a Chinese restaurant in Berkeley in 1977. Stuart Marson and Steve Peacock met frequently there to discuss building a company to do consulting in computer-aided drug design. Together with Todd Wipke, they established Molecular Design Limited, Inc. in January 1978.[80] The first product was a chemical database system called MACCS (Molecular ACCess System) which ran on the Prime computer and an IMLAC graphics system. In 1979, Chevron Chemical Company's ORTHO division in Richmond, CA licensed MACCS, becoming the first company to license an MDL software product. MDL was part of the Maxwell Communications conglomerate from 1987 to 1992 and was then publicly traded on the NASDAQ stock market in 1993 as MDL Information Systems, Inc. In 1997 MDL was acquired by Reed Elsevier. MDL remains part of Elsevier, the scientific and medical division of Reed Elsevier, and in 2004 changed its name to Elsevier MDL.[81] Over the past three decades, MDL has been a major vendor of chemoinformatics software and chemical databases.[82]

1970S: CASE EXPERT SYSTEMS

As mentioned previously, the computer-assisted structure elucidation system consists of spectral interpreter, structure generator, and candidate evaluator. Of the three phases of the overall structure elucidation process, computers are most ideally suited to the structure generation phase. Also, the structure generator has been regarded as the heart of any automated structure elucidation system. Therefore, it is not surprising that in this decade much effort went into the development of more efficient algorithms of structure generators.

In 1971, Sasaki et al. published the CHEMICS system which analyzed the structures in terms of a predefined set of multiatom substructural fragments.[83] In 1974, Masinter et al. described the first complete solution to the problem of exhaustive and irredundant structure generation based on Lederberg's vertex graph formulation, and the algorithm was implemented in their program STRGEN.[84] STRGEN formed the computational core of the DENDRAL project's original CONGEN (CONSTRAINED GENERATOR) system for structure elucidation.[85] In 1976, Serov et al. reported the structure generator MASS (Mathematical Analysis and Synthesis of Structures)[86] for Gribov's STREC structure elucidation system.[87] In 1978, Shelley et al. reported the ASSEMBLE generator algorithm[88] for Munk's CASE (Computer Assisted Structure Elucidation) system.[89]

The traditional structure generators could only generate constitutional isomers and ignored all aspects of stereochemistry. This often led to problems during evaluation of candidate structures based on the prediction of their spectral properties, because many physical properties of molecules are influenced by stereochemical factors. In 1979, Nourse et al. described the first structure generation algorithm for generating configurational stereoisomers implemented in the program STEREO.[90]

A hindrance to building expert systems is the extraction of domain knowledge to be used by the computer program. To address this important problem, the DENDRAL team developed an inductive program called META-DENDRAL (1970−1976),[91] which automatically formulated new rules for DENDRAL to use in explaining data about unknown chemical compounds. Using the plan-generate-test paradigm, META-DENDRAL successfully formulated mass spectrometry rules, some of which were known rules while others were entirely new rules. Although META-DENDRAL is no longer in use, its ideas about learning and discovery of domain knowledge are important contributions to the concept of machine learning. Among these ideas are that induction can be automated as a heuristic search; that search can be broken into two steps—approximate and refined for efficiency; that learning must be able to cope with noisy and incomplete data; and so on.

1970S: CASD EXPERT SYSTEMS

After the publication of the first CASD expert system in 1969, several groups also started to pursue approaches to computer-assisted synthesis design. In 1971, Hendrickson published his first paper on the systematic characterization of structure and reactions for use in organic synthesis.[92] In 1975, he published his famous half-reaction theory, laying the foundation for developing his synthesis design system.[93]

In 1972, two children of the OCSS system were described. First, Corey's group extended their OCSS system by incorporating a more complex strategy system which included multistep plans based on useful reactions, such as the Diels−Alder reaction, the Robinson annulation, etc. The new system was named LHASA (Logic and Heuristics Applied to Synthetic Analysis).[94] It used the knowledge base of reaction transformations and a set of rules to perceive strategic bonds. LHASA has since become one of the most famous CASD systems. It should also be mentioned that the OCSS-LHASA system was the first system to allow drawing the chemical structure using a "Rand Tablet".[95] Second, Wipke, one of the developers of OCSS-LHASA, developed his own synthesis design system called SECS (Simulation and Evaluation of Chemical Synthesis) after he left Professor Corey's group. The most important new feature of SECS was that it included stereochemical information from its inception, which the first synthesis program OCSS-LHASA did not consider. SECS 1.0 was first demonstrated publicly via teletype at a Gordon Conference July 1972.[96] At one time, a large consortium of chemical companies utilized SECS and maintained an enormous database of transforms, which made SECS the most successful CASD system. The key to LHASA and SECS is an interactive mode which allows chemists to make choices from the synthesis tree as soon as precursors are generated.

In 1972 Bersohn described an unlabeled CASD program.[97] In 1973, Gelernter, a computer scientist, reported SYNCHEM, which was designed to address some of the issues of applying artificial intelligence (AI) to the complex synthesis design problem.[98] Unlike LHASA and SECS, both Bersohn's program and SYNCHEM are noninteractive, that is, the choices are made by the program and the chemist is presented with the "best routes" when it is finished.

In 1974, Blair et al. described a pilot program for synthesis design called CICLOPS. It was based on Ugi's mathematical model called BE-matrices and R-matrix.[99] In 1978, Gasteiger and Jochum reported the EROS (Elaborations of Reactions for Organic Synthesis) system.[100] EROS was probably the first expert system that could be applied to both synthesis design and reaction prediction. An important feature of EROS was that it could generate new reactions which had not been reported in the literature. EROS is one of the best known reaction simulation and prediction systems nowadays.

The modeling of reaction mechanisms asked for an evaluation of chemical reactivity. To this effect, in 1979, Gasteiger's group came up with a simple and rapid method for the calculation of partial charges in organic molecules by Partial Equalization of Orbital Electronegativities (PEOE).[101] Later, Gasteiger's group also developed empirical methods for the calculation of other fundamental physicochemical effects such as inductive,[102] resonance,[103] and polarizability effects.[104] Attention was devoted to come up with methods that require short computation times in order to be able to process large data sets of millions of structures such as those encountered in combinatorial libraries. The above methods have been implemented in the program package PETRA (Parameter Estimation for the Treatment of Reactivity Applications) and can be applied to the calculation of physicochemical effects in large data sets.

The PEOE method, now called "Gasteiger Charges", has since seen widespread use for reactivity prediction and modeling of physical, chemical, and biological properties, and it has been integrated into all major molecular modeling packages.

WHAT ARE GASTEIGER CHARGES?

A general description of the Gasteiger Charges was recently given by Professor Gasteiger himself[105] and is cited below.

"The calculation of partial charges in organic molecules falls into two parts: the calculation of sigma-bonded systems and that of pi-bonded systems. There exist quite a few implementations of our approach to charge calculation and I do not know how good a job other people - including commercial companies - have done in reproducing our work. The calculation of charges in sigma-bonded systems (PEOE) has been explicitly published such that anybody that wanted to implement it should have been able to do so correctly. The approach is still valid and all parameters are correctly given in the paper."

"The calculation of charges in conjugated pi-systems (PEPE), however, is somehow more complicated. Here, you have to first calculate sigma charges and then relax the pi electrons to the sigma charges and sigma and pi electronegativities. Our first approach was presented at a meeting in Croatia and the proceedings were published in Croatica Chimica Acta (I agree that this is not our regular journal for publishing our research!) We further modified our approach and then published it in Angewandte Chemie (J.Gasteiger, H.Saller, Angew. Chem. Intern. Ed. Engl. 24, 687−689 (1985)). This is only a communication giving the essential ideas which makes it not straightforward for people to implement our method. However, any charge calculation in conjugated systems that only considers the sigma method (and unfortunately many people only do this) must provide bad results."

"It is true that the methods have been developed more than 20 years ago and we had thought of supplementing our methods by quantum mechanical approaches. However, with the advent of combinatorial chemistry people have to deal with really large datasets - and we have used our charge calculations on datasets with millions of structures! Thus, we have been working with these methods over all the years and we have even reimplemeted it recently in C++ as interest in fast calculation of charge continues to go strong. Our newest approach is based on a Huckel method that modifies the parameters based on the sigma charges."

"This approach is contained in our program package PETRA for the calculation of physicochemical effects in organic molecules. You can access these calculations from our web site (http://www2.chemie.uni-erlangen.de) and use it on molecules of your interest. There is also a manual of the PETRA package on our web site. The PETRA package is also commercially distributed by the company Molecular Networks (info@mol-net.de)."

1970S: 3D STRUCTURE BUILDERS

In 1972, Wipke et al. published PRXBLD, the first program that was able to generate a 3D model rapidly from a 2D drawing with stereochemistry.[106]

1980S: COMPUTER TECHNOLOGY

The 1980s experienced revolutions of computer technology. The invention of very large scale integration techniques led to dramatically reducing the size of computers. In 1981, IBM introduced the first Personal Computer (PC) based on Intel's processor and Microsoft's DOS 1.0.[107] In 1983, Apple introduced a revolutionary personal computer, Lisa, which was the first computer with a graphical user interface.[108] In 1985, Microsoft shipped Windows 1.0.[109] The PC revolution in the 1980s placed computers directly in the hands of millions of people. This decade also saw the development of large scale computer networks and the coining of the term "internet".[110]

It should also be mentioned that in 1983, Bjarne Stroustrup at Bell Labs developed the C++ programming language as an enhancement to the C programming language.[111] The major difference between C and C++ is in that C++ is an object-oriented programming language while C is not. Since the 1990s, C++ has been one of the most popular commercial programming languages.

1980S: DATABASE SYSTEMS

In the 1980s, there was important progress in the area of database systems. In 1981, Lynch et al. published their first article on the representation and searching of Markush structures.[112] Later, Lynch's group published a series of over 20 papers on this complicated subject, laying an important foundation for the development of two patent database retrieval systems supporting substructure searching—the Markush DARC system from Derwent[113] and the MARPAT system from CAS.[114]

In this decade, fragment-based screening systems were widely adopted and dramatically improved the performance of searching large structure databases.[15] In 1987, Bruck et al. described a method for substructure search on very large files using tree-structured databases that did not reply on fragment-based screening systems.[115]

In the 1980s, Willett's group reported their work on pharmacophoric pattern matching in files of three-dimensional chemical structures, including the selection of interatomic distance screens,[116] the evaluation of search performance,[117] and the comparison of geometric searching algorithms.[118] In 1987, Brint and Willett reported their studies on several algorithms for the identification of three-dimensional maximal common substructures.[119]

In 1988, Downs el al. described their work on transputer implementations of chemical substructure searching algorithms.[120] The transputer (transistor computer) was the first general purpose microprocessor designed specifically to be used in parallel computing systems developed in the 1980s.[121]

One of the most important new developments in the database area in this decade was the establishment of chemical reaction databases and related management systems. In 1980, Willett described an efficient procedure for reaction-site detection based on a maximum common subgraph isomorphism algorithm.[122] In 1982, MDL released REACCS (Reaction ACCess System), a program to store and retrieve chemical reactions and related data.[81] In 1985, MDL released ORGSYN and Theilheimer, its first synthetic chemistry databases. In 1988, Moock et al. extended REACCS to include several new features, including an atom–atom mapping algorithm and a reaction-based similarity searching approach, etc.[123] These new developments established REACCS as a vital tool for organic chemists. Several other reaction database systems were also developed in this decade, such as Synthesis Library (SYNLIB) by Chodosh and Mendelson,[124] Organic Reaction Access by Computer (ORAC) by Johnson,[125] and CASREACT from CAS.[22] In 1986, Gay described the RMS-DARC reaction management system.[126] In 1984, Picchiottino et al.[127] described their work on designing specific reaction data banks by modeling data in a logical scheme on the basis of the Entity/Relationship approach. In this system, data validation was recognized as an important step in the computer acquisition of data. Four types of validations were used: item validation, interitem validation, batch validation, and database dependent validation.

In 1989, Gasteiger and Weiske[128] described the ChemInform Reaction Database. Professor Gasteiger was the project manager for FIZ Chemie, Berlin, and the essential software development was carried out by ChemoData, a software company founded by Gasteiger's two former students, Heinz Saller and Peter Löw. ChemoData is the predecessor of InfoChem,[129] which distributes chemical structure and reaction databases and develops innovative chemical information systems, particularly for publishing houses such as Springer, Wiley-VCH, and Thieme Verlag. Presently, the ChemInform Reaction Database is the most widely distributed in-house reaction database. It is produced by FIZ Chemie, Berlin and distributed by Elsevier MDL, San Roman, CA.

One of the most exciting events in the development of chemical database systems in this decade was the conversion of the well-renowned *Beilstein Handbook of Organic Chemistry* into computer-readable form and its availability as an online database.[130] The Beilstein Database Project was started in October 1983.[131] Its goal was to build the world's largest organic factual database.[132] The electronic version of the *Beilstein Handbook of Organic Chemistry* was designed and developed by Reiner Luckenbach and Clemens Jochum. Dr. Jochum was the first student that obtained his Ph.D. with Professor Gasteiger. Professor Gasteiger himself was a consultant to the Beilstein Institute. The Beilstein Database has since become one of the most widely used organic chemical information databases. It is interesting to note that the word "Beilstein" has often been regarded as synonymous with high quality, reliability, and comprehensiveness. This reputation has been achieved through the implementation of a number of quality control mechanisms at all production stages involving the creation of all Beilstein products, including the application of both manual data selection processes and several sophisticated automatic checking approaches to each piece of data.[133]

There are several noteworthy developments in the spectral database area in this decade. In 1980, Milne and Heller described the NIH/EPA Chemical Information System for searching spectral databases.[134] Also in 1980, Gray et al.[135] described the first general applicable mass spectrum prediction method—DENDRAL's half-order theory. In 1981, Sadtler Research Laboratories entered into agreements with IR instrument manufacturers for collaborative development of search programs and an associated database to be used in conjunction with the computers that control the instruments, as described by Shaps and Sprouse.[136] In 1982, Lindley et al.[137] described a computerized approach to the verification of ^{13}C NMR spectral assignments based on the spectrum prediction. Also in 1982, Milne et al.[138] described the details of the method for determining the Quality Index (QI) from quality factors (QF) with the financial support from the U.S. Environmental Protection Agency (EPA). The QI algorithm was used to examine a mass spectrum for the occurrence of the standard errors. In addition, several major programs for mass spectrum search were described in the 1980s: MassLib was described by Domokos et al. in 1983.[139] In 1985, McLafferty et al. described the further development of their STIRS system.[140]

In 1985, Kalchhauser and Robien reported the CSEARCH NMR database system for identification of organic compounds and fully automated assignment of ^{13}C NMR spectra.[141]

Traditionally, building a spectral database is a multistep process, involving some manual operations. Errors may be introduced into the spectral data at each of these steps. For example, some spectroscopic data are still abstracted from the literature. This approach is acceptable for building databases of ^{13}C NMR-like spectra which can be represented as peak tables but not suitable for producing databases of IR-like spectra because each such a spectrum requires at least a few thousands of data points to be collected. On the other hand, among the total number of known compounds (e.g., CAS databases currently contain over 25 million compounds), only a tiny fraction of them have been associated with a certain type of spectral data. For example, the largest worldwide commercially available SpecInfo collection contains only 200 000 ^{13}C NMR spectra.[142] To overcome the above problems, in 1998, Neudert described a new electronic publishing project which was carried out by Wiley-VCH/Chemical Concepts with financial support from the German Ministry of Research.[143] Spectroscopic data acquired from the spectrometer are first converted into the JCAMP (Joint Committee on Atomic and Molecular Physical Data) format.[144] The TranSpec program[145] is used to input or import structures in MDL CTFile format.[146] Quality control modules are used to check the reliability of the information. The data in JCAMP and MDL format are then sent to the data server by e-mail. Finally, spectral software is used to automatically extract the data from the e-mail and register them into the database. It should be pointed out that in the above work flow, the quality control module plays an important role.

In another development, CAS and FIZ Karlsruhe created STN International, a scientific and technical information network in 1984. Three years later it was linked with JICST, Tokyo.[22] It allowed scientists to perform remote searches on major scientific reference databases. In 1987, STN Express, a front-end software package for searching databases on STN, was launched.[22] In 1988, Shenton et al. described the Markush DARC system from Derwent for searching databases of patents.[113] In 1989, Wade et al. reported the Sandwich Interactive Browsing and Ranking Information System (SIBRIS),[147] which gives bigger weight for the terms that appear in the title than those terms that appear in only the text of the document.

It should also be mentioned that in 1988, Weininger described the SMILES (Simplified Molecular Input Line Entry System) notation used in the Pomona College MedChem project.[148] SMILES has since become one of the most widely used modern liner notations.

It is also interesting to note that in 1985 the first commercial sale of Rubenstein's ChemDraw was made to Stu Schrieber, then at Yale University, a year earlier than ChemDraw 1.0 was released (in 1986), which has since become the most popular chemical structure drawing program.[149]

1980S: CASE EXPERT SYSTEMS

The major CASE expert systems that appeared in the 1980s still relied mainly on the traditional one-dimensional spectral data. In 1980, Dubois et al. published the DARC-EPIOS system for computer-aided elucidation of structures by ^{13}C NMR.[150] In the same year, the PAIRS IR interpretation program was published by Woodruff and Smith.[151] In 1981, Carhart et al. reported the GENOA program,[152] which was the last in the series of structure generating programs developed through the DENDRAL project. Unlike the more standard structure generation procedure, GENOA searched for ways of combining each new piece of structural information with the results of previously specified structural constraints. The most interesting feature of GENOA was that it could utilize overlapping and alternative substructures. Also in 1981, Debska et al. described the SEAC structure determination system.[153]

In 1981, Gasteiger and Marsili described the work on prediction of proton magnetic resonance shifts.[154] In 1987, Neudert et al. reported the SpecInfo system for employing mass spectral data to draw conclusions on substructural features for the given unknown compounds for aiding the determination of its structure with the help of other spectroscopic data.[155]

The primary limitation of atom-centered codes for NMR spectrum prediction was that they usually represent just the topology of resonating atoms. In 1981, Gray et al. described a method for stereochemical substructure codes for ^{13}C spectral analysis.[156] The DENDRAL coding scheme represents a significant development of the Jezl/Dalrymple[66] and Bremser[68] methods in that it incorporates configurational stereochemistry.

Traditional structure generators were based on the structure assembly algorithm, that is, the complete structures are built by linking smaller substructral fragments together. In 1988, Chritie and Munk described the first structure generator based on the structure reduction approach.[157] This method begins with building one or more hyperstructures, each of which contains a set of all bonds, and then removes inconsistent bonds as structure generation progresses. It was implemented in their COCOA program. Like GENOA, COCOA can also use required, potentially overlapping substructures.

By the end of this decade, structure generator algorithms had been quite well established. The major impediment to their further development of the CASE systems was how to derive sufficient structural constraints from spectra of the unknown compound. A new attempt to overcome this problem was to utilize the newly emerging two-dimensional NMR spectra. In 1987, Christie and Munk described the program ASSEMBLE2D,[158] which can reduce the three-bond hydrogen–hydrogen correlations derived from the COSY experiments, given one-bond hydrogen–carbon correlations derived from an HMQC (heteronuclear multiple quantum coherence correlation spectroscopy) experiment. In 1989, Funatsu et al. also extended their CHEMICS program to include the 2D NMR spectral information, such as 2D INADEQUATE experiments. [159]

1980S: CASD EXPERT SYSTEMS

The traditional CASD systems were based on the retrosynthesis approach. In the 1980s, two new expert systems of this kind were reported. In 1985, Hendrickson et al. published the SYNGEN program.[160] The main design goal of this system was to be able to automatically generate the shortest synthetic route for a given target structure. In 1986, application of the CASP (Computer-Assisted Synthesis Planning) system in Sandoz AG, Switzerland was described by Sieber.[161] On the other hand, organic synthesis planning can also be performed in the forward synthesis manner—that is, given the starting material, how one can predict the products. One of the first computerized systems for doing this was the CAMEO program published by Salatin and Jorgensen in 1980.[162] CAMEO performed the prediction of reaction outcome based on mechanistic reasoning and is one of the best known reaction prediction systems.

In 1983, Hanessian reported his Chiron program.[163] The name Chiron was derived from “chiral synthon”. Chiron was an interactive computer program for stereochemical analysis and heuristic synthesis planning. The program consisted of five modules: CARS-2D (Computer Assisted Reaction Schemes), CASA (Computer Assisted Stereochemical Analysis), CAPS (Computer Assisted Precursor Selection), CARS-3D (3D drawing and simulation), and a fifth module allowing manipulation in real time and 3D visualization. The Chiron Program decoded the stereochemical and functional complexities of a given target structure and related them to structures or substructures derived from more than 200 000 precursors or starting materials.

In 1984, Wipke and Rogers reported the SST (starting material selection strategies) program for the interactive selection of potential starting materials given a desired target molecule.[164] The program was based on the superstructure search algorithm and used hierarchical searching to rapidly select candidates from a large starting material library. It also contained a function to evaluate the appropriateness of the functionality of the starting material. In 1988, Funatsu and Sasaki described the AIPHOS program, which employed both retro- and forward synthetic strategies.[165] AIPHOS was designed to propose synthetic schemes and then to utilize a built-in reaction predictor to check that the desired product would in fact be the favored one.

1980S: 3D STRUCTURE BUILDERS

Another hot research frontier in the field of chemoinformatics in the 1980s was the development of better 3D structure builders. The 3D structure builders can be grouped into two categories: one is rule-based and the other is database-based.

(a) Rule-Based 3D Structure Builders. In 1981, Cohen et al. presented the SCRIPT program.[166] This program uses symbolic logic to construct possible ring conformations from a table of single ring templates and directly translates these symbolic representations into 3D atomic coordinates, making the calculation very fast. However, 3D coordinates thus generated are rather crude and lack an energy evaluation of the conformations at the symbolic level of conformational diagrams. This program was applied to reaction design studies with some success. In 1987, Dolata et al. developed two programs, WIZARD and COBRA, for the systematic conformational analysis using symbolic logic and artificial intelligence methods.[167] They developed a set of rules for the construction of 3D models, which were derived from the conformational units with well-known optimum geometries, such as cyclohexane chair, and the entire 3D system was built by joining conformational units. Also in 1987, Pearlman published the first description of CONCORD.[168] This program generates the conformations of ring templates from templates and then prioritizes them according to strain energy. It uses a simplified force field to perform stepwise optimization as the analysis proceeds and thus is more time-efficient than WIZARD’s strain relief approach. CONCORD produces a single conformation. It was the first 2D-to-3D structure converter to be used on a large scale. Also in 1987, Hiller and Gasteiger published the first description of CORINA (COoRdINAtes).[169] The major advantage of CORINA is that it is capable of handling certain molecular systems that cannot be very well handled by other programs. CORINA uses a table to set bond lengths and angles. It can call upon various procedures to provide fine adjustment. It has special routines to handle rigid and flexible macrocyclic rings. The most recent version of CORINA was developed by Sadowski and Gasteiger.[170] CORINA is distributed by Molecular Networks GmbH. CORINA has become one of the most widely used 3D structure builders. Presently, there have been more than 60 installations worldwide, particularly in all major pharmaceutical companies.

(b) Database-Based 3D Structure Builders. The second type of 3D structure builders is based on finding near analogies of a structure or of substructures of it in a database of 3D molecular structures. The first example of such a program is the AIMB (Analogy and Intelligence in Model Building) described by Wipke and Hahn in 1986.[171] AIMB selected organic molecules with up to 64 heavy atoms as a subset of the Cambridge Structural Database to construct a knowledge base of fragments through a process of abstraction and averaging of coordinates. This knowledge base is then used to analyze individual molecules. AIMB could be considered to be the parent of the field of protein homology modeling, since the latter also utilizes analogy to known structures as the basis of the modeling approach.

1990S: COMPUTER TECHNOLOGY

In this decade the client/server revolution sought to link PCs (“clients”) with larger computers (“servers”) that served data and applications to client computers.[172] The 1990s also witnessed the revolution in mainstream software development from structured programming to object-oriented programming.

The most exciting invention in the computer technology field in the 1990s was probably the birth of the World Wide Web. In 1990, Tim Berners-Lee, a scientist at the European Particle Physics Laboratory (CERN) in Geneva, Switzerland, started work on a hypertext GUI browser+editor based on the “hypertext” proposal he wrote in 1989. He made up “WorldWideWeb” as a name for the program and “World Wide Web” as a name for the project.[173] In 1991, the general release of WWW on central CERN machines was launched. With this new interface of the Internet, augmented by the graphics-based Web browser, such as Mosaic (the first graphics-based Web browser introduced by the National Center for Supercomputing Applications (NCSA) at the University of Illinois in 1993) and Netscape 1.0 (in 1994), the use of the Internet exploded, leading to the Internet revolution.

Another interesting thing worth mentioning is that in 1991 a Finnish college student, Linus Torvalds, posted a message to the Usenet Newsgroup comp.os.minix: “Hello everybody out there using minix: I’m doing a (free) operating system (just a hobby, won’t be big and professional like gnu) for 386(486) AT clones.”[174] This humble beginning led to the birth of Linux, which has become one of the most widely used UNIX-like operating systems today. Linux is also a favorable operating system for many computational chemists and chemoinformatics specialists.

In 1995, Microsoft released Windows 95[175] and within 4 days the software sold more than 1 million copies. In the same year, Sun Microsystems introduced the Java programming language.[176] Quickly thereafter, Java has exceeded C++ and has become the most popular computer programming language nowadays.

1990S: DATABASE SYSTEMS

Several new databases were created in the 1990s. In 1990, MDL released the Substance Model to MACCS-II, providing a mechanism for handling polymers, mixtures, and formulations. In the same year, MARPAT File, an STN database that allows generic and/or variable query structures to match against generic and/or variable file structures (“Markush structures”) in chemical patents, was made available by CAS.[22,114] In the same year, Murrall and Davies described ChemDB3D, which was the first 3D structure searching system capable of handling conformational flexibility.[177]

In 1990, Mitchell et al. described their work on using graph matching algorithm to compare secondary structure motif in proteins.[178]

In 1991, MDL released its flagship product, ISIS (Integrated Scientific Information System), based on the client/server architecture, bringing distributed computing to scientific information management. In 1995, MDL developed two new products to handle new discovery technologies: Project Library, the first desktop program for combinatorial chemistry, and MDL SCREEN, for high-throughput screening. In 1996, MDL introduced Central Library for combinatorial chemistry data management, as described by Leland et al.,[179] and the Relational Chemical Gateway, bringing chemical structures into Oracle.

In 1992, Brown et al. reported their work on exploring the possibility of using a hyperstructure model for representing a set of chemical structures to improve the substructure searching performance.[180] In the early 1990s, Artymiuk et al. described work on using graph-theoretical techniques to study the three-dimensional structural resemblance between leucine aminopeptidase and carboxpeptidase (1992)[181] and between the ribonuclease H and connection domains of HIV reverse-transcriptase (1993).[182]

In 1992, Dalby et al.[146] described a series of chemical structure file formats used for storing and transferring chemical structure information that have evolved over several years at Molecular Design Limited (now Elsevier MDL), including the MOLfile for a single (multifragment) molecule, the RGfile for a generic query, the SDfile for multiple structures and data, the RXNfile for a single reaction, and the RDfile for multiple reactions and data. These files are built using one or more connection table (Ctab) blocks. The above file formats are collectively called the V2000 format and have since been widely used and become the *de facto* standard for representation and communication of chemical structure information. However, the FORTRAN language based V2000 format has some inherent limitations. To overcome those limitations, MDL introduced an extended molfile format called the V3000 format in 1996. It offers a number of advantages over the V2000 format: (1) consolidates property information for chemical objects, (2) provides better support for new chemical properties or objects, (3) removes fixed field widths to support large structures, (4) uses free format and tagging of information for easier parsing, and (5) provides better support for backward compatibility through BEGIN/END blocks. Most recently (2003), Elsevier MDL has introduced another new file format called XDfile (XML-data file), which uses a standard set of XML (Extensible Markup Language)[183] elements that represent records of data. The XDfile format also provides the following: (1) Metadata or information about the origin of the data. (2) The ability to handle generalized data models, such as multiple structures and nonstructure fields per record, multiple reactions per record, multiline data, and binary data. None of the other CTfile formats such as SDFiles or RDFiles have this ability. (3) Very few restrictions on data formatting within the actual content. Data formatting is based on XML, which does not have restrictions on line length or blank lines. (4) Fast and easy parsing by using any XML parser. The XML data can be validated by using a DTD (Document Type Definition) or an XML schema that defines primitive rules which the data must follow. (5) Flexibility in creating application-specific XML tags. The detailed description about each of the above file formats is given in the document called MDL CTFile Formats. The latest version of this document can be downloaded at the Elsevier MDL Web site.[82]

In this decade, the second generation of the substructure search technology based on the efficient tree-structured substructure searching methods for searching large structure databases was further developed and adopted by several database systems, such as Beilstein Database system[184] and MDL ISIS.[185]

In 1993, Martin et al. described Disco, the first pharmacophore mapping system.[186] Also in 1993, Grindley described how to use a maximal common subgraph isomorphism algorithm to identify tertiary structure resemblance in proteins.[187] In 1995, Artymiuk et al. described the PROTEP program for graph-theoretic similarity searching of the 3D structures in the Protein Data Bank.[188]

In 1995, Martin et al. described probably the first use of computational methods for designing structurally diverse combinatorial libraries.[189] In the same year, 1995, Hendrickson and Sander described a rapid reaction similarity search system—COGNOS,[190] which was later renamed WebReactions.[191] Also in 1995, SciFinder, a software and information package for accessing CAS databases, was released by CAS.[22] In 1996, STN Easy, a Web site for easy searching of selected STN files, became available.[22]

In the 1990s, two solid-phase synthesis databases for combinatorial chemistry were produced: SPORE (Solid-Phase Organic REaction) from MDL and SPS (Solid-Phase Synthesis) from Synopsys Scientific Systems together with Oxford Diversity.[192] Also in this decade, InfoChem released a reaction type database—ChemReact41, containing 41 000 core reaction types, and a synthesis strategy design tool called Synthesis Tree Search (STS) containing 2.5 million reactions.[193] The InfoChem's reaction database now contains over 3.6 million reactions published since 1974.[194]

A new trend in the area of database systems was the development of data mining technologies for automatic extraction of knowledge from the existing databases. In the early 1990s, Chen and Robien significantly enhanced the CSEARCH NMR database system by introducing into it several intelligent modules, such as a new data mining module based on a fast MCSS algorithm[195] for automatic deduction of common structural features from a set of structures obtained using ^{13}C NMR spectral similarity search;[196] a new module for automatic extraction and analysis of substituent-induced chemical shift differences of ^{13}C NMR Spectra—the SCSD (Substituent-induced Chemical Shift Differences) algorithm;[197] and a novel ^{13}C NMR spectral prediction module—the OPSI (Optimized Prediction of ^{13}C NMR Spectra using Increments) method.[198] It is interesting to note that according to Chen and Robien's study, for a chemical structure there may exist, in theory, an infinite number of additivity models, and the classic additivity model[48] is only the worst special case of the OPSI method. The OPSI method is capable of dynamically generating optimized additivity models for a given structure.

To produce high quality NMR databases for the CSEARCH NMR database system,[141] Chen and Robien developed several methods for automatic detection of database errors.[199] (a) The method for detection of database errors by comparison of different spectra of a compound[200] was designed for the exact match of two identical structures with respect to their two-dimensional topology allowing a very detailed comparison. This method is definitely based on redundant data within the database itself and can only be applicable to those cases where redundant spectra of a compound exist in the database under investigation. (b) The approach based on the SCSD-algorithm[197,201] investigates database errors by comparison of all the possible similar structure pairs and their spectra in the database under consideration and further extracting and analyzing the extreme SCSD values. Unlike method (a), this technique is not restricted to identical structure pairs. (c) Another alternate method for error-detection is based on the OPSI spectrum-prediction method[198] and comparison of experimental and estimated chemical shift values using algorithms for automatic resonance line assignment.[141,202] Like the method (b), this technique is also not restricted to identical structure pairs. The above methods were implemented into the CSEARCH system for the automatic error-detection within a data collection of some 85 000 NMR-spectra including the libraries of the University of Vienna, SADTLER Research Laboratories and the German Cancer Research Center at Heidelberg.

In 1997, Allen and Hoy described IsoStar,[203] the first knowledge-base library from the Cambridge Crystallographic Data Centre (CCDC). It incorporated extensive and systematic information on noncovalent interactions derived from the CSD and the Brookhaven Protein Data Bank (PDB database), together with selected interaction energies computed using *ab initio* MO methods. IsoStar can be used in rational drug design and crystal engineering applications. In 1999, Durant et al. described Cheshire, a new scripting language and its use in characterizing databases.[204]

In the 1990s, Chemical Abstract Service (CAS) explored data mining of CAS Registry and CA File data for substance-use relationships.[22] The basic idea was to input one component of the substance-use pair and have the system automatically predict the other component. The basic procedure includes database searching, summarizing and analyzing the search retrieval, developing a prediction model, and predicting properties of a substance or new substances with a specific property. CAS also explored data mining of its CASREACT reaction database in terms of functional group transformations. Each single step reaction is encoded as one or more n-tuples describing a functional group transformation and comprised of one formed product group from 1 to 4 reactant groups that are transformed to the product group and a code indicating how the reactant groups are distributed among the reacting substances, leading to an exhaustive index to the database and a highly descriptive summary of the reaction classes in the CASREACT database.

In the 1990s, Ihlenfeldt et al. developed CACTVS,[205] a unique chemistry algorithm development environment. (Ihlenfeldt was a former student with Professor Gasteiger, obtaining his Ph.D. with him.) It should be pointed out that CACTVS is not a pure database retrieval and management system. It is a distributed client/server system for the computation, management, analysis, and visualization of chemical information of *any*, even dynamically and ad hoc defined type. CACTVS uses a worldwide network of databases with property descriptions, computational modules, data analysis tools, visualization servers, data type handlers, and I/O modules to achieve unlimited extensibility of its capacities. The system consults its network of databases to retrieve information about the necessary steps to obtain some kind of knowledge about a structure.[206] It should be noticed that on the basis of the CACTVS system, the database of the National Cancer Institute has been implemented on the Internet. With 250 000 structures it is presently the largest chemistry database accessible free of charge.

In 1997, Gardiner described their work on using clique-detection algorithms for matching three-dimensional molecular structures.[207] In 1998, Cosgrove and Willett described the SLASH program, which was designed to analyze large numbers of compounds in terms of the functional groups they contain.[208]

It is also interesting to note that the 1990s saw the revolution in the field of genomics and the birth of the biotech genomics industry,[209] which will have significant impact on the future development of chemoinformatics.

1990S: CASE EXPERT SYSTEMS

In the 1990s, there were some significant developments in the field of CASE. In 1990, a very fast structure generator, MOLGEN, was published by Kerber, a mathematics professor at Bayreuth University, Germany.[210] Also in 1990, Chen and Zhang described the ESIR (Expert System for interpretation of IR spectra) system,[211] which was capable of giving detailed explanations on its reasoning process, if requested.

In 1990, Hanebeck et al. described a method for the automatic generation of a mass spectrum from the structure of a compound.[212] In 1995, Schulz and Gasteiger described their work on the elimination of candidate structures in computer-assisted structure elucidation using the mass spectrum.[213] In 1996, Selzer et al. described a method for simulation of IR spectra with neural networks using the 3D-MoRSE code.[214] Also in 1996, Steinhauer et al. described a method for obtaining the 3D structure from infrared spectra of organic compounds using neural networks.[215]

In 1991, Christie and Munk expanded the SESAMI system to include 2D NMR.[216] In 1994, Funatsu et al. extended CHEMICS in such a way that it can utilize NOE data and distance geometry methods to create 3D coordinates.[217] This allows it to be able to distinguish between diastereomers and to provide some information about conformation behavior. Also in 1991, Bohanec and Zupan described a structure generator based on the combination of both structure reduction and structure assembly strategies.[218] In 1994, Peng et al. described the CISOC-SES system for aiding structure determination of complex natural products, which used a hyperstructure-based structure generation algorithm.[219] CISOC-SES can effectively employ 2D NMR correlation information, such as the long-range hydrogen−carbon correlations (HMBC experiment).

In 1996 Schuur et al. reported a molecular transform that allowed the representation of the 3D structure of a molecule by a fixed number of values and showed its usefulness for the simulation of IR spectra.[220] The approach was further elaborated, and the method is now available on the Internet.[221]

In 1996, Faulon[222] described a stochastic structure generator which used simulated annealing to search the space of constitutional isomers.

In 1999, Elyashberg et al. described a new expert system called Structure Elucidator for structure elucidation using a two-stage approach.[223] In the first stage, the program tries to automatically determine the structure of the unknown compound using fragments taken from a library of ^{13}C NMR spectra. It generates the structures by joining fragments that have common atoms. If the structure cannot be determined in the first stage, the system derives the molecular formula from the molecular mass and then forms the sets of fragments. The system generates the structures from non-overlapping fragments using another structure generator. The most preferable structure is selected by predicting ^{13}C and ^{1}H NMR spectra.

1990S: CASD EXPERT SYSTEMS

In the early 1990s, several CASD systems were described. In 1990, Röse and Gasteiger described their work on automatic deduction of reaction rules which were used in the reaction prediction system, EROS 6.0.[224] In 1990, Weise reported the SYNTHON program which worked by “synthon replacement”.[225] Also in 1990, Gordeeva et al. described COMPASS,[226] which utilized a more formal description of the reactions that facilitates the use of combinatorial methods with a small but powerful knowledge base. In 1992, Hippe et al. described SCANCHEM.[227] Also in 1992, Dogane et al. published the SYNSUB-MB system, which does not require guidance from a user.[228]

In 1994, two CASD programs were reported. The Sello’s LILITH program includes methods for improving the accuracy of the predictions by detecting similar group interferences.[229] Moll’s TRESOR was based on the synthon approach.[230] The reaction knowledge is organized from low level reaction to higher level synthon to make the search more efficient.

In 1996, Barberis et al. described the HOLOWin program,[231] which aimed at searching for the key step, i.e., the fundamental step, in a complex organic synthesis. In 1998, the SYNCHEM program was upgraded by Krebsbach et al. to execute in a network of multiprocessor workstations under both Linda tuple space and PVM message passing protocols.[232] Although as early as in 1984 Wipke and Rogers had already described an algorithmic parallel solution for backtracking method,[233] SYNCHEM is probably the first CASD system capable of parallelization. In 1996, Lingran Chen and Hendrickson updated the SYNGEN system by porting it from the old VAX/VMS computer to modern platforms. The most attractive feature of the SYNGEN program is that it can generate most economic synthesis routes for a given target. Their effort was aimed at applying the system to green chemistry, a newly emerging branch of chemistry for making chemistry more environmentally benign. Hendrickson’s group has made the most recent version of the SYNGEN program available from their Web site.[234] In 1998, Mehta et al. reported a simple program called SESAM for unraveling “hidden” restructured starting material skeletons for complex target structures.[235] The aim is to help the chemist in exploring new synthetic strategies for complex targets by searching simple, nonobvious but effective starting materials. The program works at the skeletal level and attempts to map simply modified starting materials into the skeletal surface of the target.

One of the most interesting CASD systems developed in the 1990s is probably the WODCA (Workbench for the Organization of Data for Chemical Applications) program developed in Gasteiger’s group.[236] WODCA is a computer program for the interactive planning of organic syntheses. The objective of WODCA is to assist the organic chemist in each step of the synthesis planning process. WODCA provides a series of methods and tools for chemical applications, in particular, for the planning of organic syntheses in a retrosynthetic approach. The connection between the target compound and available starting materials is achieved via methods of similarity searching. The algorithm for the perception of strategic bonds is based on physicochemical effects on atoms and bonds, which are calculated by published empirical methods incorporated in WODCA. This system is regarded as a second generation synthesis planning system and promises to be an important contribution in the future.[237] The WODCA system is distributed by Molecular Networks[238] and is in practical use at a few chemical companies.

Nowadays, large sets of molecular structures can easily be generated using either computational techniques or combinatorial synthesis. And thus, methodologies for automatic prioritization of molecular structures are of importance in many applications. In 1995 Gillet et al. described the CAESA (Computer Assisted Estimation of Synthetic Accessibility) program.[239] CAESA automatically ranks sets of molecules according to their ease of synthesis and displays the synthesis route. It estimates synthetic accessibility of target compounds based on a reaction knowledge base and available starting materials.[240]

A major difficulty in the development of both retrosynthesis planning and forward reaction prediction systems is the building of reaction knowledge bases. Both the model-driven and data-driven approaches for reaction classification and generalization have been explored.[241] Most of those methods are structure-topology based. An inherent limitation of such classification methods is that they cannot group together those reaction instances that have different functional groups even though these groups exert the same physicochemical effects. Another limitation common to all the reaction classification methodologies is that the classification is essentially one-dimensional. No information about the relationships between different reaction types is available. To overcome the above problems and for building reaction knowledge bases used in the EROS reaction prediction system and the WODCA synthesis planning system, Chen and Gasteiger developed the first two-dimensional method for reaction classification and prediction based on the combination of the Kohonen neuronal network and physicochemical descriptors for characterizing reaction centers in the middle of 1990s.[242] The above work has become the foundation of Gasteiger's CORA (Classification of Organic Reactions for Applications) program for deriving knowledge on chemical reaction from reaction databases.[243]

1990S: 3D STRUCTURE BUILDERS

In 1990, Chemical Design Ltd., which was purchased by Oxford Molecular and now is a part of Accelrys, developed a 2D-to-3D builder called Chem-X builder.[244] It is based on a relatively small library consisting of specific common carbocyclic and heterocyclic substructures together with generalized fragments in which the atom types are unspecified. Similar to the AIMB program by Wipke and Hahn,[171] the fragments used in the Chem-X builder were retrieved from a database.

In 1993, Gothe et al. described the generation and use of three-dimensional structures for computer assisted mechanistic evaluation of organic reactions.[245]

2000S: COMPUTER TECHNOLOGY

The year 2000 was a big test on computer technology. Many people feared that January 1st, 2000 could cause serious problems because many old computers record only the last two digits of a year, such as 95 for 1995. When the year 2000 came the computer's clock would set to 00, causing the computer to think it was 1900, not 2000. This Year 2000 bug, or Y2K problem, caused many individuals to fear for the worst. Fortunately, because of good preparation the Year 2000 caused no catastrophes at all.

Microsoft introduced C# (pronounced C-sharp) to the public in 2000 with the introduction of .NET. Although it is generally said that C# is a new object oriented language derived from C and C++, C# is actually more similar to Java than to C or C++: It borrowed several important features from Java, such as garbage collection. C# helps developers create XML Web services and Microsoft.NET-connected applications for Windows and the Internet. In 2001, Apple introduced Mac OS X 10.0. Mac OS X is based on the Mach kernel and the BSD implementation of Unix and is almost completely independent of the earlier Mac OS releases.[246] In the same year, Microsoft released the Windows XP operating system. XP is a whole new kind of Windows for consumers. It contains the 32-bit kernel and driver set from Windows NT and Windows 2000. It also offers many new features that no previous version of Windows has. Furthermore, it also supports old DOS and Windows programs, which may even run better on XP than on previous Windows. From the mid-1970s to 2002, approximately 1 billion PCs had been shipped worldwide according to a study released by the consulting firm Gartner.[247] In January 2006, Apple began shipping Macintosh computers with Intel x86 microprocessors rather than traditional PowerPC microprocessors.

There are two new trends worth mentioning in the current PC market. One is that, as of 2004, 64-bit CPUs are common in servers and have recently been introduced to the (previously 32-bit) mainstream personal computer arena in the form of the AMD64, EM64T, and PowerPC 970 (or "G5") processor architectures.[248] The other is that the multicore processor-based PC is starting to enter the mainstream PC market. In general, multicore microprocessors allow a computing device to exhibit some form of thread-level parallelism (TLP) without including multiple microprocessors in separate physical packages.[249] The main reason for this change is that the major processor manufacturers, such as Intel and AMD, have run out of room with most of their traditional approaches to boosting CPU performance. Instead of driving clock speeds and straight-line instruction throughput ever higher, they are instead turning en masse to multicore architectures.

2000S: DATABASE SYSTEMS

Failed reactions are valuable information that allows one to learn from other people's experiences and not to repeat previous mistakes. However, those reactions are seldom incorporated into reaction databases. Some examples of such failed reactions can be found in the CASREACT database.[250] Accelrys' database of Failed Reactions, first released in 2000, is probably the only database that contains solely failed reactions.[251] These reactions can be grouped into three categories: unexpected product, immediate further reaction, or simply no reaction. The failed reactions were abstracted from literature. The v.2003.1 release of the failed reaction database contained 13 500 reactions and is updated semiannually with ca. 2000 reactions/year.[252]

In 2001, He et al. described a traditional Chinese medicine database (TCMD) system,[253] which allows one to study traditional Chinese medicines and exchange related information through the World Wide Web. The program is based on ISAPI (Microsoft Internet Server Application Programming Interface), VRML (Virtual Reality Modeling Language), and JavaScript.

In 2002, Ihlenfeldt et al.[254] described a Web-based graphical user interface for conducting rapid searches by numerous criteria in the more than 250 000 structures of the Open NCI Database.[255] It is based on the chemistry information toolkit CACTVS.[205] The user can conduct 3D pharmacophore queries in up to 25 conformations precalculated for each compound. Only a Web browser is needed to use this service. The database includes nearly all structures and anticancer and anti-HIV screening data provided by NCI's Developmental Therapeutics Program. This data set has been augmented by a large amount of additional data, such as calculated logP values and predicted biological activities. Also in 2002, MDL launched a new Internet service called DiscoveryGate.[256] It provides single point access to MDL's renowned collections of synthesis, bioactivity, physical property, metabolism, toxicity, and chemical sourcing databases (over 11 million structures, 200 million associated facts, and 10 million reactions).

In recent years, Gasteiger's group has converted the information on the Poster "Biochemical Pathways", which was originally produced by Boehringer Mannheim (now Roche), into a reaction database. This project was pursued in cooperation with Spektrum Akademischer Verlag, Lion-Bioscience, and the Universities of Mannheim and Passau. Biochemical Reactions can now be analyzed with all modern structure and reaction search methods.[257] In 2003, Chen et al. described a new generation of reaction indexing and searching methodologies based on the concept of reaction hyperstructure.[305]

In 2003, Rhodes et al. described CLIP (Candidate Ligand Identification Program), a program for 3D similarity searching.[258] CLIP uses the Bron-Kerbosch clique detection algorithm to find those structures in a file that have large structures in common with a target structure. Also in 2003, Holliday et al. described a method for calculation of the similarities between pairs of substituents on ring systems using R-group descriptors.[259] It should be noticed that this similarity measurement is a local one. In the same year, Gaizauskas et al. described the PASTA system for extracting protein structures and information from biological texts.[260]

Also in 2003, Steinbeck et al. described NMRShiftdatabase,[261] an open-source, open-content database for chemical structures and their NMR data. It is based solely on free software and allows for open submission and retrieval of data sets by its user community. Both the software and the content are freely distributable.

In 2004, Elsevier MDL released MDL Isentris.[262] It is a new-generation technology platform for delivering and extending integrated scientific applications. MDL Isentris is based on open programming standards and multitier architecture. It contains four main components: MDL Base and MDL Draw—the client applications that deliver the power of the Isentris platform to the user; MDL Core Interface—the middle tier that provides the unifying logic and data integration for all applications built on Isentris; and MDL Direct—the server tier that provides specific support for storing, searching, and retrieving molecules and reactions, which was based a new reaction substructure search (RSS) algorithm developed recently by Chen et al.[263]

As mentioned previously, the algorithms for automatic detection of reaction sites have played an important role in the development of reaction databases. However, it should be noticed that such methods are far less than perfect. Many organic reactions have multiple alternative products, many of which are unbalanced reactions (missing some reactant(s) and/or product(s)). It is still a challenging task to deal with some complicated reactions belonging to the above category, which affects the quality of reaction databases. It is interesting to note that MDL's new RSS algorithm[263] has a special feature that in some cases RSS cannot match a reaction to itself because the reaction contains errors in some reacting centers. Although it cannot guarantee that it will not match a reaction to itself if the reaction has any reacting center errors, RSS can still detect some of the reactions that do contain such errors. This feature can be used to automatically detect reaction database errors.

In 2004, Elsevier MDL released MDL Patent Chemistry Database which contains more than 1.5 million structure-searchable chemical reactions, 1.6 million substances from organic chemistry, and life science patents since 1976. Production of chemical reaction databases is a multistep process, with the possibility of errors at each of these steps. Most recently, Durant et al.[264] described a tool called VET for trapping errors in the chemical reactions identified as a part of this process. VET has been designed to minimize the acceptance of incorrect reactions, while still supporting various common practices in reaction depiction, including unbalanced reactions, suppressed components, and reactions with alternative products.

Several free databases were launched in the 2000s. In 2004 NIH rolled out its publicly available chemical structure database PubChem.[265] PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative.[266] PubChem provides a fast chemical structure similarity search tool and three databases: PubChem Substance, PubChem Compound, and PubChem BioAssay, which are part of the NCBI's Entrez information retrieval system.[267] Those databases are linked to each other and also to other Entrez databases. For example, PubChem's chemical structure records have links to PubChem's bioassay database, presenting the results of biological screening. They also have links to PubMed scientific literature and NCBI's protein 3D structure resource.[268] Most recently (2006), the PubChem database of small molecule data has been cross-indexed with the Compound Index hosted on Elsevier MDL's DiscoveryGate platform,[256] making it easier and quicker for researchers to obtain more comprehensive information. By the end of year 2005, the PubChem database already contained over 850 000 compounds. It is interesting to note that the PubChem chemical structure database has drawn the concern of American Chemical Society (ACS). ACS sees PubChem as a competitive product to its CAS Registry, which currently contains more than 25 million compounds collected mainly from chemical and patent literature.[269]

In 2005, Irwin and Shoichet described ZINC (ZINC Is Not Commercial), a free database of 2.7 million commercially available compounds, prepared for use in docking programs.[270] It provides a much-needed resource for scientists in search of drug discovery leads. In 2005, Girke et al. reported ChemMine,[271] a compound mining database for facilitating drug and agrochemical discovery and chemical genomics screens. ChemMine presently contains over 2 million compounds from public and commercial resources.[272] The database is divided into a public compound mining and an internal screening domain. Registered users can access the screening domain of the database. All validated screening data will be released to the public soon.

In April 2005, IUPAC (The International Union of Pure and Applied Chemistry) released version 1 of its International Chemical Identifier (InChI).[273] InChI is an Open Source, nonproprietary, public-domain identifier for chemicals. It is encoded in a string of characters uniquely representing a specified molecular structure. It is a precise, robust, IUPAC-approved chemical substance tag, independent of the way the chemical structure is drawn. InChI can be indexed by Internet search engines. It is usable in both printed and electronic data sources. It enables reliable structure recognition and easy linking of diverse data compilations. Version 1.0 of the InChI Identifier expresses chemical structures in a standard machine-readable format, in terms of atomic connectivity, tautomeric state, isotopes, stereochemistry, and electronic charge. It deals with neutral and ionic well-defined, covalently bonded organic molecules and also with inorganic, organometallic, and coordination compounds.

It is interesting to note that open access databases as well as open access journals are among the latest examples of the Internet/World Wide Web revolution.

2000S: CASE EXPERT SYSTEMS

In 2001, Steinbeck reported a program package SENECA for computer-assisted structure elucidation of organic molecules.[274] It attempts to find the constitution of an unknown compound from spectroscopic data, in particular from NMR. SENECA is based on Jean-Loup Faulons stochastic structure generator and is guided to an optimum by simulated annealing.[222] It is claimed to be particularly capable of searching larger constitutional spaces than the common deterministic algorithms.

In 2002, Fontana and Pretsch described a new rule-based spectra interpretation system,[275] which directly processes spectral files generated by spectrometers. For small and medium sized molecules, the new program is capable of automatically reducing the solution space to under 3%. Also in 2002, Aires de Sousa et al. described the prediction of ^{1}H NMR chemical shifts using neural networks.[276] In 2004, Da Costa et al. described structure-based predictions of ^{1}H NMR chemical shifts of sesquiterpene lactones using neural networks.[277]

In 2003, Korytko et al. published a new structure generator named HOUDINI,[278] which embodies a new concept of convergent structure generation for addressing limitations of earlier methods. It uses a single integrated representation of the collective substructural information, employs parallel atom groups for efficient processing of families of alternative substructural inferences, and relies on a managed structure generation procedure to build required structural features early in the process. The HOUDINI-based SESAMI structure elucidation system appears to demonstrate a greater capacity for efficiently solving the structures of large, complex compounds than their earlier structure generator, COCOA, based SESAMI system.[279]

ACD's Structure Elucidator (StrucEluc) has recently been extended to utilize 2D homo- and heteronuclear correlation[280] as well as mass spectra.[281] It now also allows a chemist to utilize fragments stored in a fragment database as well as user-defined fragments in the structure elucidation process. 2D NMR data of COSY, HMQC/HSQC, and HMBC have become the primary form of spectral data used for structure elucidation nowadays. The common lengths of the connectivities, 1-bond for COSY and 1- or 2-bond for HMBC derived from 2D NMR data, are set as the default in the StrucEluc system (formerly, ACD/Structure Elucidator). However, if COSY and HMBC connectivities of lengths greater than those default values exist, contradictions can appear. In 2004, Molodtsov et al. described algorithmic methods for the detection and removal of such contradictions in 2D NMR data.[282] The methods are based on the analysis of molecular connectivity diagrams, which are claimed to be able to detect 90% of contradictory cases and to automatically remove 50% of these problems. These approaches and a method of "fuzzy" structure generation in the presence of contradictions have been implemented in the StrucEluc system.

2000S: CASD EXPERT SYSTEMS

In 2000, Höllering et al. reported the further development of the EROS reaction prediction system.[283] In the most recent version of EROS 7.0, the knowledge base and the problem solving techniques have been completely separated. The knowledge base consists of two parts: the methods for calculating important electronic and energy effects in molecular structures and rules for evaluating the course of elementary chemical processes. Presently, EROS can be used to investigate the following organic reaction types: (1) electrophilic substitution, (2) amide hydrolysis, and (3) general hydrolysis.[284] Another interesting program for aiding organic synthesis is the SystematiChem system from SysChem.[285] Unlike the traditional CASD systems, SystematiChem uses neither bond strategies in an attempt to break up the target compound nor general reactions. Rather, it builds synthetic routes for a target compound through searching reaction databases.

In 2004, Rucker et al. reported that there existed certain correlations between the LHASA rules for finding strategic bonds in polycyclic target structures and indices of molecular complexity, at least for the more general rules.[286] This result indicates that it is possible to identify the bonds most useful for retrosynthetic disconnection by a simple calculation rather than by application of a body of rules and thus has potential application in the development of CASD systems.

In 2005, Zhu et al. reported a new tool for aiding synthesis planning based on a superstructure searching algorithm.[287] The program first uses a molecular key screening approach to retrieve a set of generic reactions from a reaction knowledge base and then uses a superstructure searching algorithm to compare the target molecule with the product of each generic reaction candidate obtained from the screening. The system employs two other approaches to enhance the generic reaction searching. One is the preprocessing of the target structure using a rule-based approach for strategic bond recognition. The other is the R-group similarity comparison between target and the products of the generic reaction candidates.

Prediction of synthetic feasibility/accessibility of molecular structures has become a useful technique during drug discovery. Most recently, Schürer et al. described a novel approach to direct the exploration of chemical space in an effort to balance synthetic accessibility and medicinal relevancy prior to experiments.[288] The reaction sequences are dynamically generated from reaction transformations, with empirical reactivity and compatibility data, and commercially available starting materials. The reaction sequences are evolved and optimized using a genetic algorithm, which leverages fitness functions based on predicted properties of generated molecular products.

In 2005, Socorro et al. reported an interesting program for prediction of organic reactions called ROBIA.[289] The ROBIA system not only employs the traditional rule-based mechanism and reaction transformations for reactive sites perception and intermediates generation but also has quite comprehensive filters to screen for the best probable products based on rules, 3D conformation search, molecular mechanics, and even quantum chemistry calculation. It is a new trend to integrate the traditional chemoinformatics technologies with the traditional computational chemistry methods in recent years. ROBIA is a new example of this type of hybrid systems.

2000S: 3D STRUCTURE BUILDERS

It seems that there has been no active work on the development of 3D structure builders since year 2000. This is a good indication that the existing 3D structure builders have already been well established and served the needs of many applications. In 2000, Schönberger et al. described their work on molecular modeling of fullerene dendrimers based on CORINA. They showed that CORINA successfully processed molecules with about 700 and more non-hydrogen atoms without problems.[290]

However, research interests have shifted to other 3D structure related topics. For example, in 2001, Feuston et al. reported a knowledge-based approach for generating conformations of molecules,[291] that provides a good sampling of a molecule's conformational space by restricting the generated conformations to those consistent with the reference database. In 2004, Bywater et al. described the *de novo* generation of molecular structures using optimization to select graphs on a given lattice.[292] In 2005, Mekenyan et al. reported a new system for automated 2D-3D migration of chemicals in large databases with conformer multiplication.[293] The module for conformer multiplication within the 2D-3D migration system is based on a new formulation of the genetic algorithm for computing populations of possible conformers.

SUMMARY AND FUTURE PROSPECTS: COMPUTER TECHNOLOGY

The punched cards and vacuum tubes based "first generation" computers invented in the 1940s represent one of the greatest inventions in the history, leading to the revolution in the human society. In the 1950s, the transistors and printed circuits based "second generation" computers were built and became available for general use by a handful of scientists. The integrated circuits based "third generation" computers built in the 1960s were more powerful; the minicomputers became available at major research laboratories, leading to the explosion in the use of computers. In the 1970s, the development of Large Scale Integration (LSI) circuits led to the wide availability of "fourth generation" computers, such as VAX 11/780. In the 1980s, the invention of very large scale integration techniques led to dramatically reducing the size of computers; the PC revolution in this decade placed computers directly in the hands of millions of people. The 1990s experienced the Internet revolution. Also, the speed of PCs was improved rapidly, as predicted by Moor's law.[27]

At the turn of the new century, computers and computer technology have become an integral part of our lives. The PCs in use today have more processing power than the mainframe computers used in 1969 to put men on the moon. The power of the traditional silicon-based computer will continue to increase by shrinking device sizes. However, there is a speed limit to silicon-based computers. The current method to make faster processors is to reduce the size of chips. It is estimated that in 25 years, the size of the chip will reach its final limit—the size of atoms.

It is interesting to note that about 5 years ago, Intel predicted that its Pentium 4 microarchitecture design should enable the chipmaker to make microprocessors that can eventually achieve 10 GHz in chip clock speed. However, Intel has failed to break even the 4-GHz barrier using the Pentium 4 design because of the heat problem.[294] Now, the new mantra in the chip industry is multicore. In the coming years, the multicore, especially the dual-core processor-based PC will become popular. Simply speaking, multicore processor architecture entails silicon design engineers placing two or more processor-based "execution cores", or computational engines, within a single processor. The idea behind this implementation of the chip's internal architecture is that by divvying up the computational work performed by the single microprocessor core in traditional microprocessors and spreading it over multiple execution cores, a multicore processor can perform more work within a given clock cycle. However, to enable this improvement, the software running on the multicore platform must be written such that it can spread its workload across multiple execution cores. This functionality is called thread-level parallelism or "threading". It is expected that this will lead to the biggest sea change in mainstream software development since the object-oriented programming revolution in the 1990s: the concurrency revolution.[295] In computer science, concurrency occurs when two or more execution flows are able to run simultaneously. Concurrent programming encompasses the programming languages and algorithms which are used to implement concurrent systems. Unlike parallel systems which generally have a predefined and well-structured communications pattern, concurrent programming is usually considered to be more general than parallel programming because it can involve arbitrary and dynamic patterns of communication and interaction. The base goals of concurrent programming include correctness, performance, and robustness.[296]

However, it should be pointed out that multicore architectures and hyperthreading approaches are only the short term solutions. To completely overcome the above limitation of silicon chips, computers based on completely different material and/or mechanism will be invented in this century. For example, quantum computers have the ability to solve problems digital computers cannot solve, by using superposition (entangled) states. Highly parallel quantum algorithms can decrease the computational time for some problems by many orders of magnitude.[297] Since the middle of the 1990s, the DNA computer has been tried for solving hard computational problems: applying molecular computation to the data encryption standard, running dynamic programming algorithms on a DNA computer, and active transport in biological computing.[298] One of the most promising new types of computer is the molecular computer. Some pioneering work in this area has been done in the last century. In his last State of the Union address on January 27, 2000, President Clinton predicted the existence of “molecular computers the size of a tear drop with the power of today’s fastest supercomputers”.[299]

Most recently, the transition to a post-silicon era is forecast in a report called the “International Technology Roadmap for Semiconductors (2005 Edition)”.[300] The report was produced cooperatively by semiconductor industry associations from Europe, Japan, Korea, Taiwan, and the United States. What has changed in the industry’s road map is the growing confidence in new technologies that make electronic switches from single molecules or even single electrons. The transition to new nanotechnology techniques could occur around 2015.[301] In addition, Intel predicted that “fifteen years hence, the capabilities that users expect of PCs are certain to change as dramatically as they have in the past 15 years. The evolution is likely to include magnitudes-better recognition applications and search functions that enable seamless mining of information and support knowledgeable, data-based decision-making”.[302]

It is interesting to recall that in 1982, Japan’s Ministry of International Trade and Industry launched the Fifth Generation Computer Systems project (FGCS) for creating an “epoch-making computer” with supercomputer-like performance and usable artificial intelligence capabilities,[303] surprising the world. However, unsurprisingly, this ambitious project failed a few years later because of the lack of some key technology foundations, such as the suitable programming language supporting concurrency and the CPU performance. It can be expected that at a certain point in this century, people will resume the interest in the “fifth generation” computer. The dream of building the intelligent computer that is able to listen, talk, think, reason, and learn will become true, which will lead to a new computer revolution.

Finally, it can also be expected that there will be a human-machine interface revolution in this century. For instance, the future user interface will allow databases and knowledge bases to be used as a “natural” extension of a human’s memory. This will make human beings much smarter and more powerful than we are today. In that time, we will no longer fear that human beings will be defeated and controlled by humanoid computers in the future, as described in some science fictions.

SUMMARY AND FUTURE PROSPECTS: DATABASE SYSTEMS

Database systems are among the most mature and most widely used chemoinformatics techniques. Great varieties of chemical databases have been built that are accessible worldwide, such as literature databases, patent databases, molecular databases, reaction databases, spectral databases, and chemical safety databases. Although the value of the results of the failed experiments has been recognized for a long time, only very few databases that collect the results of solely failed experiments have appeared fairly recently.

Most of the current databases are factual databases that contain the experimental reports on the inherent properties of substances. For example, the CAS Registry currently contains more than 25 million compounds, CASREACT contains over 10 million single- and multistep reactions,[304] and the Beilstein Database contains more than 260 million experimental data (facts). In the future, more databases for storing calculation results from expensive, accurate, *ab initio* calculations and knowledge deduced from factual databases will appear. These new databases will become important resources, complementing the factual databases. All scientific journal and books will be offered mainly in electronic format. The chemical literature databases will be fully integrated with molecular databases, reaction databases, protein databases, and so on. Public information databases will become more valuable. The size and importance of databases will continue to grow until properties can be accurately and quickly predicted from the structure alone. Information retrieval will be vastly improved through the development of faster search algorithms,[305] data mining, and other analysis techniques. Database, knowledge base, and intelligent software will hold the key to the future in the chemical database areas.

SUMMARY AND FUTURE PROSPECTS: EXPERT SYSTEMS

Chemical expert systems have experienced a series of dramatic gains and setbacks. In the early days of Stanford’s DENDRAL and Harvard’s OCSS there was great enthusiasm about the possibilities in this area. Some chemists even feared that one day their jobs might be replaced by those intelligent computer systems. However, it turned out that expert systems with real-problem solving capability were more difficult to build, to maintain, and to use than one had expected. Although a variety of expert systems have been developed for different purposes—structure elucidation, synthesis design, catalyst design, drug design, fault analysis and control, and so on—many of them are still too “young” to be useful.

One major difficulty in developing chemical expert systems is the lack of efficient and accurate prediction methods for sorting and selecting the final candidates. In the structure elucidation system, we need to predict the spectra of all the structure candidates generated; in the synthesis design system, we need to predict reactivity, even toxicity, of the precursors and starting materials; and in the ligand design system, we need to predict binding affinity and synthetic accessibility of ligand candidates. The other challenging problem is how to build a knowledge base automatically from known data. In particular, we need to know how to build reaction knowledge bases from reaction databases for CASD systems. For the CASE systems, more efficient data mining tools are also of great importance for automatically deriving structural features of a given unknown compound from its spectral data in the corresponding spectral databases.

Although in the early 1980s some pioneering work had been done in integrating IR spectral database systems with the computers that control the instruments, the new generation of smart, fully automatic and remotely controllable spectroscopic instruments armed with the technologies of automatic measurement, data collection, and analysis based on the fully integrated spectral databases and CASE systems will appear in the future.

Chemical expert systems, especially the CASD systems, are probably the least used chemoinformatics systems. In recent years, however, they have gradually gained acceptance, especially the CASE systems. On the other hand, it must be pointed out that many methods used in other chemoinformatics systems, such as database retrieval systems, molecular modeling, and so on, were originally developed in the CASE and CASD areas. For example, the most widely used chemical drawing program, ChemDraw,[306] has a deep root of the structure input program of the LHASA system.[149]

It is interesting to note that the traditional CASD system has recently found a new application in drug discovery by estimating the synthetic accessibility of ligand candidates. In fact, this area has become so important that a symposium at the ACS National Meeting in March 2006 has been organized by Gasteiger to show the present state of affairs in the *de novo* design of lead structures and in methods for estimating the synthetic accessibility of organic structures.[307]

The CASD systems which aim at searching for the "ideal synthesis" using the shortest and most atom economical route from starting materials to the target will have real applications in green chemistry in the future.

It is interesting to recall that in the spring of 1997 in New York, IBM Deep Blue,[308] a 32-node IBM RS/6000 supercomputer, stunned the world by defeating grandmaster Garry Kasparov, one of the greatest chess players in human history. It was the first time a computer had beaten a reigning world chess champion, and this event ignited a public debate on how close electronic computers could come to approximating human intelligence. The above example indicates that it is indeed possible to develop expert systems for solving the problems in a narrow chemical domain. However, this will take much longer time to achieve because of several reasons. We can take synthesis design as an example. First, solving synthesis design problems is usually much harder than playing chess, thanks to the much larger search space. Second, chemists usually do not have access to the most advanced computer systems such as Deep Blue for their CASD work. Third, although there have been several reaction databases containing over several million reactions (e.g., both the CASREACT and Beilstein databases contains nearly 10 million reactions), the current CASD systems can utilize only a tiny fraction of them. For example, by 1994, the number of reactions known to the LHASA system was only 2100.[309] As an interesting comparison, the Deep Blue was armed with a database of more than 700 000 grandmaster games, and, more importantly, this chess computer had the ability to extract useful knowledge from that game database.[310] With the rapid progress in computer technology, the second condition will no longer be a big issue in the near future. Although we cannot change the complexity of the synthesis design problem, it is possible to reduce the search space by employing better heuristics and more efficient searching algorithms. And thus, the remaining challenge is how to teach the CASD systems to learn and use, effectively, all known reactions in the ever growing reaction databases. It can be expected that when this occurs the CASD system will be better than the average synthetic organic chemist.

SUMMARY AND FUTURE PROSPECTS: 3D STRUCTURE BUILDERS

The 3D structure builders are designed to perform the automatic conversion of a chemical structure from a 2D connection table into a 3D molecular model. There are two types of 3D structure builders: rule-based builders and database-based builders. A variety of 3D structure generation programs have been developed. The most well-known ones among them are CONCORD and CORINA. Both programs perform a robust and reasonably good 3D conversion. CONCORD is somewhat faster than CORINA, while CORINA has a better conversion rate than CONCORD. The 3D builders have become a standard chemoinformatics technique routinely used in many applications, such as drug discovery and determination of 3D structures in solution by NMR techniques. The 3D structure builders derive their knowledge mainly from experimental data, computed geometries, or from rules about the construction of 3D models. The future development of 3D structure builders will continue to focus on the improvement of computing speed and molecular model accuracy. For building 3D molecular models whose experimental data are not available, molecular mechanics or even *ab initio* calculation techniques will play a more important role.

SUMMARY: GASTEIGER'S CONTRIBUTIONS TO CHEMOINFORMATICS

Professor Johann Gasteiger is a pioneer in the field of chemoinformatics. He has played an important role in the formation of the discipline through his significant contributions to several of the important research areas in chemoinformatics in the past 30 years: computer-assisted synthesis design (EROS, WODCA), reaction simulation and prediction (EROS), reaction classification (CORA, HORACE[311,312]), 3D structure building (CORINA), simulation of spectra (TeleSpec project[313]), reaction database development (ChemInform), applications of neural networks, and genetic algorithms in chemistry.

Professor Gasteiger has also made important contributions to the areas of molecular modeling and drug design. For example, besides the well-known Gasteiger Charges, his group has also developed methods for the treatment of conformational flexibility of molecules (ROTATE and GAMMA) and approaches for the discovery of lead structure, the optimization of drugs, the definition of similarity and diversity of combinatorial libraries,[314] the establishment of structure−activity relationships,[315] virtual screening,[316] data mining,[317] chemotaxonomy,[318] prediction of the metabolism of drugs,[319] and modeling of biochemical pathways.[320] These methods are based on structure coding techniques (SURFACE, AUTOCORR, ARC) developed in his group.[257]

Professor Gasteiger is a coeditor of the prestigious *Encyclopedia of Computational Chemistry*,[321] and the editor of the authoritative *Handbook of Chemoinformatics*.[11] In addition, he has edited four volumes in the *Software Development in Chemistry* series: #1 (1987), #2 (1988), #4 (1990), and #10 (1996). He is also the author of several influential books, such as *Neural Networks in Chemistry and Drug Design*[322] and *Chemoinformatics − A Textbook*.[323] Professor Gasteiger is also a cofounder of the well-known Computer-Chemistry-Center at Erlangen, Germany.[5] Last but not least, he has been a mentor for many chemoinformatics specialists in the chemical and pharmaceutical industry and in academe.

Professor Gasteiger is one of the leading chemoinformatics scientists. His contributions to chemoinformatics as well as other related fields have been widely recognized. He is the recipient of the Gmelin-Beilstein Medal of GDCh (1997), the Herman Skolnik Award of the ACS Division of Chemical Information (1997), the Mike Lynch Award of the Chemical Structure Association Trust (2005), and most recently, the ACS Award for Computers in Chemical & Pharmaceutical Research (2006).[313]

SUMMARY AND FUTURE PROSPECTS: CHEMOINFORMATICS

After its development over the last half-century, a variety of chemoinformatics methods has been established and become “must have tools”, such as chemical structure drawing programs and database searching systems. Chemoinformatics has emerged as a new discipline, as indicated by Gasteiger’s *Handbook of Chemoinformatics*. Besides the traditional mainstream areas of chemoinformatics, such as database systems, computer-assisted structure elucidation systems, computer-assisted synthesis design systems, and quantitative structure−activity relationship (QSAR), several new research areas of chemoinformatics have appeared recently, such as *in silico* library design, virtual screening, docking, prediction of ADME (absorption, distribution, metabolism, and excretion) and toxicity, and so on. The scope of this rapidly developing field will certainly continue to expand.

Because of the specific focus of this contribution and also the limitations of space, the new research frontiers described above have not been covered here. However, it should be noticed that many of the new chemoinformatics systems have been built on the base of the key technologies (such as chemical structure representation, substructure and maximum common substructure searching,[324] etc.) invented during the development of “traditional” chemoinformatics systems. For example, the traditional structure database systems have been extended to support combinatorial chemistry;[179] conventional computer-assisted synthesis design systems have been modified to estimate synthetic accessibility of ligand candidates; and docking, a computational method that addresses the problem of the formation of noncovalent ligand−receptor complexes (in other words, the fitting and binding of small molecules (ligands) at the active sites of biomacromolecules (e.g., enzymes)), may be generally regarded as an integration of the structure database searching and QSAR. However, this does not mean that these problems have already been solved or can be easily solved using only the existing technologies. In the near future, docking, virtual screening, combinatorial library design, *de novo* design, and so on will continue to be active research areas in chemoinformatics.

It is interesting to notice that at the end of the 20th century, almost all the major foundations and theories of chemistry had been well understood and established. Chemistry has already evolved from largely a study of the elements to a study of molecules to currently a study of molecular interactions, especially those involving biomacromolecules. This offers an excellent opportunity for chemoinformatics to grow in this new direction.

On the other hand, even the traditional research areas of chemoinformatics continue to attract great attention. For example, cyber-enabled chemistry has recently been identified as a big thing to come, a core area in chemistry by the U.S. National Science Foundation.[325] The main focus is on the development of integrated databases, data mining tools, molecular visualization and computational capabilities, and remote and networked use of instrumentation. This has the potential to more efficiently tackle difficult scientific problems previously thought to be intractable. Since the above areas are heavily overlapping with some areas in chemoinformatics, the cyber-enabled chemistry may be regarded as a branch of chemoinformatics.

It is worth mentioning that there is a new trend of integration of chemoinformatics with bioinformatics. This is because many sectors of the chemical and pharmaceutical industries are interdisciplinary by nature, and major progress and developments in those industries are occurring in both bioinformatics and chemoinformatics side by side. For example, Genetics Computer Group (GCG), a subsidiary of Pharmacopeia, created the Wisconsin Package, the bioinformatics industry standard for sequence analysis. In 2001, GCG’s vice president of R&D, Joseph King, pointed out that one of the big trends in bioinformatics was integration across the research spectrum, from bioinformatics to molecular modeling to small-molecule cheminformatics.[326] There were even conferences specifically focusing on this topic.[327]

It should also be pointed out that there is another type of integration: integration of computational chemistry approaches with chemoinformatics methods. For example, molecular mechanics and even quantum chemistry calculations were employed in the reaction prediction expert system ROBIA.[289] In another example, Clark recently demonstrated that complete databases of tens of thousands of compounds can be treated with economical quantum mechanical techniques, showing the advantages of detailed quantum mechanical descriptions of molecules for QSPR[328] and QSAR.[329] He even referred to the above work as “Quantum Chemoinformatics”. It should be pointed out that the integration of computation chemistry with chemoinformatics is a bidirectional process. Chemoinformatics methodologies will also be used to develop new computational chemistry systems. For instance, structure generation algorithms and expert system technologies will be employed to build more intelligent computational chemistry programs for studying transition states of chemical reactions in a more systematic and automatic way. It can be expected that the above trend will continue to grow. In the future, the division between chemoinformatics and computational chemistry may disappear. The combination and integration of computational methods and informatics technologies (*compuinformatics* for short) will lead to the development of much more powerful and much more intelligent methodologies and compuinformatic systems. And thus, the integrated new field of computational chemistry and chemoinformatics may be called *compuinformatic chemistry*, a field of interdisciplinary science in which mathematics, computer science, informatics, and chemistry merge into a single discipline. The major goals of compuinformatic chemistry are to develop chemistry-oriented mathematical models, computational and information technologies, and computer programs and apply these programs to studying and solving chemical problems. It can be expected that compuinformatic chemistry will become one of the most active research frontiers in chemistry in the future. It should also be noticed that, in the recent years, the classic two approaches to scientific research, theoretical/analytical and experimental/observational approaches, have been extended to *in silico* modeling and simulation to explore new possibilities and to achieve new precision. The emerging *compuinformatic science* (including compuinformatic chemistry, compuinformatic biology,[330] etc.), whose major goals are to develop mathematical models, computational and information technologies, and computer programs and apply these programs to studying and solving scientific and engineering problems, will become a key technology field of the 21st century. It will bring a new generation of much more powerful yet intelligent modeling and simulation methods and other revolutionary technologies, allowing scientists to study real-life complex phenomena and processes that are experimentally difficult, if not impossible, to characterize and to solve.

For example, molecular design (such as drug design, catalyst design, and material design) is aimed at designing a molecular structure in such a way that the resulting molecule or a collection of such molecules (such as in the material) will possess the desired physical, chemical, and/or biological properties. Up until now, molecular design is still done mainly using the trial-and-error method. Take drug design as an example. Drug discovery is still a long, extremely expensive process with a very low rate of new therapeutic discovery. Countless molecules have to be “designed”, synthesized, and tested before a new drug molecule is discovered. However, it should be noticed that good progress continues to be made in this important area of molecular design. For instance, as of this writing, Wang et al. described a novel molecular-design method based on a linear combination of atomic potentials (LCAP), a general framework that creates a continuous property landscape from an otherwise unlinked set of discrete molecular-property values.[331] Theoretically, this LCAP can represent any molecules, and for some properties, it can troll through a billion trillion possible structures.[332] It can be expected that in the future, compuinformatic chemistry will play a key role in molecular design, making the design of the desired molecule as simple and efficient as designing an automobile or a home. The dream of such genuine molecular design will become true in this century.

As mentioned previously, the invention of electronic computers in the middle of the last century has revolutionized human society. It can be expected that in the next half-century, the new generation of non-silicon computers, such as molecular computers, will replace the existing silicon-based computers. Those extremely powerful yet tiny computers, supported with the tiny yet huge capacity storage device and the revolutionary computer-human interface, will bring a new revolution to the entire human society. For example, the paper notebooks used for centuries by chemists will eventually be replaced with the electronic notebooks armed with the truly advanced technologies of hand-writing recognition and voice recognition. Chemists will become more and more computer dependent, Internet dependent, and chemoinformatics dependent.

Chemoinformatics, through its development in the *past* half a century, has reached in the *present* wide acceptance, *and* will have a bright *future*!

ACKNOWLEDGMENT

The author gratefully acknowledges Dr. James G. Nourse for many constructive discussions, suggestions, and support during the development of this paper. Thanks also go to Professor Johann Gasteiger, Dr. Phil McHale, and all referees for their valuable comments on the manuscript.

REFERENCES AND NOTES

(1) Chemometrics World: http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,1181,2-4-741-0-741-directories-0-0,00.html.
(2) Chemometrics and Intelligent Laboratory Systems: http://www.elsevier.com/wps/find/journaldescription.cws_home/502682/description#description.
(3) *Journal of Chemometrics*: http://www3.interscience.wiley.com/cgi-bin/jhome/4425.
(4) *Journal of Computer Chemistry*: http://www.sccj.net/publications/JCCJ/.
(5) Computer-Chemistry-Center, Erlangen, Germany: http://www.chemie.uni-erlangen.de/ccc/.
(6) Marsili, M. *Computer Chemistry*; CRC Press: Florida, 1989.
(7) *Computer Chemistry (Topics in Current Chemistry)*; Ugi, I. Ed.; Springer-Verlag: Berlin, 1993.
(8) Hogeweg, P. Simulation of Cellular Forms. In *Frontiers in System Modelling*; Zeigler, B. P., Ed.; Simulation Councils, Inc.: 1978; pp 90−95.
(9) The Human Genome Project (HGP) was a 13-year project (1990−2003) coordinated by the U.S. Department of Energy and the National Institutes of Health. The goals of the Project were to identify all the approximately 20 000−25 000 genes in human DNA, determine the sequences of the 3 billion chemical base pairs that make up human DNA, store this information in databases, improve tools for data analysis, transfer related technologies to the private sector, and address the ethical, legal, and social issues (ELSI) that may arise from the project. (b) Human Genome Project Information: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.
(10) Brown, F. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annu. Rep. Med. Chem.* **1998**, *33*, 375−384.
(11) *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003.
(12) Gasteiger, J. The Scope of Chemoinformatics. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; pp 3−5.
(13) *Journal of Chemical Documentation*: http://pubs3.acs.org/acs/journals/toc.page?incoden=jci001.
(14) *Journal of Chemical Information and Modeling*: http://pubs.acs.org/journals/jcisd8/index.html.
(15) Willett, P. A History of Chemoinformatics. In *Handbook of Chemoinformatics*; Gasteiger, J. Ed.; Wiley-VCH: Weinheim, Germany, 2003; pp 6−20.
(16) A Brief History of Computing: http://ox.compsoc.net/∼swhite/history.html.
(17) King, G. W.; Cross, P. C.; Thomas, G. B. The Asymmetric Rotor. III. Punched-Card Methods of Constructing Band Spectra. *J. Chem. Phys.* **1946**, *14*, 35−42.
(18) IBM Archives: Valuable resources on IBM’s history: http://www-03.ibm.com/ibm/history/.
(19) *Mathematical Challenges from Theoretical/Computational Chemistry*; The National Academies Press: 1995; Chapter 2.
(20) Zemany, P. D. Punched Card Catalog of Mass Spectra Useful in Qualitative Analysis. *Anal. Chem.* **1950**, *22*, 920−922.

(21) Kuentzel, L. E. New Codes for Hollerith-Type Punched Cards. *Anal. Chem.* **1951**, *23*, 1413−1418.

(22) Fisanick, W.; Amaral, N. J.; Metanomski, W. V.; Shively, E. R.; Soukup, K. M.; Stobaugh, R. E. Chemical Abstract Service Information System. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 277−315.

(23) Ray, L. C.; Kirsch, R. A. Finding Chemical Records by Digital Computers. *Science* **1957**, *126*, 814−819.

(24) Opler, A.; Baird, N. Display of Chemical Structural Formulas as Digital Computer Output. *Am. Doc.* **1959**, *10*, 59−63.

(25) DEC PDP 8 (Programmed Data Processor) 1964: http://www.computermuseum.li/Testpage/DEC−PDP-8.htm.

(26) Hewlett-Packard: http://www.hp.com/

(27) Moore, G. E. Cramming more components onto integrated circuits. *Electronics.* April 19, 1965. (b) Wikipedia: Moore's law: http://en.wikipedia.org/wiki/Moores_Law.

(28) Sparks, R. A. *Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using an IBM 1401 Computer*; ASTM: Philadelphia, PA, 1964.

(29) Heller, S. R. Mass Spectrometry Databases and Search Systems. In *Computer-Supported Spectroscopic Databases*; Zupan, J., Ed.; Ellis Horwood Limited: New York, 1986; Chapter 6, pp 118−132.

(30) Allen, F. H.; Hoy, V. J. Cambridge Structural Database. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 155−167.

(31) Dubois, J. E.; Laurent, D.; Viellard, H. Système de Documentation et d'Automatisation des Recherches de Corrélations (DARC), Principes généraux. *C. R. Seances Acad. Sci., Ser. C* **1966**, *263*, 764−767. (b) Dubois, J. E.; Sobel, Y. DARC System for Documentation and Artificial Intelligence in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 326−333.

(32) Dubois, J. E. Principles of the DARC Topological System. Applications Pointing to Structural Influence on Oxidation of Hydrocarbons. *Entropie* **1969**, *25*, 5−13.

(33) Gluck, D. J. A Chemical Structure Storage and Search System Developed at Du Pont. *J. Chem. Doc*. **1965**, *5*, 43−51.

(34) Morgan, H. L. The Generation of Unique Machine Description for Chemical Structures − A Technique Developed at Chemical Abstracts Services. *J. Chem. Doc.* **1965**, *5*, 107−113.

(35) Vleduts, G. E. Concerning One System of Classification and Codification of Organic Reactions. *Inf. Stor. Retriev.* **1963**, *1*, 117−146.

(36) Armitage, J. E.; Lynch, M. F. Automatic Detection of Structural Similarities among Chemical Compounds. *J. Chem. Soc. (C)* **1967**, 521−528.

(37) Sussenguth, E. H., Jr. A Graph-Theoretic Algorithm for Matching Chemical Structures, *J. Chem. Doc*. **1965**, *5*, 36−43.

(38) Warr, W. A. In *Chemical Structure Information Systems. Interfaces, Communication, and Standards*; ACS Symposium Series 400. Warr, W. A., Ed.; American Chemical Society: Washington DC, 1989; Chapter 1, pp 1−9.

(39) Dyson, G. M.; Lynch, M. F.; Morgan, H. L. A Modified IUPAC−Dyson Notation System for Chemical Structures. *Inf. Storage Retriev.* **1968**, *4*, 27−83. (b) Dammers, H. F.; Polton, D. J. Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures. *J. Chem. Doc*. **1968**, *8*, 150−160.

(40) Smith, E. G. *The Wiswesser Line-Formula Chemical Notation*; McGraw-Hill: New York, 1968.

(41) Wiswesser, W. J. Historic Development of Chemical Notations. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258−263.

(42) Thomson, L. H.; Hyde, E.; Matthews, F. W. Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation. *J. Chem. Doc.* **1967**, *7*, 204−209.

(43) Ledley, R. S. Automatic Coding of Chemicals Directly from Structure Pictures. In *Automation and Scientific Communication*; Short Papers of the 26th Annual Meeting of the American Documentation Institute, Chicago, 6−11 October 1963. American Documentation Institute, Washington, DC, 1963, Pt. 2, pp 201−202.

(44) Lederberg, J. Topological Mapping of Organic Molecules. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 134−139.

(45) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley & Sons: New: York, 1986.

(46) Munk, M. E.; Sodano, C. S.; McLean, R. L.; Haskell, T. H. Actinobolin. I. Structure of Actinobolamine. *J. Am. Chem. Soc.* **1967**, *89*, 4158−4165. (b) Nelson, D. B.; Munk, M. E.; Gash, K. B.; Herald, D. L., Jr. Alanylactinobicyclone. Application of Computer Techniques to Structure Elucidation. *J. Org. Chem.* **1969**, *34*, 3800−3805.

(47) Sasaki, S.; Abe, H.; Ouki, T.; Sakamoto, M.; Ochiai, S. Automated Structure Elucidation of Several Kinds of Aliphatic and Alicyclic Compounds. *Anal. Chem.* **1968**, *40*, 2220−2223.

(48) Grant, D. M.; Paul, E. G. Carbon-13 Magnetic Resonance. II. Chemical Shift Data for the Alkanes. *J. Am. Chem. Soc.* **1964**, *86*, 2984−2990.

(49) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178−192.

(50) Hendrickson, J. B. Molecular Geometry. I. Machine Computation of the Common rings. *J. Am. Chem. Soc*. **1961**, *83*, 4537−4547.

(51) Boyd, D. B. Drug Design. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 795−804.

(52) Merrifield, R. B. Solid-Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *J. Am. Chem. Soc*. **1963**, *85*, 2149−2154.

(53) Wikipedia: Digital Equipment Corporation: http://en.wikipedia.org/wiki/Digital_Equipment_Corporation.

(54) Wikipedia: C Programming Language: http://en.wikipedia.org/wiki/C_programming_language.

(55) Wikipedia: Microsoft: http://en.wikipedia.org/wiki/Microsoft.

(56) Apple Computer, Inc.: http://www.apple.com.

(57) Inventors Apple Computer History: http://inventors.about.com/library/inventors/blapplecomputer.htm.

(58) Wikipedia: History of Apple Computer: http://en.wikipedia.org/wiki/History_of_Apple_Computer.

(59) Oracle Corporation: http://www.oracle.com.

(60) Wikipedia: Oracle Corporation: http://en.wikipedia.org/wiki/Oracle_Corporation.

(61) Crowe, J. E.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. I: Non-Cyclic Fragments. *J. Chem. Soc. (C)* **1970**, 990−996.

(62) Graf, W.; Kaindle, H. K.; Kniess, H.; Schmidt, B.; Warszawski, R. Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 51−55.

(63) *Nature New Biol.* **1971**, *233*, 223. (b) Sussman, J. L. Protein Data Bank (PDB): A Database of 3D Structural Information of Biological Macromolecules. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 2160−2168.

(64) Gund, P.; Wipke, W. T.; Langridge, R. Computer Searching of a Molecular Structure File for Pharmacophoric Patterns. *Comput. Chem. Res. Educ., Proc. Int. Conf.*; Elsevier: Amsterdam, 1974; Vol. 3, pp 5/33−5/38.

(65) Johnson, L. F.; Jankowski, W. C. *Carbon-13 NMR Spectra*; John Wiley: New York, 1972.

(66) Jezl, B. A.; Dalrymple, D. L. Computer Program for the Retrieval and Assignment of Chemical Environments and Shifts to Facilitate Interpretation of Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1975**, *47*, 203−207.

(67) Bremser, W.; Klier, M.; Meyer, E. Mutual Assignment of Subspectra and Substructures Structure Elucidation by ^{13}C-NMR Spectroscopy. *Org. Magn. Reson.* **1975**, *7*, 97−105.

(68) Bremser, W. HOSE − A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(69) Erni, E.; Clerc, J. T. Strukturaufklärung Organischer Verbindungen durch Computerunterstützen Vergleich Spektraler Daten. *Helv. Chim. Acta* **1972**, *55*, 489−500.

(70) Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. Computer-Aided Interpretation of Mass Spectra. III. Self-Training Interpretive and Retrieval System. *J. Am. Chem. Soc*. **1973**, *95*, 4185−4194.

(71) Dubois, J. E.; Bonnet, J. C. The DARC Pluridata System: The ^{13}C NMR Data Bank. *Anal. Chim. Acta* **1979**, *112*, 245−252.

(72) Speck, D. D.; Venkataraghavan, R.; McLafferty, F. W. A Quality Index for Reference Mass Spectra. *Org. Mass Spectrom.* **1978**, *13*, 209−213.

(73) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. Quality Control of Chemical Data Bases. *J. Chem. Inf. Comput. Sci*. **1976**, *16*, 232−233.

(74) Adamson, G. W.; Bush, J. A. A Method for the Automatic Classification of Chemical Structures. *Inf. Storage Retriev.* **1973**, *9*, 561−568.

(75) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(76) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci*. **1986**, *26*, 36−41.

(77) Ullmann, J. R. An Algorithm for Subgraph Isomorphism, *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.

(78) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci*. **1978**, *18*, 154−159.

(79) Evans, L. A.; Lynch, M. F.; Willett, P. Structural Search Codes for On-line Compound Registration. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 146−149.

(80) *Molecular Connection*, The MDL Newsletter for Communicating with Customers, 20th Anniversary Issue, January 1998; Vol. 17, No. 1, p 2.

(81) Elsevier MDL: Corporate History: http://mdl.com/company/about/history.jsp.

(82) Elsevier MDL: http://www.mdl.com.

(83) Sasaki, S.; Kudo, Y.; Ochiai, S.; Abe, H. Automated Chemical Structure Analysis of Organic Compounds: An Attempt to Structure Determination by the Use of NMR. *Mikrochim. Acta* **1971**, 726−742.

(84) Masinter, L. M.; Sridharan, N. S.; Lederberg, J.; Smith, D. H. Applications of Artificial Intelligence for Chemical Inference. XII. Exhaustive Generation of Cyclic and Acyclic Isomers. *J. Am. Chem. Soc.* **1974**, *96*, 7702−7714.

(85) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. XVII. Approach to Computer-Assisted Elucidation of Molecular Structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755−5762.

(86) Serov, V. V.; Elyashberg, M. E.; Gribov, L. A. Mathematical Synthesis and Analysis of Molecular Structures. *J. Mol. Struct.* **1976**, *31*, 381−397.

(87) Gribov, L. A.; Elyashberg, M. E.; Serov, V. V. Computer System for Structure Recognition of Polyatomic Molecules by IR, NMR, UV and MS methods. *Anal. Chim. Acta* **1977**, *95*, 75−96.

(88) Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. An Approach to Automated Partial Structure Expansion. *Anal. Chim. Acta* **1978**, *103*, 121−132.

(89) Shelley, C. A.; Woodruff, H. B.; Snelling, C. R.; Munk, M. E. Interactive Structure Elucidation. In *Computer-Assisted Structure Elucidation*; Smith, D. H., Ed.; American Chemical Society: Washington, DC, 1977; Vol. 92, Chapter 7.

(90) Nourse, J. G.; Carhart, R. E.; Smith, D. H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. 29. Exhaustive Generation of Stereoisomers for Structure Elucidation. *J. Am. Chem. Soc.* **1979**, *101*, 1216−1223.

(91) Stanford University: Historical Projects: http://smi-web.stanford.edu/projects/history.html#METADENDRAL.

(92) Hendrickson, J. B. A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis. *J. Am. Chem. Soc.* **1971**, *93*, 6847−6854.

(93) Hendrickson, J. B. Systematic Synthesis Design. IV. Numerical Codification of Construction Reactions. *J. Am. Chem. Soc.* **1975**, *97*, 5784−5800.

(94) Corey, E. J.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94*, 421−431.

(95) Wipke, W. T. Evolution of Molecular Graphics. In *Graphics for Chemical Structures. Integration with Text and Datta*; Warr, W. A., Ed.; American Chemical Society: Washington, DC, 1987; pp 1−8.

(96) Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. SECS − Simulation and Evaluation of Chemical Synthesis: Strategy and Planning. In *Computer-Assisted Organic Synthesis*; Wipke W. T., Ed.; American Chemical Society: Washington, DC, 1977; pp 97−127.

(97) Bersohn, M. Automatic Problem Solving Applied to Synthetic Chemistry. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 1897−1903.

(98) Gelenter, H.; Sridharan, N. S.; Hart, H. J.; Yen, S. C.; Fowler, F. W.; Shue, H. J. The Discovery of Organic Synthetic Routes by Computer. *Topics Curr. Chem.* **1973**, *41*, 113−150.

(99) Blair, J.; Gasteiger, J.; Gillespie, C.; Gillespie, P. D.; Ugi, I. Representation of the Constitutional and Stereochemical Features of Chemical Systems in the Computer-Assisted Design of Syntheses. *Tetrahedron* **1974**, *30*, 1845−1859.

(100) Gasteiger, J.; Jochum, C. EROS − A Computer Program for Generating Sequences of Reactions. *Topics Curr. Chem.* **1978**, *74*, 93−126.

(101) Gasteiger, J.; Marsili, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrehedron Lett.* **1978**, 3181−3184.

(102) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity − An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.

(103) Gasteiger, J.; Saller, H. Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomeriekonzeptes. *Angew. Chem.* **1985**, *97*, 699−701. (b) Gasteiger, J.; Saller, H. Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687−689.

(104) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559−564.

(105) CCL.NET: Computational Chemistry List, ltd.: (From: Wolf-Dietrich Ihlenfeldt <Wolf-Dietrich.Ihlenfeldt $#at#$ ccc.chemie.un; Date: Wed, 16 Jan 2002 14:34:02 +0100; Subject: Re: CCL: Gasteiger charges questions) http://www.ccl.net/chemistry/resources/messages/2002/01/16.005-dir/index.html.

(106) Wipke, W. T.; Verbolis, J.; Dyott, T. Three-Dimensional Interactive Model Building. Presented at the 162nd National Meeting of the American Chemical Society, Los Angeles, August, 1972.

(107) IBM Archives: Valuable resources on IBM's history: http://www-03.ibm.com/ibm/history/.

(108) Wikipedia: Apple Lisa: http://en.wikipedia.org/wiki/Apple_Lisa.

(109) Wikipedia: Windows 1.0: http://en.wikipedia.org/wiki/Windows_1.0.

(110) Wikipedia: History of the Internet: http://en.wikipedia.org/wiki/Internet_history.

(111) Wikipedia: C++ programming language: http://en.wikipedia.org/wiki/C%2B%2B.

(112) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148−150.

(113) Shenton, K.; Norton, P.; Fearns, E. A. Generic Searching of Patent Information. In *Chemical* Structures − the International Language of Chemistry; Warr, W. A., Ed.; Springer: Berlin, 1988.

(114) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145−155.

(115) Bruck, P.; Nagy, M. Z.; Kozics, S. Substructure Search on Hierarchical Trees. In *Online Information 87; Proceedings of the 11th International Online Information Meeting*, London, Dec 8−10, 1987; Learned Information: Oxford, 1987; pp 41−43. (b) Nagy, M. Z.; Kozies, S.; Veszpremi, T.; Bruck, P. Substructure Search on Very Large Files Using Tree-Structured Databases. In *Chemical Structures: The International Language of Chemistry*; *Proceedings of an International Conference at the Leeuwenhorst Congress Center*, Noordwijkerhout, The Netherlands, May 3 I-June 4, 1987; Warr, W. A., Ed.; Springer: Heidelberg, 1988; pp 127−130.

(116) Jakes, S. E.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures. Selection of Interatomic Distance Screens. *J. Mol. Graphics* **1986**, *4*, 12−20.

(117) Jakes, S. E.; Watts, N.; Willett, P.; Barden, D.; Fisher, J. D. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Evaluation of Search Performance. *J. Mol. Graphics* **1987**, *5*, 41−48.

(118) Brint, A. T.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Geometric Searching Algorithms. *J. Mol. Graphics.* **1987**, *5*, 49−56.

(119) Brint, A. T.; Willett, P. Algorithms for the Identification of Three-Dimensional Maximal Common Substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.

(120) Downs, G. M.; Lynch, M. F.; Willett, P.; Manson, G. A.; Wilson, G. A. Transputer implementations of chemical substructure searching algorithms. *Tetrah. Comput. Method.* **1988**, *1*, 207−217.

(121) Wikipedia: INMOS Transputer: http://en.wikipedia.org/wiki/Transputer.

(122) Willett, P. The Evaluation of an Automatically Indexed, Machine-Readable Chemical Reactions File. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 93−96.

(123) Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The Implementation of AAM and Related Reaction Features in the Reaction Access System (REACCS). In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 303−313. (b) Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe G.; Cronin, K.; Nourse, J. G. Similarity Searching in the Organic Reaction Domain. *Tetrah. Comput. Method.* **1988**, *1*, 117−128.

(124) Chodosh, D.; Mendelson, W. L. SYNthesis LIBrary-Graphics at the Bench. *Drug Inf. J.* **1983**, *17*, 231−238.

(125) Johnson, A. P. Computer Aids to Synthesis Planning. *Chem. Br.* **1985**, *21*, 59−67.

(126) Gay, J. P. RMS-DARC Reaction Management System: A New Software Produced by Telesystemes DARC. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Aldershot: Gower, 1986; pp 87−91.

(127) Picchiottino, R.; Georgoulis, G.; Sicouri, G.; Panaye, A.; Dubois, J. E. DARC-SYNOPSYS. Designing Specific Reaction Data Banks: Application to KETO-REACT. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 241−249.

(128) Gasteiger, J.; Weiske, C. ChemInform − An Integrated Information System on Chemical Reactions. *Proceed. 13th Internat. Online Inform. Meet.*; Learned Information, Oxford, 1989; pp 147−154. (b) Parlow, A.; Weiske, C.; Gasteiger, J. ChemInform − An Integrated Information System on Chemical Reactions, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400−402.

(129) InfoChem GmbH: http://www.infochem.de/eng/index.htm.

(130) Heller, S. R. The Beilstein Online Database, an Introduction. In *The Beilstein Online Database: Implementation, Content, and Retrieval*;

Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1990; Chapter 1, pp 1−9.

(131) Luckenbach, R.; Jochum, C. Beilstein-Institut für Literatur der Organischen Chemie, Frankfurt am Main, Achema-Jahrbuch 88, Band 1: Forschung und Lehre des Chemie-Ingenierswesens, Seite 245−247. (b) Jochum, C. Building a Structure-Oriented Numerical Factual Database. In *Chemical Structure*; Warr, W. A., Ed.; Springer-Verlag, 1988; pp 187−193.

(132) Jochum, C. Computerizing Beilstein. In *The Beilstein Online Database: Implementation, Content, and Retrieval*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1990; Chapter 2, pp 10−23.

(133) Luckenbach, R. Past Perfect, Present Perfect, Future Perfect − Quality Assessment and Quality Control Mechanisms at Beilstein. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 923−929.

(134) Milne, G. W. A.; Heller, S. R. NIH/EPA Chemical Information System. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204−211.

(135) Gray, N. A. B.; Carhart, R. E.; Lavanchy, A.; Smith, D. H.; Varkony, T.; Buchanan, B. G.; White, W. C.; Creary, L. Computerized Mass Spectrum Prediction and Ranking. *Anal. Chem.* **1980**, *52*, 1095−1102.

(136) Shaps, R. H.; Sprouse, J. F. Fast Matching with IR Spectral Search and Display. *Ind. Res. Devel.* **1981**, *23*, 168−173.

(137) Lindley, M. R.; Gray, N. A. B.; Smith, D. H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. Part 40. A Computerized Approach to the Verification of Carbon-13 NMR Spectral Assignments. *J. Org. Chem.* **1982**, *47*, 1027−1035.

(138) Milne, G. W. A.; Budde, W. L.; Heller, S. R.; Martinsen, D. P.; Oldham, R. G. Quality Control and Evaluation of Mass Spectra. *Org. Mass Spectrom.* **1982**, *17*, 547−552.

(139) Domokos, L.; Henneberg, D.; Weimann, B. Optimization of Search Algorithm for a Mass Spectral Library. *Anal. Chim. Acta* **1983**, *150*, 37−44.

(140) McLafferty, F. W.; Stauffer, D. B. Retrieval and Interpretive Computer Programs for a Mass Spectrometry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 245−252.

(141) Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103−108.

(142) Neudert R.; Davies, A. N. Spectroscopic Databases. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; pp 700−721.

(143) Neudert, R. Spectroscopic Databases. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 2632−2646.

(144) Davies, A. N. Standard Exchange Formats for Spectral Data. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 2692−2699.

(145) TranSpec is a program developed by Chemical Concepts, Boschstrasse 12, D-69469 Weinheim, Germany.

(146) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(147) Wade, S. J., Willett, P.; Bawden, D. SIBRIS: The Sandwich Interactive Browsing and Ranking Information System. *J. Inform. Sci.* **1989**, *15*, 249−260.

(148) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(149) Rubenstein, S. Electronic Documents in Chemistry: from ChemDraw 1.0 to Present: http://images.cambridgesoft.com/powerpoint/ACS%-202004-08-24.ppt.

(150) Dubois, J. E.; Carabedian, M.; Ancian, B. Elucidation structurale, automatique par RMN du carbone 13: Méthode DARC-EPIOS. Recherche d'une relation discriminante structure-déplacement chimique. *C. R. Seances Acad. Sci., Ser. C* **1980**, *290*, 369−372.

(151) Woodruff, H. B.; Smith, G. M. Computer Program for the Analysis of Infrared Spectra. *Anal. Chem.* **1980**, *52*, 2321−2327.

(152) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. 37. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *46*, 1708−1718.

(153) Debska, B.; Duliban, J.; Guzowska-Swider, B.; Hippe, Z. Computer-Assisted Structure Analysis of Organic Compounds by an Artificial Intelligence System. *Anal. Chim. Acta* **1981**, *133*, 303−318.

(154) Gasteiger, J.; Marsili, M. Prediction of Proton Magnetic Resonance Shifts: The Dependence on Hydrogen Charges Obtained by Iterative Partial Equalization of Orbital Electronegativity. *Org. Magn. Reson.* **1981**, *15*, 353−360.

(155) Neudert, R.; Bremser, W.; Wagner, H. Multidimensional Computer Evaluation of Mass Spectra. *Org. Mass Spctrom.* **1987**, *22*, 321−329.

(156) Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. Stereochemical Substructure Codes for ^{13}C Spectral Analysis. *Org. Magn. Reson.* **1981**, *15*, 375−389. (b) Gray, N. A. B.; Crandell, C. W.; Nourse, J. G.; Smith, D. H.; Dageforde, M. L.; Djerassi, C. Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data. *J. Org. Chem.* **1981**, *46*, 703−715.

(157) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.

(158) Christie, B. D.; Munk, M. E. The Application of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Assisted Structure Elucidation. *Anal. Chim. Acta* **1987**, *200*, 347−361.

(159) Funatsu, K.; Susuta, Y.; Sasaki, S. Introduction of Two-Dimensional NMR Spectral Information to an Automated Structure Elucidation System, CHEMICS. Utilization of 2D-INADEQUATE Information. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6−11.

(160) Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. A Logic-Based Program for Synthesis Design. *J. Am. Chem. Soc.* **1985**, *107*, 5228−5238.

(161) Sieber, W. In *Artificial Intelligence: Towards Practical Application*; Bernold, T., Albers, G., Eds.; Elsevier: Amsterdam, 1986; pp 107−109.

(162) Salatin, T. D.; Jorgensen, W. L. Computer Assisted Mechanistic Evaluation of Organic Reactions. l. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2051.

(163) Hanessian, S. *Total Synthesis of Natural Products − The 'Chiron' Approach*; Pergamon Press: Oxford, U.K., 1983.

(164) Wipke, W. T.; Rogers, D. Artificial Intelligence in Organic Synthesis. SST: Starting Material Selection Strategies. An Application of Superstructure Search. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71−81.

(165) Funatsu, K.; Sasaki, S. Computer-Assisted Organic Synthesis Design and Reaction Prediction System, "AIPHOS". *Tetrah. Comput. Method.* **1988**, *1*, 27−37.

(166) Cohen, N. C.; Colin, P.; Lemoine, G. SCRIPT: Interactive Molecular Geometrical Treatments on the Basis of Computer-Drawn Chemical Formula. *Tetrahedron* **1981**, *37*, 1711−1721.

(167) Dolata, D. P.; Leach, A. R.; Prout, K. WIZARD: AI in Conformational Analysis. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73−85.

(168) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Automation News* **1987**, *2*, 5−7.

(169) Hiller, C.; Gasteiger, J. Ein Automatisierter Molekülbaukasten. In *Software-Entwicklung in der Chemie*; Gasteiger, J., Ed.; Springer: Berlin, 1987; Vol 1, pp 53−66.

(170) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581

(171) Wipke, W. T.; Hahn, M. A. Analogy and Intelligence in Model Building. In *Application of Artificial Intelligence in Chemistry*; Pierce, T., Hohne, B., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 136−146.

(172) IBM Archives > Exhibits > History of IBM > 1990s: http://www-03.ibm.com/ibm/history/history/decade_1990.html.

(173) W3C: A Little History of the World Wide Web: http://www.w3.org/History.html.

(174) Computer History: History of UNIX / Linux and other variants: http://www.computerhope.com/history/unix.htm.

(175) Microsoft Corporation: http://www.microsoft.com.

(176) Sun Microsystems, Inc.: The Source for Java Developers: http://java.sun.com.

(177) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.

(178) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motif in Proteins. *J. Mol. Biol.* **1990**, *212*, 151−166.

(179) Leland, B. A.; Christie, B. D.; Nourse, J. G.; Grier, D. L.; Carhart, R. E.; Maffett, T.; Welford, S. M.; Smith, D. H. Managing the Combinatorial Explosion, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 62−70.

(180) Brown, R. D.; Downs, G. M., Willett, P.; Cook, A. P. F. A Hyperstructure Model for Chemical Structure Handling: Generation and Atom-by-atom Searching of Hyperstructures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 522−531.

(181) Artymiuk, P. J.; Grindley, H. M.; Park, J. E.; Rice, D. W.; Willett, P. Three-Dimensional Structural Resemblance between Leucine

Aminopeptidase and Carboxpeptidase As Revealed by Graph-Theoretical Techniques. *FEBS Lett.* **1992**, *303*, 48−52.

(182) Artymiuk, P. J.; Grindley, H. M.; Kumar, K.; Rice, D. W.; Willett, P. Three-Dimensional Structural Resemblance between the Ribonuclease H and Connection Domains of HIV Reverse-Transcriptase Revealed Using Graph Theoretical Techniques. *FEBS Lett.* **1993**, *324*, 15−21.

(183) W3C: Extensible Markup Language (XML): http://www.w3.org/XML/

(184) Hicks, M. G.; Jochum, C. Substructure Search Systems. 1. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry, MVSSS, and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191−199.

(185) Christie, B. D.; Leland, B. A.; Nourse, J. G. Structure Searching in Chemical Databases by Direct Lookup Methods. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 545−547.

(186) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.

(187) Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm. *J. Mol. Biol*. **1993**, *229*, 707−721.

(188) Artymiuk, P. J.; Grindley, H. M.; MacKenzie, A. B.; Rice, D. W.; Ujah, E. C.; Willett, P. PROTEP: A Program for Graph-Theoretic Similarity Searching of the 3-D Structures in the Protein Data Bank. In *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbo, R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 123−140.

(189) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(190) Hendrickson, J. B.; Sander, T. COGNOS: A Beilstein-Type System for Organizing Reactions. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 251−260.

(191) WebReactions: http://www.webreactions.net.

(192) Warr, W. A. Combinatorial Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 407−417.

(193) Springer: http://www.springer.de/newmedia/chemist/infochem.

(194) Synthesis Tree Search: http://www.infochem.de/eng/texte/software_sts.htm.

(195) Chen, L.; Robien, W. MCSS: A New Algorithm for Perception of Maximal Common Substructures and its Application to NMR Spectral Studies. 1. The Algorithm. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 501−506.

(196) Chen, L.; Robien, W. Application of the Maximal Common Substructure Algorithm to Automatic Interpretation of ^{13}C-NMR Spectra. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 934−941.

(197) Chen, L.; Robien, W. Sophisticated Algorithm for Automatic Extraction and Analysis of Substituent-Induced Chemical Shift Differences on ^{13}C-NMR Spectra. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 441−446.

(198) Chen, L.; Robien, W. OPSI: A Universal Method for Prediction of ^{13}C NMR Spectra Based on Optimized Additivity Models. *Anal. Chem*. **1993**, *65*, 2282−2287.

(199) Chen, L. Computer-Assisted Structure Elucidation of Organic Compounds from ^{13}C NMR Spectra. Ph.D. Thesis, University of Vienna, Vienna, Austria, 1993; Chapter 7: Automatic Detection of Database Errors, pp 162−166.

(200) Chen, L.**;** Robien, W. MCSS: A New Algorithm for Perception of Maximal Common Substructures and its Application to NMR Spectral Studies. 2. Applications. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 507−510.

(201) Chen, L.; Robien, W. The CSEARCH-NMR Data Base Approach to Solve Frequent Questions Concerning Substituent Effects on ^{13}C NMR Chemical Shifts. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 217−223.

(202) Robien, W. Computerunterstützte Zuordnung von ^{13}C NMR Spektren. *Mh. Chem.* **1983**, *114*, 365−372.

(203) Allen, F. H.; Hoy, V. J. Cambridge Structure Database. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998; pp 155−167.

(204) Durant, J. L.; Briggs, R. L., Jr.; Nassau, D.; Leland, B. A.; Sasse, T.; Nourse, J. G. Cheshire: A New Scripting Language and its Use in Characterizing Databases. Fifth International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June, 1999.

(205) Ihlenfeldt, W. D.; Takahasi, Y.; Abe, H.; Sasaki, S. CACTVS: A Chemistry Algorithm Development Environment. In *Daijuukagakutouronkai Dainijuukai Kouzoukasseisoukan Shinpojiumu Kouenyoushishuu*; Machida, K., Nishioka, T., Eds.; Kyoto University Press: 1992; pp 102−105. (b) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(206) What is CACTVS ? http://www2.ccc.uni-erlangen.de/software/cactvs/whatis.html.

(207) Gardiner, E. J.; Artymiuk, P. J.; Willett, P. Clique-Detection Algorithms for Matching Three-Dimensional Molecular Structures. *J. Mol. Graph. Modell*. **1997**, *15*, 245−253.

(208) Cosgrove, D. A.; Willett, P. SLASH: A Program for Analysing the Functional Groups in Molecules. *J. Mol. Graph. Modell.* **1998**, *16*, 19−32.

(209) Wikipedia: Human Genome Project: http://en.wikipedia.org/wiki/Human_Genome_Project.

(210) Kerber, A.; Laue, R.; Moser. D. Ein Strukturgenerator für Molekulare Graphen. *Anal. Chim. Acta* **1990**, *235*, 221−228.

(211) Chen, L.; Zhang, M. A New Expert System for Structural Interpretation of IR Spectra. *Chem. J. Chin. Univ. (Ser. B)* (English version), **1990**, *6*, 289−294.

(212) Hanebeck, W.; Rafeiner, K.; Schulz, K.-P.; Röse, P.; Gasteiger, J. Towards the Automatic Generation of a Mass Spectrum from the Structure of a Compound. In *Software-Development in Chemistry 4*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1990; pp 187−195.

(213) Schulz, K.-P.; Gasteiger, J. The Elimination of Candidate Structures in Computer-Assisted Structure Elucidation Using the Mass Spectrum. In *Software Development in Chemistry 9*; Moll, R., Ed.; GDCh: Frankfurt/Main, 1995; pp 319−326.

(214) Selzer, P.; Schuur, J.; Gasteiger, J. Simulation of IR Spectra with Neural Networks Using the 3D-MoRSE Code. In *Software Development in Chemistry 10*; Gasteiger, J., Ed.; GDCh: Frankfurt/Main, 1996; pp 293−303.

(215) Steinhauer, L.; Steinhauer, V.; Gasteiger, J. Obtaining the 3D Structure from Infrared Spectra of Organic Compounds Using Neural Networks. In *Software Development in Chemistry 10*; Gasteiger, J., Ed.; GDCh: Frankfurt/Main, 1996; pp 315−322.

(216) Christie, B. D.; Munk, M. E. The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation. *J. Am. Chem. Soc*. **1991**, *113*, 3750−3757.

(217) Funatsu, K.; Nishizaki, M.; Sasaki, S. Introduction of NOE Data to an Automatic Structure Elucidation System, CHEMICS. Three-Dimensional Structure Elucidation Using the Distance Geometry Method. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 745−751.

(218) Bohanec, S.; Zupan, J. Structure Generation of Constitutional Isomers from Structural Gragments. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 531−540. (b) Bohanec, S. Structure Generation by the Combination of Structure Reduction and Structure Assembly. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 494−503.

(219) Peng, C.; Yuan, S.; Zheng, C.; Chen, L. From Spectra to Structure by Computer: Dreams and Reality. *Comput. Appl. Chem. 1994−1995*, 26−33. (b) Peng, C.; Yuan, S.; Zheng, C.; Hui, Y. Efficient Application of 2D NMR Correlation Information in Computer-Assisted Structure Elucidation of Complex Natural Products, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 805−813.

(220) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure − Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(221) TeleSpec: IR Spectra Simulation on the WorldWideWeb: http://www2.chemie.uni-erlangen.de/services/telespec.

(222) Faulon, J.-L. Stochastic Generator of Chemical Structure. 2. Using Simulated Annealing to Search the Space of Constitutional Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 731−740. (b) Stochastic Generator of Chemical Structures and Reactions: http://www.cs.sandia.gov/MICS/newsnotes/faulon_000202.htm.

(223) Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. A New Approach to Computer-Aided Molecular Structure Elucidation: the Expert System Structure Elucidator. *Lab. Automation Inf. Manage.* **1999**, *34*, 15−30.

(224) Röse, P.; Gasteiger, J. Automatic Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chim. Acta* **1990**, *235*, 163−168.

(225) Weise, A. Synthesis Simulation by Synthon Substitution. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 490−491.

(226) Gordeeva, E.; Lushinikov, D.; Zevirov, N. COMPASS Program: Combination of Empirical Rules and Combinatorial Methods for Planning of Organic Synthesis. *Tetrah. Comput. Method.* **1990**, *3*, 445−459.

(227) Hippe, Z.; Fic, G.; Mazur, M. A Preliminary Appraisal of Selected Problems in Computer-Assisted Organic Synthesis. *Recl. Trav. Chim. Pays. Bas.* **1992**, *111*, 255−261.

(228) Dogane, I.; Takabatake, T.; Bersohn, M. Computer-Executed Synthesis Planning, a Progress Report. *Recl. Trav. Chim. Pays. Bas.* **1992**, *111*, 291−296.

(229) Sello, G. Lilith: From Childhood to Adolescene. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 120−129.

(230) Moll, R. Context Description in Synthesis Planning. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 117−119.

(231) Barberis, F.; Barone, R.; Chanon, M. HOLOWin: A Fast Way to Search for Tandem Reactions with Computer. Application to the Taxane Framework. *Tetrahedron* **1996**, *52*, 14625−14630.

(232) Krebsbach, D.; Gelernter, H.; Sieburth, S. M. Distributed Heuristic Synthesis Search. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 595−604.

(233) Wipke, W. T.; Rogers, D. Rapid Subgraph Search Using Parallelism. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 255−262.

(234) The SynGen program for organic synthesis design: http://syngen2.chem.brandeis.edu/syngen.html.

(235) Mehta, G.; Barone, R.; Chanon, M. Computer-Aided Organic Synthesis − SESAM: A Simple Program to Unravel "Hidden" Restructured Starting Materials Skeleta in Complex Targets. *Eur. J. Org. Chem.* Volume 1998, Issue 7, 1409−1412.

(236) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P.; Wanke, R. Computer-Assisted Reaction Prediction and Synthesis Design. *Anal. Chim. Acta* **1990**, *235*, 65−75.

(237) Ihlenfeldt, W. D.; Gasteiger, J. Computergestützte Planung Organisch-Chemischer Synthesen: Die Zweite Programmgeneration. *Angew. Chem.* **1995**, *107*, 2807−2829. (b) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613−2633.

(238) Molecular Networks, GmbH: http://www.mol-net.de/.

(239) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO, and CAESA: Tools for De Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34−50.

(240) SimBioSys, Inc.: CAESA: Computer Assisted Estimation of Synthetic Accessibility: http://www.simbiosys.ca/caesa/.

(241) Chen, L. Reaction Classification and Knowledge Acquisition. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 1, pp 348−388.

(242) Chen, L.; Gasteiger, J. Organic Reactions Classified by Neural Networks: Michael Additions, Friedel−Crafts Alkylations by Alkenes, and Related Reactions. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 763−765. (b) Chen, L.; Gasteiger, J. *Angew. Chem.* **1996**, *108*, 844−846. (c) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4042.

(243) CORA: Classification of Organic Reactions for Applications: http://www2.chemie.uni-erlangen.de/software/cora/index.html.

(244) Davies, K.; Upton, R. Experiences Building and Searching the Chapman & Hall Dictionary of Drugs. *Tetrah. Comput. Methodol.* **1990**, *3*, 665−671.

(245) Gothe, S. A.; Helson, H. E.; Houdaverdis, I.; Lagerstedt, I.; Sinclair, S.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 22. The Generation and Use of Three-Dimensional Structures. *J. Org. Chem.* **1993**, *58*, 5081−5094.

(246) Wikipedia: Mac OS X: http://en.wikipedia.org/wiki/Mac_OS_X.

(247) Computer History: History for 2000 to today: http://www.computerhope.com/history/2000.htm.

(248) Wikipedia: 64-bit: http://en.wikipedia.org/wiki/64-bit_computers.

(249) Wikipedia: Multi-core (computing): http://en.wikipedia.org/wiki/Dual_core_processor.

(250) CAS STN Workshops: Reaction Searching in CASREACT: http://www.cas.org/training/reaction.pdf.

(251) Vaschetto, M. E. Personal communication, December 15, 2005.

(252) Accelry: Home > All Products > Chemical Databases > Failed Reactions: http://www.accelrys.com/products/chem_databases/databases/failed_reactions.html.

(253) He, M.; Yan, X.; Zhou, J.; Xie, G. Traditional Chinese Medicine Database and Application on the Web. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 273−277.

(254) Ihlenfeldt, W. D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C. Enhanced CACTVS Browser of the Open NCI Database. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 46−57.

(255) Frederick/Bethesda Data and Online Services: http://cactus.nci.nih.gov/.

(256) Elsevier MDL's DiscoveryGate: https://www.discoverygate.com.

(257) A Portrait of the Research Group of Professor Dr. Johann Gasteiger: 25 years of Research and Development in Chemoinformatics: http://www2.chemie.uni-erlangen.de/presentations/symposium/torvs_e.pdf.

(258) Rhodes, N.; Willett, P.; Calvert, A.; Humblet, A. CLIP: Similarity Searching of 3D Databases Using Clique Detection. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 443−448.

(259) Holliday, J. D.; Jelfs, S. P.; Willett, P. Calculation of Intersubstituent Similarity Using R-group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406−411.

(260) Gaizauskas, R.; Demetriou, G.; Artymiuk, P. J.; Willett, P. Protein Structures and Information Extraction from Biological Texts: the PASTA System. *Bioinformatics* **2003**, *19*, 135−143.

(261) Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftdatabase-Constructing a Free Chemical Information System with Open-Source Components. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733−1739.

(262) Elsevier MDL: Home > Products > MDL Discovery Framework: MDL Isentris: http://www.mdl.com/products/framework/index.jsp.

(263) Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L. Over 20 Years of Reaction Access Systems from MDL: A Novel Reaction Substructure Search Algorithm. *J. Chem. Inf.* Comput. Sci. **2002**, *42*, 1296−1310.

(264) Durant, J. L.; Leland, B. A.; Nourse, J. G. VET: A Tool for Reaction Plausibility Checking. *J. Chem. Inf. Model.* **2006** ASAP Web Release Date: 2006−02−18.

(265) *Chem. & Eng. News* April 25, 2005, p 5.

(266) NIH Roadmap for Medical Research: Molecular Libraries and Imaging: http://nihroadmap.nih.gov/molecularlibraries/.

(267) NCBI: Welcome to the Entrez Cross-Database Search Page: http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi.

(268) NCBI: PubChem Text Search: http://pubchem.ncbi.nlm.nih.gov/.

(269) *Chem. & Eng. News* April 25, 2005, p 5.

(270) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182. (b) A Free Database for Virtual Screening: ZINC is not Commercial: http://zinc.docking.org.

(271) Girke, T.; Cheng, L.-C.; Raikhel, N. ChemMine. A Compound Mining Database for Chemical Genomics. *Plant Physiol.* **2005**, *138*, 573−577.

(272) ChemMine: Center for Plant Cell Biology at UC Riverside: http://bioweb.ucr.edu/ChemMine/Documents/README.php.

(273) IUPAC: Current Project: Chemical Nomenclature and Structure Representation Division (VIII): IUPAC International Chemical Identifier (InChI): Promotion and Extension: http://iupac.org/projects/2004/2004-039-1-800.html.

(274) Steinbeck C. SENECA: A Platform-Independent, Distributed, and Parallel System for Computer-Assisted Structure Elucidation in Organic Chemistry. *J Chem Inf Comput Sci.* **2001**, *41*, 1500−1507. (b) Seneca: A Platform-Independent, Distributed and Parallel System for Computer-Assisted Structure Elucidation in Organic and Bioorganic Chemistry: http://almost.cubic.uni-koeln.de/jrg/software/seneca/.

(275) Fontana, P.; Pretsch, E. Automatic Spectra Interpretation, Structure Generation, and Ranking. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 614−619.

(276) Aires de Sousa, J.; Hemmer, M.; Gasteiger, J. Prediction of ^{1}H NMR Chemical Shifts Using Neural Networks. *Anal. Chem.* **2002**, *74*, 80−90.

(277) Da Costa, F. B.; Binev, Y.; Gasteiger, J.; Aires-de-Sousa, J. Structure-Based Predictions of ^{1}H NMR Chemical Shifts of Sesquiterpene Lactones Using Neural Networks. *Tetrahedron Lett.* **2004**, *45*, 6931−6935.

(278) Korytko, A.; Schulz, K. P.; Madison, M. S.; Munk, M. E. HOUDINI: A New Approach to Computer-Based Structure Generation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1434−1446.

(279) Schulz, K. P.; Korytko, A.; Munk, M. E. Applications of a HOUDINI-Based Structure Elucidation System. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1447−1456.

(280) Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. An Expert System for Automated Structure Elucidation Utilizing ^{1}H-^{1}H, ^{13}C-^{1}H and ^{15}N-^{1}H 2D NMR Correlations. *Fresenius J. Anal. Chem.* **2001**, *369*, 709−714.

(281) Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Molodtsov, S. G.; Martin, G. E.; Martirosian, E. R. Structure Elucidator: A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 771−792.

(282) Molodtsov, S. G.; Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Martirosian, E. E.; Martin, G. E.; Lefebvre, B. Structure Elucidation from 2D NMR Spectra Using the StrucEluc Expert System: Detection and Removal of Contradictions in the Data. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1737−1751.

(283) Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schulz, K.-P.; Herwig, A. Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 482−494.

(284) EROS: Elaboration of Reactions for Organic Synthesis: A Program for Reaction Prediction. http://zabib.chemie.uni-erlangen.de/software/eros/.

(285) SysChem: http://www.syschem.com/solution.htm.

(286) Rucker, C.; Rucker, G.; Bertz, S. H. Organic Synthesis − Art or Science? *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 378−386.

(287) Zhu, Q.; Yao, J.; Yuan, S.; Li, F.; Chen, H.; Cai, W.; Liao, Q. Superstructure Searching Algorithm for Generic Reaction Retrieval. *J. Chem. Inf. Model.* **2005**, *45*, 1214−1222.

(288) Schürer, S. C.; Tyagi, P.; Muskal, S. M. Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space. *J. Chem. Inf. Model.* **2005**, *45*, 239−248.

(289) Socorro, I. M.; Taylor, K.; Goodman, J. M. ROBIA: A Reaction Prediction Program. *Org. Lett*. **2005**, *7*, 3541−3544. (b) Socorro, I. M.; Goodman, J. M. The ROBIA Program for Predicting Organic Reactivity. *J. Chem. Inf. Model*. **2006**, *46*, 606−614.

(290) Schönberger, H.; Schwab, C. H.; Hirsch, A.; Gasteiger, J. Molecular Modeling of Fullerene Dendrimers. *J. Mol. Model.* **2000**, *6*, 379−395.

(291) Feuston, B. P.; Miller, M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K. Comparison of Knowledge-Based and Distance Geometry Approaches for Generation of Molecular Conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 754−763.

(292) Bywater, R. P.; Poulsen, T. A.; Røgen, P.; Hjorth, P. G. De Novo Generation of Molecular Structures Using Optimization to Select Graphs on a Given Lattice. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 856−861.

(293) Mekenyan, O.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45*, 283−292.

(294) Yi, M. Intel Likely to Tout Multicore Chips at Forum: New Microprocessor likely to Dominate Moscone Conference. San Francisco Chronicle, Monday, March 6, 2006: http://www.sfgate.com/cgi-bin/article.cgi?file=/chronicle/archive/2006/03/06/BUGH4HIG021.DTL&type=printable.

(295) Sutter, H. The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. Dr. Dobb's Journal, 30(3), March 2005: http://www.gotw.ca/publications/concurrency-ddj.htm.

(296) Wikipedia: Concurrency (computer science): http://en.wikipedia.org/wiki/Concurrency_%28computer_science%29.

(297) Doolen, G. D.; Mainieri, R.; Tsifrinovich, V. I.; Berman, G. P. *Introduction to Quantum Computers*; World Scientific Publishing Co. Pte. Ltd.: River Edge, NJ, 1998.

(298) Dimacs (Group), Nsf Science and Technology Center in Discrete Mathematics and theoreti (Corporate Author), Laura F. Landweber (Editor), Eric B. Baum (Editor), *DNA Based Computers II: Dimacs Workshop*, June 10−12, 1996, ISBN: 0821807560 (Dimacs Series in Discrete Mathematics and Theoretical Computer Science, V. 44.)

(299) President Clinton's Address (U.S.A.): http://www.pbs.org/newshour/bb/white_house/jan-june00/sotu5.html.

(300) International Technology Roadmap for Semiconductors (2005 Edition): http://www.itrs.net/Common/2005ITRS/ExecSum2005.pdf.

(301) (a) Markoff, J. Future of Chips Charted Report Forecasts Nanotechnologies in a Post-Silicon Era. *New York Times* Thursday, December 29, 2005. (b) Future of Chips Charted: Report Forecasts Nanotechnologies in a Post-Silicon Era: http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2005/12/29/BUGDEGEBRV1.DTL.

(302) Intel Multi-Core Processor Architecture Development Backgrounder: http://cache-www.intel.com/cd/00/00/20/57/205707_205707.pdf.

(303) Wikipedia: Fifth Generation Computer: http://en.wikipedia.org/wiki/Fifth_generation_computer

(304) CAS Databases: CASREACT - Answers Your Chemical Reaction Questions: http://www.cas.org/CASFILES/casreact.html.

(305) Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L.; Taylor, K. T. A New Generation of Reaction Indexing and Searching Methodologies. The CINF Session on Advances in Reaction Searching, ACS National Meeting, New York, U.S.A. September 7−11, 2003.

(306) Cambridgesoft: http://www.cambridgesoft.com/.

(307) Gasteiger, J. Symposium on De Novo Design and Synthetic Accessibility at the ACS Meeting in Atlanta, GA, U.S.A. on March 26−30, 2006.

(308) IBM: Deep Blue: http://www.research.ibm.com/deepblue/.

(309) Van Rozendaal, E. L. M.; Ott, M. A.; Scheeren, H. W. A LHASA Analysis of Taxol. *Recl. Trav. Chim. Pays-Bas*, **1994**, *113*, 297−303.

(310) (a) Campbell, M Knowledge Discovery in Deep Blue: A Vast Database of Human Experience can be Used to Direct a Search. *Commun. ACM* **1999**, *42*, 65−67. (b) http://archive.computerhistory.org/projects/chess/related_materials/text/5−3%20and%205−4.Knowledge_discovery_in_deep_blue/Knowledge_discovery_in_deep_blue.campbell-murray.1997.ACM.062303048.pdf.

(311) Rose, J. R.; Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74−90.

(312) Chen, L.; Gasteiger, J.; Rose, J. R. Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon Hydrogen Bonds to Carbon−Carbon Double Bonds. *J. Org. Chem*. **1995**, *60*, 8002−8014.

(313) *Chem. & Eng. News* Jan. 9, 2006, p 50.

(314) Pascual, R.; Mateu, M.; Gasteiger, J.; Borrell, J.; Teixidó, J. Design and Analysis of a Combinatorial Library of HEPT Analogues: Comparison of Selection Methodologies and Inspection of the Actually Covered Chemical Space. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 199−207.

(315) Vracko, M.; Gasteiger, J. A QSAR Study on a Set of 105 Flavonoid Derivatives Using Descriptors Derived from 3D Structures. *Internet Electronic J. Mol. Des.* **2002**, *1*, 527−544.

(316) Polanski, J.; Zouhiri, F.; Jeanson, L.; Desmaële, D.; d'Angelo, J.; Mouscadet, J.-F.; Gieleciak, R.; Gasteiger, J.; LeBret, M. Use of Kohonen Neural Network for Rapid Screening of Ex Vivo Anti-HIV Activity of Styrylquinolines. *J. Med. Chem.* **2002**, *45*, 4647−4654.

(317) Gasteiger, J. Data Mining in Drug Design. In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, E, 2001; pp 459−474.

(318) Spycher, S.; Nendza, M.; Gasteiger, J. Comparison of Different Classification Methods Applied to a Mode of Toxic Action Data Set. *QSAR Comb. Sci.* **2004**, *23*, 779−791.

(319) Behrendt, H.; Altschuh, J.; Gasteiger, J.; Kostka, T. Model Calculations to Assess the Fate of Triazines and Their Metabolites in Soil-Plant Systems. *Proceed. ECO-INFORMA'*97, Eco-Informa Press: 1997; pp 559−565.

(320) Reitz, M.; Sacher, O.; Tarkhov, A.; Trümbach, D.; Gasteiger, J. Enabling the Exploration of Biochemical Pathways. *Org. Biomol. Chem.* **2004**, *2*, 3226−3237.

(321) *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; J. Wiley & Sons: Chichester, 1998.

(322) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.

(323) Gasteiger, J.; Engel, T. *Chemoinformatics − A Textbook*; Wiley-VCH: Weinheim, Germany, 2003.

(324) Chen, L. Substructure and Maximal Common Substructure Searching. In *Computational Medicinal Chemistry and Drug Discovery*; Bultinck, P., Winter, H. D., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker: New York, 2004; pp 483−513.

(325) *Chem. & Eng. News* April 18, 2005.

(326) Walkins, K. J. BIOINFORMATICS: Making Sense of Information Mined from the Human Genome is a Massive Undertaking for the Fledgling Industry. *Chem. & Eng. News* February 19, 2001, pp 26−45: http://pubs.acs.org/cen/coverstory/7908/7908bus3.html.

(327) (a) Integrating Bioinformatics and Cheminformatics, Two-Day National Conference, November 14−15, 2001: http://www.iqpc.co.uk/binary-data/IQPC_CONFEVENT/pdf_file/1904.pdf. (b) Pharmainformatics: Integration of Bioinformatics & Cheminformatics At Spring ACS meeting in San Diego, April 1−5, 2001: http://www.chem.ac.ru/Chemistry/Conf/Apr01/ACSPHAR.en.html.

(328) (a) Clark, T. Quantum Cheminformatics: An Oxymoron? (Part 1). In *Chemical Data Analysis in the Large: The Challenge of the Automation Age*; Hicks, M. G., Ed.; *Proceedings of the Beilstein-Institut Workshop*, May 22nd − 26th, 2000, Bozen, Italy. (b) Beilstein-institut Home: http://www.Beilstein-institut.de/bozen2000/proceedings.

(329) Clark, T. Quantum Cheminformatics: An Oxymoron? (Part 2). In *Rational Approaches to Drug* Design; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, 2001.

(330) Compuinformatic biology is a counterpart of compuinformatic chemistry in biology. It is the integrated new field of computational biology and bioinformatics. In detail, compuinformatic biology is an interdisciplinary science in which mathematics, computer science, informatics, and biology merge into a single discipline. The major goals of compuinformatic biology are to develop biology-oriented mathematical models, computational and information technologies, and computer programs and apply these programs to studying and solving biological problems.

(331) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing Molecules by Optimizing Potentials. *J. Am. Chem. Soc*. **2006**, *128*, 3228−3232.

(332) *Chem. & Eng. News* March 13, 2006, p 33.

CI060016U