Chapter 1

The CAOS/CAMM Center: The Dutch national academic facility for computer-assisted organic synthesis and modeling

Jan H. Noordik
CAOS/CAMM Center, University of Nijmegen, the Netherlands

1. Introduction

The development of networks and PCs in the last two decades of the 20th century has fundamentally changed the practice of chemistry, and the union of chemistry and computer science has produced extremely powerful chemical research and educational tools. From the computer technology side much of the innovation was driven by the increasing availability, since the mid 1980s, of small desktop computers with graphics capabilities. From the chemical sciences side the development, around the same time, of algorithms to deal with chemical structures was a major step forward. Communication between chemists and computers now became possible in the basic language of the chemist, the structural formula. The keyboard as input device for long tables with atom names and bonds was replaced by the terminal screen for input of structures and reactions, mimicking a long-time practice in inter-chemist communication in the pre-computer era. By the mid 1980s many papers on computer applications in the more traditional fields of chemistry, like organic synthesis, began to appear not only in the Journal of Chemical Information and Computer Sciences but also in more traditional chemical journals. In analogy to its biological counterpart, bioinformatics, the term cheminformatics, or chemoinformatics, came into use to indicate the field. The techniques of computational chemistry, the chemical discipline where computers had been used from the beginning, were often excluded from the definition of the new discipline.

Cheminformatics, and more specifically the structure-based consultation of databases with factual chemical data, brought the power of computer technology into the laboratory and to the laboratory bench. The computer search for structural and molecular data, reactions, spectra and other chemical information needed to plan a chemical experiment, gradually replaced the paperwork involved in the usual preliminary library literature search and study. To cope with the rapidly growing amount of electronically-stored chemical information and to efficiently retrieve it in the Dutch academic environment of the early 1980s, considerable resources in terms of money, hardware, and expertise were required, and licensing costs for chemical databases presented a special financial problem. Scientific journals are copyright-protected and most chemical factual data files (X-ray data, spectral data, reactions) were constructed by extracting from the printed primary literature and then converting the data into digital files. Moreover, many of these information sources are either owned by publishers and/or copyright-protected too, making access generally expensive. This was a problem typically experienced for chemical data. Because, for example, biological data such as DNA and protein sequences have been in the public domain since the beginning of sequencing projects, mainly because communication in printed form was impossible and because early large-scale sequencing projects were publicly funded, the problem of expensive data access has been experienced much less in bioinformatics. In the Dutch academic environment of the mid 1980s the bench chemist lacked sufficient funding, hardware, and software resources, and the knowledge to implement and use electronically stored chemical information. Centrally-funded service activities for the academic community would help to overcome these difficulties.

Against this background, in 1984, in a bottom-up initiative at the faculty of science at the University of Nijmegen, a proposal for a Dutch academic cheminformatics center was formulated. The proposal stated that the center's main task would be, "to provide Dutch academic chemists with all the databases, software tools and computing facilities needed for molecular design, by making them on-line accessible at all Dutch universities". The idea quickly gained ground among Dutch academic chemists and, in January 1985, the CAOS/CAMM (Computer Aided Organic Synthesis/Computer Aided Molecular Modeling) Center was founded at the Faculty of Science of the University of Nijmegen, financially supported by the Dutch Ministry of Education, the Dutch National Science Foundation, and the University of Nijmegen.

2. A short history

2.1. The early days

The Center started its activities in 1985, at a time when funding of computerization of information had the warm sympathy of science funding bodies in the Netherlands. At the government level it was felt that Dutch academic science had to make up for a backlog in this field and one of the government initiatives to this end was the funding of 'Expertise Centers', i.e. centers where expertise in computer science and expertise in another scientific discipline were brought together. The CAOS/CAMM Center proposal fitted this government initiative remarkably well. In the Center, computer science techniques and chemical expertise were going to be combined to create an infrastructure for Dutch academic chemical information services to support academic chemical research. This required the creation of a hardware environment and a software library that could facilitate the use of a variety of techniques, ranging from relatively simple search procedures to complicated computational chemistry calculations. Initially the focus was on chemical services only, mainly because of the licensing costs problem mentioned above. Soon it became clear that with only small additional investments sequence data services could be accommodated too and the scope of activities was expanded accordingly.

To house the required programs and techniques and to provide the bench chemists at remote locations with the necessary computer resources and means of access, considerable computer power was required. At a time when scientific computing was the hallmark of centralized time-sharing systems, the choice was made to build the infrastructure around a 'large' central system, which at that time was a VAX 11/785. This machine could accommodate database applications as well as provide the computing power needed for modeling purposes. For external access and data communication, the Dutch academic network SURFnet (see www.surfnet.nl), the first phase of which had just become operational, was used. Still living in the pre-PC era, relatively low-cost graphics terminals were installed by the Center at all Dutch academic chemical laboratories. The Center's initial services 'toolkit' contained organic chemistry databases, structure databases, and molecular modeling programs, and, within a year from its start, chemical database services (reactions, structures and spectra) could be offered nationwide. Biomolecular services (DNA and protein sequences and protein structures) were added in the next year and soon the Center's software library offered access to several organic reaction-retrieval and synthesis-planning systems, organic, inorganic, and protein crystal structure databases, DNA and protein sequence databases, and several molecular modeling and molecular mechanics and dynamics programs. At the Tübingen Workshop, *Computer in der Chemie*, in November 1988, the Center contributed with a paper in which the CAOS/CAMM Center services were presented as "An Integrated System of Computer Assisted Chemistry Tools" [1]. In this context the indication 'Integrated System' referred to the graphics user interface [2] that had been developed to simplify access to the various available databases and programs.

2.2. The 1990s

By the end of the 1980s, with the mass production of personal computers, the availability and capabilities of these machines increased rapidly and prices of hardware and software went down quickly. Dedicated (graphics) computer terminals vanished and most laboratories gradually replaced the graphics terminals, installed by the Center, with the much more versatile desktop computers. Simultaneously, the concept of time-sharing on one large system rapidly lost ground in favor of local workstations, often operating under (freeware versions of) the operating system UNIX. However, because of the license costs of chemical databases, as pointed out above, and because of the storage capacity required by many database applications which was then not available on PCs, centrally-offered services were still very much needed, particularly in the Dutch academic environment. In line with these developments, the Center therefore replaced its central VAX computer with a cluster of dedicated computer systems for the different applications and by 1992 the Center's computing environment consisted of a cluster of MicroVaxes for, for example, reaction retrieval, a pipelined-vector Convex system for very time-consuming modeling and computational chemistry calculations, and a cluster of DECstations, running UNIX, for software development and network access control. In the following years this configuration was basically kept as such, except that, with the disappearance of Digital and Convex as computer manufacturers, other manufacturers' products, mainly Silicon Graphics, replaced the older components in the Center's computer cluster.

2.3. Users

The Center's first users were mainly organic chemists. However, with the expansion of the services to other chemical disciplines, notably the biomolecular sciences, the Center's user groups were going to span a broader spectrum of researchers and students and, in the early 1990s, the Center's services reached into all Dutch university laboratories, with more than 500 scientists in over 80 departments holding a subscription. Because different user groups required different services, the software and database library of the Center was organized in a number of different packages, tailored to organic chemists, structural and computational chemists and biochemists and biologists. Thanks to major government funding, subscriptions to these packages could be issued at nominal fees.

2.4. User training

Together with the data services, an elaborate practical courses program was initiated to teach efficient use of the chemical data storage and retrieval methods. For hands-on training a lecture room, initially equipped with terminals and one advanced color display graphics workstation and in later years with powerful PCs, was established at the Center. In an extensive documentation and manual production effort, training and instruction material was produced. The practical courses program, aiming at initial user training, proved to be very successful and in the first five years of the Center's existence many hundreds of chemistry students and researchers from all Dutch universities followed these courses. Additionally, a high-level user training program, focusing on the scientific methodology used in the different databases and programs, supported all services. In this way the Center provided the chemist who had neither a detailed knowledge of computers, operating systems, and networks, nor of all available chemical databases and programs, with access to essential data storage and retrieval 'laboratory equipment' as well as with the knowledge of how to use these tools. As a spin-off from the training programs of the Center new teaching materials have been developed in which modern electronic aids like PC, the Web, and CDs (see also Section 4 below and Chapter 8 in this issue) are used.

2.5. Research

To maintain the service tasks and the help desk support for data access, data exchange, and data interpretation at an academic level, an appropriate services-supporting research program was implemented. Information science knowledge and techniques were used to develop user interfaces and data conversion and formatting routines [3], and database search programs were optimized to use the specific hardware configuration of the Center [4]. Research on cheminformatics topics in the fields of modeling, crystal structure prediction, reaction retrieval, and synthesis planning, and on a variety of bioinformatics topics, has been published in many papers and Ph.D. theses (see paragraphs on research below). By the mid 1990s the CAOS/CAMM Center group numbered about fifteen people (see acknowledgements below).

2.6. From CAOS/CAMM to CMBI

Over the years the Center's user community for *chemical* data services remained rather constant while the *molecular biology* (mainly sequence data) user group showed a steady growth rate. In 1998 more than half of the Center's service activities concerned sequence data-related services. As a consequence of this success in attracting new service users for the 'booming' bio-related research, and despite the Center's achievements in cheminformatics services and research in the past decade, both university and national funding policies changed. The warm sympathy of science funding bodies which enabled the start of the Center as a cheminformatics facility in 1985 now shifted to bioinformatics and the Center had to follow this change. By the end of the millennium this policy change culminated in the renaming of the CAOS/CAMM Center into CMBI, the Center for Molecular and Biomolecular Informatics, with a very strong emphasis on bio. This renaming effectively marked the disappearance of the CAOS/CAMM Center as a *national* service center for cheminformatics and its change into a *local* Nijmegen University bioinformatics research group. Although some of the original national cheminformatics services are still available, most research and development in this field has been frozen at the 1999 level.

3. Services and research

As cited above, the CAOS/CAMM Center was established *to provide Dutch academic chemists with all the databases, software tools, and computing facilities needed for molecular design*. This mission statement emphasized the service tasks but involved also a significant research component, both to maintain the necessary scientific level of the services and to contribute to the chemical research projects of the faculty. In the following paragraphs attention will be given to both aspects.

3.1. User interfacing

From the beginning of the Center's operations, the emphasis has been on offering an integrated system of tools for chemical information retrieval and computing aiming at academic bench chemists as users. To use the Center's tools only minimal expert computer system knowledge should be required. With regard to computer resources, the only choice for offering these type of services was a central computer system with options for remote access. Remember that around the mid 1980s a common reaction retrieval system required about 400 MB in storage capacity for a set of about 100,000 reactions, and several databases of this size had to be stored. Although now hardly imaginable, because any modern PC has one or more GB-size disks, at that time only large central servers could provide such an amount of storage capacity. Given a central server and many simultaneous users, one of the success-determining factors for a data service facility was going to be server speed and network speed and reliability. The Center's VAX11/785 provided storage capacity and server speed. Network facilities could be provided by the just-founded academic computer network SURFnet (www.surfnet.nl). Another success-determining factor would be the ease of access to the services. Because most of the systems and programs that were to be implemented by the Center were developed elsewhere, mostly as a stand-alone academic or commercial software packages, interconnection of the different database and/or program systems and user interfacing were going to be major tasks for the new facility. Therefore much effort was invested by Center staff to design and develop an easy-to-use interface. In more than ten years of development and usage this CAOS/CAMM user interface has been the main entrance to all databases and program systems available at the Center. Over the years many new elements, offered by new software technology, were integrated, constantly improving the basic intention of shielding the bench chemist from the specific computer and operating system environments housing the different information services.

In the pre-Web era our user interface was a keyboard-oriented menu system. With just a few mouse clicks or keyboard-controlled cursor movements, even on the older graphics terminals it provided the users with a connection to a specific group of services, such as reaction retrieval. Initially, when all services were provided by one server system, the interface could be correspondingly simple. But in an expanding and changing hardware and software environment the interface soon had to cope with more and more different computer platforms at one end and with a broad spectrum of databases and computational tools at the other end. Typically, around 1990, the interface functionality offered a remote user system access via a Telnet [5] session to the Center's Internet domain name. The Telnet connection opened a terminal window on the remote server computer, displaying it at the user's terminal and supplying him with a menu screen. Depending on the service wanted, different layers of sub-menus were offered and, by selecting a specific service with the keyboard cursor control keys and activating the selected service by a carriage return, this service executed interactively at the dedicated hardware box of the remote server system. Initially the VAX/VMS Fortran-based interface used simple ASCII menu definition files to define the menu buttons. The user interface options as described might now seem rather trivial; they certainly were not in the late 1980s and the early 1990s and the simple and easy access to the rapidly-growing amount of research data, provided by this interface, explains much of the popularity of the Center's services among bench chemists at that time. In the mid 1990s the spectacular growth of WWW-based services and the proliferation of personal workstations and computer networks urged a complete overhaul of the user interface. X11 graphics options were added and a WWW menu version was developed. This new Web interface proved to be an excellent access mechanism to the Center's chemistry tools and has enhanced its user-friendliness once again. The user's familiarity with his preferred Web browser, allowing the seamless integration of supporting HTML documents in a data search and retrieval session, and the options to view, copy, delete, and upload files from the server to the user's local file space, now allowed the bench chemist to focus completely on chemistry.

3.2. Computer-Assisted Organic Synthesis (CAOS): Services

3.2.1. Barriers to use

Although CAOS software has been around since the mid 1960s, its impact on organic chemical research was very limited up to the last decade of the 20th century [6]. Against this background, the contrast with the present day situation, where bench chemistry without prior reaction database consultation is unthinkable, is the more remarkable. A major reason for the rather slow rate of penetration of computerization in the organic laboratory was undoubtedly the lack of affordable graphics hardware and the lack of software which could communicate in the chemist's natural language, the structural formula. When both these barriers were leveled by the early 1990s, CAOS programs and, in particular, reaction retrieval systems rapidly found their way on to the organic chemist's workbench [7]. In Dutch academic organic research CAOS/CAMM Center services have played a decisive role in this breakthrough. From its very start the Center has implemented reaction databases and search software and provided access to these systems as a service for all Dutch academic chemists. At the same time a user training and a synthesis design software development project were started to support these services.

The use of computers to solve problems in organic chemistry requires the machine to be able to communicate in the complex language of organic chemistry and this requires the formal description of many fundamental chemical concepts which are often intuitive to the chemist. Major algorithmic developments since the mid 1980s enabled direct encoding of these concepts in computer programs. In his 1995 thesis [8] Jan-Willem Boiten, one of the CAOS/CAMM Center's Ph.D. students, mentioned the following topics highlighted in the literature.

- Man-machine communication.

- Structure storage.
- Structure comparison.
- Substructure searching.
- Ring identification.

To better understand the importance of these developments for chemical data storage and retrieval, a short comment on each of these topics is given.

Man-machine communication. With the development of graphical user interfaces and the corresponding development of appropriate and affordable hardware, the keyboard input of structures in the widely-used Wiswesser line notation [9] came to an end in the early 1980s. Ongoing developments in computer graphics [10] have since then resulted in options for structural diagram input and output in all synthesis-related programs.

Structure storage. Reaction retrieval and synthesis programs extract chemical information from the input structure and store this information in computer-manageable form, mostly a connection table [11,12]. This table contains a detailed topological description of a molecule in a well-defined format. The connection table is used for both permanent storage, when written to a file, and for exchange of structural information with other programs. This latter usage is hampered by the fact that connection table formats still differ from program to program, despite several attempts at standardization [13–16]. Translation of connection tables from one format into another has always been one of the services provided by the CAOS/CAMM Center.

Structure comparison. The identification of identical structures is an important prerequisite of any synthesis-related computer program. An exhaustive one-to-one comparison of all atoms and bonds in two molecules, or of all atoms and bonds involved in a reaction, is an extremely laborious task. When N atoms are involved this requires N! permutations. Canonicalization algorithms [17–21] have been developed to solve this problem. These algorithms assign a unique numerical code to each structure. The code is derived from a systematic ordering of the atoms in a molecule, based on atomic properties like connectivity, atom type, and bond order. One-to-one comparison of structures is only necessary if they end up with the same canonicalization number, which usually means that they are identical. At the Center we have used the coding of a canonicalization algorithm by our students as a very instructive method in teaching computer-assisted organic synthesis.

Substructure searching. Searching of reaction or structure databases is nowadays standard practice in daily laboratory activities. Most often, queries concern searches for reactions or structures with only partial resemblance to the search structure. To enable this substructure searching in extensive databases in a time acceptable in an interactive process, development of efficient searching algorithms [22–28] has been a research topic for many years. Computational problems associated with substructure searching were considerable because exhaustive mapping of a query and a target structure was impossible in databases of any practical size. The solution has been found in structural screens, small structural fragments defined for each database compound, and one-bit coded. With efficient screens, a substructure search consists of a rapid screen search, i.e. a comparison of query structure and database structure screens, followed by a mapping search. If screens are properly chosen the vast majority of database entries will be skipped in the first pass and only few structures will enter the more time-consuming second step. Atom-to-atom mapping is essential to specify that a particular substructure in the product is derived from a corresponding substructure in the reactant. With mapping it can be indicated, for example, that in a reaction query for selective ester reduction in the presence of an aldehyde, the resulting primary alcohol derives from the ester and not from the aldehyde.

Ring identification. A special case of structure comparison is the recognition of rings. Molecules with rings form a major part of all organic compounds and ring size is a key feature. The development of ring perception algorithms therefore was one of the earliest research topics [29,30] in computer-assisted organic synthesis research. The complexity of the problem is easily illustrated by the example of fused ring systems, which has spawned the concept of the smallest-set-of-smallest-rings (SSSR) [31].

In addition to software developments for the topics above, many other chemical concepts had to be translated into efficient algorithms. Without being exhaustive we can mention aromaticity, tautomerism, valency checking, and perception of stereochemistry. These issues are addressed in many publications which have appeared from the beginning of computer-assisted organic chemistry research in the early 1970s [32].

3.2.2. Fields of application

If CAOS software is divided according to functionality, two main groups of programs can be distinguished:

- Reaction retrieval programs.
- Synthetic analysis programs.

Both groups share a number of basic underlying algorithms but differ in their application. Reaction retrieval is nowadays an indispensable tool for studying almost any organic synthetic problem. Synthetic analysis is used much less frequently, mostly as a complementary tool to support human creativity to design a synthesis. Reaction retrieval has become a leading computer tool in organic chemistry. Synthetic analysis is much more a sometimes helpful generator of ideas. Both fields have been actively supported at the CAOS/CAMM Center. A third group of CAOS programs, briefly commented on later, deals with spectral data, mainly mass spectra, IR and UV spectra, and 13C NMR spectra.

3.2.2.1. Reaction storage and retrieval

The number of publications on organic chemistry and organic reactions is historically so large that collective series summarizing the primary literature have been the traditional means to assist the chemist to keep up with this avalanche of information. With the development of the computer technology described above, reaction databases gradually took over this role. Several organic reaction database systems have been introduced since the early and mid 1980s. Some of these offer complete primary literature coverage, others aim at selections of useful and carefully-compiled reactions.

CASREACT [33] was the main database of the former type up to the mid 1990s when it got competition from the CROSSFIRE [34] system, constructed by the Beilstein Institute on the basis of the information in its printed volumes.

Examples of early systems with selections of reactions are SYNLIB [35,36], ORAC [37] and REACCS [38]. These databases were considerably less memory-consuming than the comprehensive files, and were intended for in-house usage, whereas systems like CASREACT used external hosts and network access. Although much smaller than the comprehensive data files, the selective systems still required 0.5 KB (SYNLIB) to 5 KB (ORAC) disk space per reaction. With databases with up to 100,000 reactions, only larger central computer systems could provide so much memory space and for Dutch academic usage the CAOS/CAMM Center offered this. This situation changed only by the end of the 1990s when PCs with GB-size disks became affordable. Although from then on PCs could technically accommodate reaction databases, database license costs generally prevented individual laboratories starting their own services. Cost reduction from sharing license costs for a central service caused a continued demand for such a central service and, even after its transformation into a bioinformatics research group in 1999, network access to all relevant organic reaction databases has continued.

In a sense, some of the selective reaction retrieval systems were developed as a spin-off from earlier-developed synthesis design programs: ORAC from LHASA and REACCS from SECS. Using the display and structure manipulation techniques developed for these systems, answers to queries about scope and limitations and reaction conditions, indispensable for synthesis design programs, could easily be provided by searching the stored selections of actually-performed and published reactions.

Nowadays both the network-accessible comprehensive reaction retrieval systems and specialized in-house databases are a part of standard laboratory equipment. Through mergers and takeovers only a few reaction retrieval database systems and vendors are left. In 1991 ORAC Ltd and MDL (REACCS) merged and, with this takeover, ORAC effectively disappeared as a retrieval system. After MDL's development of a new generation of software products (ISIS, IRDAS), which integrated PC and host-based systems during the mid and end 1990s, REACCS also disappeared from the market as system. MDL has frozen all ORAC- and REACCS-specific databases at the 1992 level and has joined them into one file, the Reference Library of Synthetic Methodology, containing about 210,000 reactions searchable with ISIS. Since then several new databases, such as CSM (Current Synthetic Methodology) and JSM (Journal of Synthetic Methods), have been developed by MDL, now owned by Elsevier.

At about the same time – the beginning of the 1990s – former ORAC staff started to operate as an independent reaction database supplier under the name SYNOPSIS. They compiled recent chemistry in files such as MOS (Methods in Organic Chemistry), PG (Protective Group) chemistry, SPS (Solid Phase Chemistry), and BioCatalysis (biomolecule-catalyzed reactions). Because of the careful reaction selection process, these files have a high information content. Around the year 2000, in a new merger, Molecular Simulations (MSI), Oxford Molecular (OML), Genetics Computer Group (GCG), and SYNOPSIS together formed ACCELRYS (see www.accelrys.com), which in turn is a subsidiary of Pharmacopeia Inc.

3.2.2.2. Synthetic analysis

This topic will be treated in more detail in Chapter 4 and in this paragraph only some introductory remarks and the use of the methodology at the CAOS/CAMM Center are presented. While planning or designing a synthesis, organic chemists are confronted with some fundamental questions:

- How does a reaction proceed, given one or more chemical compounds and certain reaction conditions?
- How does one find the optimal sequence of reactions that will result in the desired product from available starting materials?

Synthetic analysis programs dealing with the first question are generally classified as reactivity prediction programs. The second question is addressed in synthesis design programs.

A further distinction along similar lines into logic-oriented [39] and information-oriented systems refers to the chemical knowledge forming the reasoning basis of the systems. Information-oriented programs rely on a library of known chemistry. This reaction library or knowledge base contains the chemical knowledge of the system. As mentioned above, as a spin-off, some of these reaction libraries have evolved into reaction retrieval systems. The contents of the underlying library necessarily limit the application. Reactions which are not represented in the knowledge base cannot be generated. Information-oriented systems, a typical representative of which is LHASA [40], are mostly used for synthesis design.

Logic-oriented systems use schemes of bond-breaking and bond-forming steps, mechanistic and thermodynamic considerations, or even mathematical models to generate reactions. Unprecedented chemistry or hitherto unknown reactions may result from these programs. To separate possible and impossible reactions, evaluation functions with parameters like heat of formation, electronegativity, bond dissociation energy, and polarizability are used. Logic-oriented systems are mostly used for reaction prediction and mechanistic evaluation. Well-known examples come from the groups of Ugi [41], Gasteiger [42], Ihlenfeldt [43], Hendrickson [44], and Jorgensen [45].

A completely different approach to synthetic analysis is found in programs applying AI (Artificial Intelligence) methodology, comprehensively reviewed in 1991 [46].

Many of the synthetic analysis programs mentioned above have been actively supported at the CAOS/CAMM Center, either as an on-line service or as a local stand-alone program, and the methods have been used in many internal and external collaborations [47].

3.2.2.3. Spectral data

Spectral databases can provide useful information to bench chemists and as such they belong to the library of valuable cheminformatics tools. For that reason the CAOS/CAMM Center has, in its CAOS package, offered access to different spectral files and programs. For the interpretation and analysis of mass spectra, mass spectrometers were already equipped with stored reference spectra in the 1980s but similar data files were also available as a stand-alone database. Identification software in which the spectrum of an unknown compound could be compared with stored spectra, and software to analyze or to interpret unknown spectra, completed the database. For several years the Center has offered on-line access to such programs based on the development work of McLafferty at Cornell University. However, by the mid 1990s mass spectrometer computers and data files had become so large and advanced that Dutch academic interest in the stand-alone versions of these databases disappeared.

Also around the end of the 1980s, the Center experimented for a few years with the provision of on-line access to a 13C NMR spectral database system but, although this command-driven program had a graphical structure and spectra output module, interest from bench chemists did not reach the expected level and services were terminated.

A spectral database system that was available for many years as an on-line service to CAOS/CAMM Center services subscribers was SpecInfo (www.chemicalconcepts.com). A recent release of this system is now available as an Internet-based application providing database searching for NMR spectral libraries with, for example, C13, 1H, 31P, and other spectra, for IR and MS spectral libraries and an NMR spectral prediction option.

3.3. Computer-Assisted Organic Synthesis (CAOS): Research

3.3.1. Reaction retrieval database comparison studies

When, in 1985, network access to the Center's organic reaction databases had been realized and services had started, a research program supporting the CAOS services was initiated. For better support of reaction retrieval services, there was felt to be a need for a uniform access option to databases from different systems and suppliers. On the one hand this would enable Center staff to perform comparative evaluations for database users, focusing on database content. On the other hand it would enable the Center to be, as a service provider, as independent as possible from a specific search system. In the period 1987–1995 a database comparison research program to develop the required options resulted in several database comparison and reformatting studies [8,48–51].

3.3.2. Synthesis planning

For synthetic analysis research the Center started a successful collaboration with the development group of the LHASA system. In the LHASA program, under development at Harvard University since the early 1970s, E.J. Corey's approach to retrosynthetic analysis [40] was implemented. Synthesis planning is the topic of Chapter 4. Here only some of the main lines of the research performed at the Center will be mentioned.

In retrosynthetic analysis the structure of a target molecule is transformed into simpler molecules. Reactions viewed in the retrosynthetic direction start with the product and go backwards to the reactants. Retro-reactions or transforms are the imaginary counterparts of reactions. They are thought processes. The introduction of this concept opened the way for a general and systematic approach to complex syntheses, first on paper and then with the advent of computers with the LHASA program. The Harvard LHASA research was joined by a group at Leeds University and by the Dutch CAOS/CAMM group (Dr. Ott) in the early and mid 1980s.

LHASA consists of a control program and a knowledge base containing the chemistry in the form of formal descriptions of transforms. In the collaboration the Harvard group mostly concentrated on control program development while both the Leeds and the CAOS/CAMM group concentrated on the development of the knowledge base. Important CAOS/CAMM group contributions are the Quinone Diels-Alder transform, a new module for steroid synthesis, and the implementation of stereochemical aspects.

- **Steroid synthesis**. With desogestrel as a model compound and LHASA as the synthesis planning program, different synthesis strategies were evaluated. Condensation reactions and cycloaddition reactions were among the frequently occurring desogestrel synthesis routes generated. Polyene cyclizations were missing as a possible synthesis strategy, a shortcoming in the LHASA methodology. Through the development of a new program module which could recognize strategic polyene cyclizations in various ring systems and with the functionality to generate the required initiations and terminations, the applicability of LHASA for steroid synthesis planning could be significantly enhanced.
- **Stereochemical aspects.** To improve LHASA's capabilities with respect to 3D perception of 2D input structures, new algorithms for optimization of acyclic fragments, for better perception of symmetry and for the management of different conformations, were developed and implemented. A pilot database extension with 3D molecular fragments of steroids in different conformations has been constructed and tested.

The LHASA development research at the Center has been described in two Ph.D. theses (Boiten, 1995 and Ott, 1996 [8,51].

Apart from the research application at the Center, the LHASA system has also found significant use in Dutch academic organic chemistry.

3.4. Computer-Assisted Molecular Modeling (*CAMM*)*: Services*

3.4.1. Visualization and computing

Modeling involves a number of different aspects of structure handling by computers, many of which are excellently reviewed in the book series *Reviews in Computational Chemistry* [52]. In the early days of the computer era computers were merely used as computing machines and as such were perfectly suited to finding solutions for, for example, the Schrödinger equation. Quantum chemical calculations therefore were among the first computer applications in chemistry. Started in the 1960s, computational chemistry has evolved from this application as a new chemical discipline. Although nowadays *ab initio* and semi-empirical quantum chemical methods belong to the chemist's standard equipment, computational chemistry programs as such are generally not considered to be typical cheminformatics tools. Storage of the results of quantum chemical calculations and development of database tools for retrieval, however, are. The Quantum Chemical Literature Database (QCLDB, see below) therefore has long been one of the databases to which access has been provided by the CAOS/CAMM Center.

Visualization of the results of quantum chemical calculations is another typical cheminformatics application. One of the successful developments in this field at the CAOS/CAMM Center is the program system MOLDEN (described later). Because of the complexity and the inherent computer time requirements of quantum chemical calculations, even on modern supercomputers, these methods allow only calculations on 'small" molecules, up to, say, 50 carbon atoms. For calculations on biomolecules such as proteins, faster methods had to be developed and in the mid 1960s the first molecular mechanics (MM) calculations were published. In the beginning these were also only for small molecules but soon the methodology proved to be sufficiently robust to tackle biomolecules. The further development of these fast empirical calculation techniques in the following years, and their combination with the possibilities offered by the graphical display options which became available ten to fifteen years later, created in the 1980s hitherto unknown modeling possibilities. Features arising from the three dimensional architecture of molecules could now be explored and studied interactively.

3.4.1.1. MM and MD

Modeling based on molecular mechanics (MM) and molecular dynamics (MD) techniques uses an empirical interaction function (see Chapter 7 in this issue), the force field, describing the potential energy of a molecular system as a function of the atomic coordinates. The potential energy is calculated as a sum of parameterized force field terms, the parameters of which are determined from experimental (empirical) data. The covalent structure of molecules is represented with harmonic potentials in terms of equilibrium bond lengths, bond angles, torsion angles, and dihedral angles. Interactions between non-bonded atoms are represented by, for example, a Lennard-Jones potential (van der Waals interactions) and a Coulomb potential (interaction between charged particles). Depending on available computer power and the required degree of sophistication, other terms can be included in the force field.

Although atomic behavior fundamentally follows quantum mechanical laws, the classical mechanics description of (macro)molecules is valid as long as no chemical reactions (electronic rearrangements) take place. MM methods calculate molecular equilibrium geometries by minimization of the molecular potential energy. Different mathematical minimization techniques, based on variation of atomic coordinates, are used to find the energy minimum. A problem arises if the potential energy surface has several local minima, as is always the case with biomolecules. Then the global energy minimum is most often not found and minimization ends in one of the local minima. To overcome this problem MD techniques can be used. MD simulates the dynamic behavior of molecules by calculating a series of atomic positions as a function of time. Beginning with a starting structure, for example a crystal structure, all atoms in the system are given a temperature-dependent initial velocity, after which the Newtonian equations of motion are solved by numerical integration. New atomic positions and velocities are calculated and, with the atoms moved to these new positions, a new calculation cycle is started. From the resulting trajectories, i.e. the atomic positions as a function of time, interesting dynamic molecular properties can be calculated, and so can the equilibrium conformation. MD calculations at sufficiently high temperatures and with a sufficiently large number of time steps will overcome all energy barriers between local minima, thus allowing searching of conformational space. Selection of structures at regular time intervals and minimization to the associated minimum energy structure, should lead to the global minimum.

3.4.1.2. Software toolkit

To provide access to, and to lower the barrier for the usage of, computational and graphics display tools, from its start in 1985 the CAOS/CAMM Center has implemented several modeling and structure visualization programs based on the MM and MD methodology. Access for Dutch academic chemists to, for example, QUANTA (www.accelrys.com) CHEMX [53], MACROMODEL [54], and SYBYL (www.tripos.com), and to user training courses for these programs, were covered by the Center’s ‘modeling package’ to which laboratories and individual researchers or students could subscribe. Several research projects (see below) were initiated to support these CAMM services. In recent years interest in centrally-offered services has markedly decreased. Where professional visualization until some years ago was the exclusive domain of expensive central graphics facilities, now local PCs and workstations offer such advanced graphics capabilities that only software licensing costs form the limiting factor for implementation.

3.4.2. Storage and retrieval of 3D data

Besides calculation and visualization of structural features, computer storage and retrieval of structural data are important aspects of modeling. Soon after the discovery of X-rays as a phenomenon and of the possibility of using this short-wavelength radiation as a means to determine the three-dimensional structure of molecules, the first papers with sets of atomic coordinates appeared. From this time, around 1915, up to the 1960s, when computer methods became available to perform the complex calculations necessary for a structure determination, the number of structures published was limited. Thereafter it increased almost exponentially. To study and use the wealth of information that became available with this rapidly increasing number of paper publications of structural data, new technologies had to be used. An important initiative, taken in Cambridge in the early 1960s, marked the start of computer storage of structural data of organic and metal-organic compounds. Similar initiatives for inorganic structures and structures of biomolecules have followed and it is now safe to say that the structural data of all compounds that have ever been determined by X-ray or neutron diffraction techniques are stored in and retrievable from databases. The Cambridge Structural Database (see Chapter 6 in this issue) is a comprehensive database containing three-dimensional coordinates of organic and metal-organic compounds. The Inorganic Crystal Structure Database, also comprehensive, contains the structures of inorganic compounds and metals. The Protein Database holds the atomic coordinates of mainly proteins and some other biomolecules. Some basic information on all three databases is given below.

3.4.2.1. Databases

To provide Dutch academic chemists with the possibilities offered by direct on-line access to this enormous and continuously updated information resource, facilitation of access to these databases, supported by user training, a help desk, and documentation of the services, has been a major task of the CAOS/CAMM Center, right from its foundation in 1985. In recent years PC storage capacity and computation power have increased to such an extent that local database implementations are now possible, thus reducing the need for central facilities. A similar trend was noticed for visualization services. Common PC storage capacity is now GB-size and with a size of, for example, less than 1 GB for the CSD, purchase on CD and local storage are a real alternative for a central facility. Some introductory comments on the main CAMM database services offered by the Center follow.

- **QCLDB** (**Quantum Chemical Literature Database**) [55]. This database (www.jaici.or.jp) references the literature on ab initio quantum chemical calculations and covers publications from 1978 on. The 2002 release already contains over 57,000 entries. A hard copy version of the database is published in the *Journal of Molecular Structure: Theochem* (Elsevier). The database is maintained by the Quantum Chemistry Database Group of the Institute for Molecular Science and the National Center for Science Information Systems in Japan. The file contains literature citations of ab *initio* computational work on atomic and molecular electronic structures. Semi-empirical calculations are excluded. Data are extracted by expert scientists from over 100 journals, about one-third of which are searched exhaustively. Because the quantum chemical calculations cited provide useful information to a wide range of scientists, varying from chemists and physicists to biologists, the CAOS/CAMM Center has included the QCLDB in its CAMM services. In its original form, the file allowed keyboard-oriented alphanumeric searches only. At the Center the file has been made accessible through a relational database system equipped with an easy to use Web browser (www.cmbi.kun.nl), but updating of this service ended with the conversion of the Center into the CMBI. New QCLDB releases are now accessible as a Web service in Japan.

– **CSD** (**Cambridge Structural Database**) [56]. Begun in the early 1960s, the CSD (www.ccdc.cam.ac.uk) is the electronic compilation of evaluated experimental small molecule structural data. It is comprehensive and covers the structures of all organic and organometallic compounds determined by X-ray and neutron diffraction techniques. Each entry in the database, identified by a unique reference code, consists of:

  ∗ a chemical and bibliographic summary (chemical name, molecular formula, chemical classification by structure type, literature citation, and comments);
  ∗ a two-dimensional structure description (atom types, connectivity, bond types, cyclicity);
  ∗ a three-dimensional structure description (x,y,z atomic coordinates, unit cell, space group, and R-factor).

During database building, bit screens are added to each entry in the database. These screens are a bit-wise compact summary of all searchable information and used to speed up the search procedure. For each search query, bit screens generated from the query are compared with the screens stored in the database and atom-by-atom searching follows only in case of hits. Over the years the Cambridge Crystallographic Data Center has upgraded the CSD interrogation program from an alphanumeric and keyboard-input-driven program to an easy-to-use graphical user interface allowing two- and three-dimensional substructure query input. Data analysis programs are now also part of the CSD software library. The latest development is the new interface software, ConQuest. Query construction and viewing and combining of search results are now much more intuitive than before. ConQuest and the CSD database are also available on CD for local storage and retrieval on a PC.

The history of CSD data services at Nijmegen University is a long one and shows the contrast between the early days of structural data storage and retrieval in the early 1970s and the current situation. Many years before the CAOS/CAMM Center started its chemical information retrieval services, CSD searches for Dutch academic chemists were already provided on a voluntary basis by the author of this chapter. From the mid 1970s he mounted a tape copy of the Cambridge master file and the search software on the Nijmegen University computer. Queries, submitted on paper or by telephone, were entered and answers were communicated by surface mail.

With the advent of remote access possibilities and the foundation of the CAOS/CAMM Center, the Center became CSD's Dutch Affiliated Data Center and CSD searches were among the first on-line chemical information retrieval services offered by the new facility. By 1990 demand for this service had grown so much that a link to a dedicated high-speed search procedure on the Center's vectorized computer was built. In an on-line search session, where the user communicated with the Center's central VAX server, the data retrieval action was actually executed on a Convex system through a task-to-task communication protocol [57]. With this system even complex queries could be answered within 10 seconds – at that time a speed record. Now, less than fifteen years later, the same result is obtained by a chemist using his local PC and the viewing of results is much improved.

- **PDB** (**Protein Databank**). The PDB (www.rcsb.org) is the CSD's counterpart for biological macromolecular structures. Started in 1971, the database was developed at Brookhaven National Laboratory and was maintained at that site up to 1999. It stores the three-dimensional structure data of biological macromolecules. At first only protein structures were archived but, more recently, nucleic acid structures and structures of protein-nucleic acid complexes have also been deposited. In 1999 the file moved from Brookhaven to RCSB, the Research Collaboratory for Structural Bioinformatics. Under the new management PDB has developed from merely a repository of structure data into a tool for researchers to investigate the biological function of proteins. Several mirror sites provide easy access to PDB. Like CSD services, PDB access was also one of the initial tools in the Center's CAMM package.
- **ICSD** (**Inorganic Crystal Structure Database**). The ICSD (www.stn-international.de) contains records with cell parameters, space group information and atomic coordinates of all inorganic crystal structures published since 1915 – currently over 61,000 entries. The file is produced and maintained in an international collaboration between FIZ in Karlsruhe, Germany, and the National Institute of Standards in Gaithersburg, USA. A complementary database, CRYSTMET, contains similar data on metals and alloys. This database, originally developed at the National Research Council in Ottawa, Canada, is now maintained by Toth Information Systems (www.tothcanada.com). As is the case with CSD data, all ICSD and CRYSTMET data records are checked, evaluated, and recorded by experts, thus guaranteeing high data quality and reliability. New search software, with calculation options and graphical representation of structures in both Internet and PC versions has replaced the older keyboard-orien programs.

Access to both inorganic data files used to be provided in the Center's CAMM services package.

3.5. Computer-Assisted Molecular Modeling (*CAMM*)*: Research*

3.5.1. Research topics

To provide the Center's modeling services with high-level help desk support, several research projects have supported knowledge and expertise development. In the approximately 15 years of the Center's existence before the transition to the current CMBI, topics in the following fields were studied.

- The process of racemate resolution by crystallization of diastereomeric salts (see also Chapter 7 in this issue).
- Modeling software development, i.e. the MOLDEN program.
- Crystal structure prediction.
- Description of crystal symmetry.

3.5.2. Racemate resolution

This project started in 1988 and was concluded in 2001. In many papers [58–60] and two Ph.D. theses [61,62], the study of the fundamentals of the resolution process was reported. The first thesis (Leusen; *Rationalization of Racemate Resolution*) was completed in 1993. It could be shown that the solubility difference between two salts of a diastereomeric pair is related to the lattice enthalpy difference of the two salts. From that it follows that a computational model that accurately calculates lattice enthalpy differences can be used as a predictive tool for the resolution of optical isomers, an important industrial process. Various force field methods were used to calculate relative lattice energies of diastereomers of experimentally very-well-studied cyclic phosphoric acid-ephedrine complexes, and it had to be concluded that these methods could not reproduce experimental data. One of the main reasons for this failure was thought to be the inadequate description of the electrostatics in the complexes studied. In a follow-up study, concluding in a Ph.D. thesis completed in 2002 (Schaftenaar; *Computational Chemistry Methods*), an attempt was made to improve the force field calculations with a better electrostatic model. New molecular point charge models were developed, and distributed multipoles were used to model the electrostatic interactions. Some improvement was achieved, in particular in the calculation of the stability order of some complexes, but sensitivity of the calculations to scaling of the electrostatic contribution to the lattice energy and to conformational energy corrections of the rigid molecular ions made the overall results unsatisfactory. Complete abandonment of force field models and the use of quantum chemical density functional calculations was then tried. Depending on the sampling methodology, correct structures could be calculated and, in a number of cases, also a correct stability order for the diastereomeric salts in a pair. However, the final conclusion of this study had to be that current computational models for the very large molecular-ionic complexes studied are still insufficiently accurate to predict the outcome of a resolution.

3.5.3. *Modeling software development*

The program MOLDEN, a pre- and postprocessing program for molecular and electronic structures, under development since 1989, is a spin-off of the Center's CAMM research. Although different modeling programs were available to visualize the results of the calculations on the diastereomeric salts described earlier, during the research it was felt that display of electronic features in particular should be easier and more informative than that offered by the graphical interfaces of common computational chemistry programs like GaussView, SYBYL, or CERIUS. Where these programs feature sophisticated routines for the display of structural details, MOLDEN is unique in its coverage of features of the electronic structure of molecules. Some of these properties are processed directly from the output of computational chemistry programs like GAMESS, GAUSSIAN, and MOPAC, to which MOLDEN is tightly interfaced. Others are calculated by MOLDEN before display. Quantities visualized from the output of computational chemistry programs are, for example, geometry optimization results or saddle point location and related data. Many options for the display of structural features from entries in the PDB and CSD data files are also provided. Quantities not directly available from the output of other computational chemistry programs and calculated by MOLDEN are, for example, the electrostatic potential and a fit of point charges to reproduce the electrostatic potential, a distributed multipole analysis, orbitals and molecular density, or the difference density and the Laplacian of the electronic density.

MOLDEN has acquired a very large user base worldwide [63] and is used in the development of new chemistry teaching material at the Center (see also Section 4 and Chapter 8).

3.5.4. *Crystal structure prediction*

In the period 1995–2000 the CAOS/CAMM Center coordinated a Dutch national collaboration research program, CMS-c (Computational Materials Science-Crystallization). In this research program, in which the frontiers of computer modeling of crystallization processes were explored, crystal structure prediction research had an important place. The previously described racemate resolution study was one of the research projects in this program. Polymorph prediction of steroids and the prediction of structure and morphology of paraffin and fat crystals were two others.

The steroid studies focused on the prediction of the crystal packing in moderately flexible molecules. Steroids were chosen as model compound because of the available experimental data. Standard state-of-the-art software was used in most of the calculations (CERIUS; www.accelrys.com). With the introduction of novel methods to handle conformational flexibility and with new force field parameterization for rotatable groups, encouraging structure and polymorph prediction results on the model compounds could be obtained [64]. A review paper on crystal structure prediction methods was published in 1998 [65].

The paraffin and fat studies focused on the combined use of powder diffraction data, morphological data, and model building to predict structures and morphology of alkanes and triacylglycerols. It could be shown that discrepancies in structure, density, and stability between theory and experiment in long n-alkanes are influenced by thermal motion [66]. A possible polymorph of a triacylglycerol crystal structure could be predicted [67]. This research has been documented in a Ph.D. thesis which appeared in 2001 (van der Streek; *Molecular Modeling of the Crystal Structures of Long Chain Compounds*) [68]. Limits of crystal structure prediction software have been explored in some crystal structure prediction studies in which the CSD (see above and Chapter 6) was used as a reference [69,70].

3.5.5. Crystal symmetry environment

The regular shape of crystals is the result of lattice periodicity, an endless repetition of identical building blocks – the unit cell. This lattice periodicity, together with additional symmetries like rotation axes and mirror planes, is consistently described using group theory. All possible combinations of symmetry elements in three dimensions give rise to 230 space groups. These space groups are tabulated in the International Tables for Crystallography [71]. The discovery of new classes of crystal structures like incommensurately-modulated crystals and quasi-crystals that do not exhibit the lattice periodicity of ordinary structures have led to the development of a description in a space of greater than three dimensions. Symmetry elements such as 5-fold or 8-fold rotation axes, which do not exist in 3-dimensional crystallographic space, are legitimate in superspace and form superspace groups. The number of combinations of superspace group elements used to form superspace groups is much larger than for 3 dimensions and grows rapidly with superspace dimensionality. A descriptive summary such as the International Tables for the 230 3-dimensional space groups is not feasible. This gives rise to the need to visualize patterns with higher dimensional symmetry and to reveal relations between large numbers of space groups.

In 1991, in a collaboration between the Institute of Theoretical Physics of the University of Nijmegen and the CAOS/CAMM Center, a research project, CSE (Crystal Symmetry Environment), was started to develop a programming environment that would be the electronic equivalent of the International Tables for Crystallography. In this project the emphasis was on dimensions larger than 3, but the computer program was designed to include the lower-dimensional cases as well. In a first phase, an expandable prototype program environment using object-oriented programming techniques [72] was developed [73]. The information on crystal symmetry and derived structural properties in 3-dimensional space, as covered by the International Tables for Crystallography, was incorporated and made numerically, symbolically, and graphically presentable. Further development of this prototype program had to be stopped with the disappearance of cheminformatics research, caused by the transition of the CAOS/CAMM Center into the bioinformatics research group CMBI in 1999 (see Section 2.6). However, part of the software developed for calculations with tensors and characters appeared on CD-ROM as a supplement with a new edition of the International Tables for Crystallography in 2003.

3.6. Computer-Assisted MacroMolecular Sequence Analysis: (*CAMMSA*)

Bioinformatics is the biological counterpart of cheminformatics. It can be defined as the field in molecular biology dealing with nucleotide and amino acid sequences and the information they carry, including the computer methodology to handle this information. Bioinformatics evolved as a new molecular biology discipline in the beginning of the 1980s, together with the beginning of massive sequencing. The sequences themselves became a rich source of information through the application of various computational techniques. Comparison and analysis of sequence data in proteins and nucleic acids opened up the possibility of investigating the relationship between sequence, primary, secondary, and tertiary structure and function, and the evolutionary relations between proteins. Because bioinformatics is now a scientific discipline in its own right, its discussion is actually beyond the scope of this book, but with the sequence services belonging to the Center's initial service tools, a brief, mainly historic, discussion is justified.

When the CAOS/CAMM Center started its database services for chemists in 1985, it was clear that similar services would be even more needed by biochemists and molecular biologists. Where bench chemists could continue to synthesize molecules without referring to a computer, although much information would probably be overseen by doing so, a molecular biologist could never perform, for example, a sequence comparison by hand. With the Center's foundation, the hardware and organizational infrastructure for on-line database services was being set up and, from an information sciences point of view, storage and retrieval of reaction data is not much different from storage and retrieval of sequence data. Therefore the Center's management decided to also cover sequence databases, and when NWO, the Dutch science funding organization, honored a proposal to expand services in this direction, biochemists and molecular biologists became another target group for the Center's database servic

3.6.1. Sequence analysis of proteins and nucleic acids

Sequence analysis may serve different purposes:

- Identification by matching an unknown sequence with sequences stored in a database.
- Analysis to identify functionality, such as coding regions in a nucleic acid sequence.
- Analysis of functional domains and secondary structure determination in, for example, proteins.

The database content and the storage and retrieval software of the many sequence databases that have been developed during the past two decades are tailored to these different purposes. It is far beyond the scope of this book and this chapter to present a complete overview of the very many and often freely-accessible databases. Interested readers can consult, for example, the Web site of the EBI (the European

Bioinformatics Institute; www.ebi.ac.uk) or that of EMBnet (the European Molecular Biology network; www.embnet.org). Only a short introduction to some of the most frequently used sequence analysis tools that have been offered as CAOS/CAMM Center services will be presented here.

– **Sequence identification.** Databases with all known protein and nucleic acid sequences are among the most frequently used databases in molecular biology. The growth rate of these databases has been spectacular, mainly caused by the many genome sequencing projects in the last decade of the past century. The two main nucleic acid databases are the EMBL (now EBI) database, a European effort, and GenBank, a USA initiative. Close collaboration between both organizations has, from a user's point of view, resulted in more or less identical files. Spawned by the Human Genome Project, the GDB (Human Genome Database), a database with gene information and information on genetic diseases, was developed.
  A database search generally aims at finding sequences in the databases that significantly match the query sequence. Such a match might indicate an evolutionary relationship or, in the case of proteins, a relation between structure and function. Search algorithms take into account the evolutionary changes that have taken place over the centuries, such as insertions and deletions, and the statistical probabilities of accidental sequence similarity. Often a three-step procedure is used. After user specification of the required number of successive identical nucleotides in a query sequence and a database entry, step one produces a table with hits. In the second step scores for the best homologous pieces are calculated and finally the sequences with the highest scores are aligned. In an alignment procedure the homologous residues in different sequences are arranged in the best possible way.
– **Nucleic acid sequence analysis** tries to locate functional domains, i.e. domains with nucleotide sequences involved in protein coding. Nucleic acid research has revealed the repeating occurrence of substructural units and patterns which must be involved in protein coding and algorithms have been developed to identify these patterns statistically or on the basis of pattern recognition techniques.
– **Protein sequence analysis** generally aims at characterizing the relationship between the primary and the secondary and tertiary structure of proteins. By using sequence information and 3-dimensional structural information from X-ray crystallographic and, more recently, from NMR studies, more or less reliable structure prediction of proteins or protein fragments with known sequences and unknown 3D structure comes within reach. Although the algorithms for predicting, for example, alpha helices or beta sheets are still not successful in many cases, research efforts in this field certainly will improve the success rate within the next few years.

3.6.2. Services

All database-searching and sequence-analysis techniques described above, developed either as stand-alone programs or as part of academic or commercial software packages, have been implemented and offered as on-line service in the Center's bioinformatics tools package. Continuous updating of data files and software libraries were part of the Center's additional services and (re)formatting of data and user interfacing were also provided by the Center. Because of the simplicity of accessing and using all these tools, within a few years after its foundation the Center had developed into the main national hub for academic sequence data services and even for several commercial users. After the Center's transformation from a national facility into a university bioinformatics research group in 1999, some of the bioinformatics services have been continued, mainly as a local service. Most others have been transferred to a new national facility, BioASP (www.bioasp.nl) in Amsterdam.

3.6.3. EMBnet

For several years computer tapes, prepared by EMBL and GenBank were the media used to update the Center's local sequence databases. With the foundation of EMBnet in 1988, a European bioinformatics infrastructure for exchange of data and expertise was created and one of the first operational EMBnet services was database updating over the network. EMBnet could be established thanks to financial support of the European Commission. As a science-based collaboration of national nodes throughout Europe, the combined molecular biology and computer sciences expertise of the nodes allowed EMBnet to provide services which encompassed the possibilities of the individual nodes. Over the years EMBnet members and member nodes have participated in:

- projects to improve data and software accessibility for users;
- molecular biology and computer science research projects;
- upgrading of data distribution functionality;
- data networking;
- development of user training programs.

Through these projects EMBnet has been instrumental in the development of bioinformatics in Europe to such an extent that even several non-European groups have joined the collaboration. For further information see www.embnet.org and the links to the national nodes given there.

4. Educational services and chemistry teaching

Chemical research without computers is nowadays unthinkable. Therefore education and training of students to familiarize them with the newest programs available or being developed for their specific chemical discipline is imperative for any academic chemical curriculum. As a national service facility the CAOS/CAMM Center has focused attention on some aspects of this specific chemistry teaching right from the start. User training courses, as part of the normal use of the Center's databases and programs, have already been mentioned in Section 2.4. Two projects, aimed at the development of new educational material with an emphasis on the use of new computer methodology for teaching chemistry, originated from these courses in later years:

- Project SAMSAM (Students Access Molecular Structures And Modeling) where, in a collaboration with the Royal Dutch Chemical Society, KNCV (www.kncv.nl), access to parts of the Center's service databases and programs has been realized via the Center's Web menu (see Section 3.1) for graduate students.
- Projects WeTChe (Web Tutorials in Chemistry) and Mol4D (Molecules in 4 Dimensions), where course material that can be used entirely from within a Web-browser has been written for undergraduate students.

Chapter 8 is entirely devoted to these efforts. Here both initiatives will only be summarized.

SAMSAM has been developed as a student version of the regular CAOS/CAMM Center services. It differs from the latter by the limitation of the number of available modules and the extent of the contents. The target group used to be Dutch university chemistry students. SAMSAM content was offered at three levels: Basic, Plus, and Menu.

Basic offered freely-accessible Web pages with options for visualizing molecules. Plus, also freely-accessible, offered the same options with an extension to manipulate structures using a Web–browser plug-in. Explanatory text clarified the pictures. The Menu level was only accessible for registered users. In addition to the display of preselected structures and text, students could retrieve data from the Center's databases and use it for further investigation. Elaborate tutorials with manuals and examples completed the material presented. In the period 1996–2000, a number of Dutch universities incorporated one or more items from the SAMSAM program in their curricula and many chemistry students and teachers upgraded their cheminformatics knowledge through this program. With the discontinuation of the Center in 2000, further development under the SAMSAM umbrella came to an end but material developed before is still available (www.cmbi.kun.nl/samsam).

Mol4D has its origin in the earlier WeTChe project that aimed at the investigation of the feasibility of the usage of Web technology in chemistry teaching. In Mol4D exercises, a Web browser plug-in or Java applets are used for the animation of 3D structures in Web pages. Where SAMSAM aimed at graduate students, Mol4D was developed for undergraduates. For these students chemistry tuition in which computers are used should focus on teaching the chemistry and ideally combine maximum interactivity with minimal required computer-specific knowledge. Visualization and animation of structures and reactions are used to realize these goals. New educational material was provided mainly for topics from the undergraduate Organic Chemistry curriculum, such as cycloaddition reactions, molecular vibrations, and electrophilic aromatic substitution reactions, in which structural changes are animated in order to expand the student's in-depth knowledge of the chemistry. Through the use of scripts, a simple graphical structure editor, a computational chemistry program, and MOLDEN (see Section 3.5.3) in the background, interactive and dynamic Web pages with energy plots, orbital display, and electrostatic potential are generated and displayed in seconds. Mol4D exercises are freely accessible at www.cmbi.kun.nl/mol4d.

5. Summary

Founded in 1985, the CAOS/CAMM Center has, for a period of about 15 years, served as the Dutch national academic facility for on-line cheminformatics (and bioinformatics) services, supporting academic chemical research in the Netherlands. As such the Center has been instrumental in introducing and stimulating chemical data storage and retrieval methods and molecular modeling techniques in Dutch academic research and teaching. Many of the Center's research projects supporting the cheminformatics services have resulted in Ph.D. theses and tens of other papers. The development of the MOLDEN program, for example, resulted not only in a very-frequently-cited paper [63] but also in a worldwide implementation of the code with a user base of over 1300 installations at the moment.

Owing to a change in local and national funding policies followed by the transfer of the main bioinformatics services to a newly created facility (BioASP) in Amsterdam, at the end of the millennium the CAOS/CAMM Center was transformed into the CMBI, a local Nijmegen University bioinformatics research group, and cheminformatics research ended.

Acknowledgements

I am most grateful to the Center's initial staff members, whose efforts were essential in its development: Dr. H. Bruins Slot (modeling and crystallography); Dr. M.A. Ott (reaction retrieval and synthesis planning); Dr. J.A.M. Leunissen (bioinformatics); Dr. H. Borkent (chemistry teaching and training); and Dr. A. Thiers (system and network development). Of course, co-workers who joined later should not be forgotten: Dr. H. de Hilster (systems); Dr. F. Leusen (crystal structure prediction);

Dr. J.W. Boiten (synthesis planning); Dr. G. Schaftenaar (computational chemistry and modeling);

Dr. P. Verwer (crystallography); and Dr. C. van Gelder (teaching and SAMSAM project).

References

[1] J.H. Noordik, *CAOS/CAMM Services. An Integrated System of Computer Assisted Chemistry Tools*, in: Software Entwicklung in der Chemie 3. Proceedings des Workshops Computer in der Chemie, Springer Verlag, G. Gauglitz, ed., 1988.

[2] A.H.M. Thiers, J.A.M. Leunissen, T.M. Miller, G. Schaftenaar and J.H. Noordik, Computational Chemistry Network Services and User Interfacing, *J. Chem. Inf. Comput. Sci.* **33** (1993), 858–862.

[3] H. de Hilster, A.H.M. Thiers and J.H. Noordik, Computational Chemistry Network Services and User Interfacing, 2. *J. Chem. Inf. Comput. Sci.* **38** (1998), 775–779.

[4] A.H.M. Thiers and J.H. Noordik, Use of Vector Processing to search the Cambridge Structural Database, *J. Chem. Inf. Comput Sci.* **30** (1990), 19–22.

[5] Generally available network protocol, This action has some similarity with the selection of a WWW address.

[6] For an elaborate list of literature references see Martin A. Ott: Computer Methods in Synthetic Analysis. PhDthesis, Nijmegen University, 1996.

[7] M.A. Ott and J.H. Noordik, Computer tools for reaction retrieval and synthesis planning in organic chemistry. A brief review of their history, methods and programs, *Recl. Trav. Chim. Pays-Bas* **111** (1992), 239–246.

[8] J.W. Boiten, *Computer Methods in Organic Synthesis Design*, PhD thesis Nijmegen University.

[9] W.J. Wiswesser, 107 years of Line-Formula Notations 1861–1968, *J. Chem. Doc.* **8** (1968), 146–150.

[10] S.V. Kasparek, *Computer Graphics and Chemical Structures*, New York; John Wiley & Sons, 1990.

[11] E.J. Corey, W.T. Wipke, R.D. Cramer and W.J. Howe, Computer Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics, *J. Am. Chem. Soc.* **94** (1972), 421–430.

[12] H.L. Morgan, The Generation of a Unique Machine Description for Chemical Structures. A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.* **5** (1965), 107–113.

[13] H. Bebak, C. Buse, W.T. Donner, P. Hoever, H. Jacob, H. Klaus, J. Pesch, J. Romelt, P. Schilling, B. Woost and C. Zirc, The Standard Molecular Data Format(SMD Format) as an Integration Tool in Computer Chemistry, *J. Chem. Inf. Comput. Sci.* **29** (1989), 1–5.

[14] J.M. Barnard, Draft Specification for Revised Version of the Standard Molecular Data(SMD) Format, *J. Chem. Inf. Comput. Sci.* **30** (1990), 81–96.

[15] D.I. Cooke-Fox, G.H. Kirby, M.R. Lord and J.D. Rayner, Computer Translation of IUPAC Systematic Organic Chemical Nomenclature.4. Concise Connection Tables to Structure Diagrams, *J. Chem. Inf. Comput. Sci.* **30** (1990), 1222–1127.

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

[17] Brandt, A. von Scholley, An Efficient Algorithm for the Computation of the Canonical Numbering of Reaction Matrices, *Comput. & Chem.* **77** (1983), 51–59.

[18] D.J. Gluck, Chemical Structure Storage and Search System Developed at du Pont, *J. Chem. Doc.* **5** (1965), 43–51.

[19] S. Fujita, Canonical Numbering and Coding of Imaginary Transition Structures. A Novel Approach to the Linear Coding of Individual Organic Reactions, *J. Chem. Inf. Comput. Sci.* **28** (1988), 128–137.

[20] V. Kvasnicka and J. Pospichal, Canonical Indexing and Constructive Enumeration of Molecular Graphs, *J. Chem. Inf. Comput. Sci.* **30** (1990), 99–105.

[21] R.E. Valdes-Perez, A Canonical Representation of Multistep Reactions, *J. Chem. Inf. Comput. Sci.* **31** (1991), 554–556.

[22] M.G. Hicks, C. Jochum and H. Maier, Substructure Search Systems for large chemical databases, *Anal. Chim. Acta.* **235** (1990), 87–92.

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

[24] T.R. Hagadone, Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases, *J. Chem. Inf. Comput. Sci.* **322** (1992), 515–521.

[25] J.M. Barnard, Substructure Searching Methods: Old and New, *J. Chem. Inf. Comput. Sci.* **333** (1993), 532–538.

[26] R.P. Sheridan, R. Nilakantan, A. Rusinko, N. Bauman, K.S. Haraki and R. Venkataraghavan, 3DSEARCH; A System for Three Dimensional Substructure Searching, *J. Chem. Inf. Comput. Sci.* **29** (1989), 255–260.

[27] W. Fisanick, K.P. Cross, J.C. Forman and A. Rusinko, Experimental System for Similarity and 3D Searching of CAS Registry Substances.1. 3D Substructure Searching, *J. Chem. Inf. Comput. Sci.* **333** (1993), 549–559.

[28] M.G. Bures, E. Danaher, J. DeLazzer and Y.C. Martin, New Molecular Modeling Tools Using Three Dimensional Chemical Substructures, *J. Chem. Inf. Comput. Sci.* **34** (1994), 218–223.

[29] E.J. Corey and G.A. Petersson, An Algoritm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Organic Structures, *J. Am. Chem. Soc.* **94** (1972), 460–465.

[30] W.T. Wipke and T. Dyott, Use of Ring Assemblies in a Ring Perception Algorithm, *J. Chem. Inf. Comput. Sci.* **15** (1975), 140–114 4.

[31] B.T. Fan, A. Panaye, J.P. Doucet and A. Barbu, Ring Perception. A New Algorithm for Directly Finding the Smallest Set of Smallest Rings from a Connection Table, *J. Chem. Inf. Comput. Sci.* **333** (1993), 657–662.

[32] For an overview up to 1995 see ref. 8.

[33] E. Blake and R.C. Dana, CASREACT: More than a Million Reactions, *J. Chem. Inf. Comput. Sci.* **30** (1990), 394–399.

[34] C. Jochum, The Beilstein Information System Is Not a Reaction Database, or Is It? *J. Chem. Inf. Comput. Sci.* **34**, (1994), 71–773.

[35] D.F. Chodosh, *SYNthesis LIBrary Modern Approaches to Chemical Reaction Searching*, P. Willett, ed., Aldershot; Gower, 1986, pp. 118–145.

[36] D.F. Chodosh, J. Hill, L. Shpilsky and W.L. Mendelson, SYNthesis LIBrary, an expert system for chemical reaction knowledge-base management, *Recl. Trav. Chim. Pays-Bas* **111** (1992), 2477–2254.

[37] ORAC; Organic Reactions Accessed by Computer. MDL Information Systems Inc. 14600 Catalina Street, San Leandro CA 94577, USA.

[38] REACCS; Reaction ACCess System MDL Information Systems Inc. 14600 Catalina Street, San Leandro CA 94577, USA.

[39] I.K. Ugi, J. Bauer, R. Baumgartner, E. Fontain, D. Forstmeyer and S. Lohberger, Computer Assistance in the Design of Syntheses and a New Generation of Computer Programs for the Solution of Chemical Problems by Molecular Logic, *Pure & Appl. Chem.* **60** (1988), 1573–1586.

[40] E.J. Corey, General Methods for the Construction of Complex Molecules, *Pure & Appl. Chem.* **14** (1967), 19–37.

[41] I. Ugi, J. Bauer, C. Blomberger, J. Brandt, A. Dietz, E. Fontain, B. Gruber, A.v. Scholley-Pfab, A. Senff and N. Stein, Concepts, Theories, and Formal Languages in Chemistry and Their Use as a Bais for Computer Assistance in Chemistry, *J. Chem. Inf. Comput. Sci.* **34** (1994), 3–16.

[42] J. Gasteiger, W.D. Ihlenfeldt and P.A. Roese, A Collection of Computer Methods for Synthesis Design and Reaction Prediction, *Recl. Trac. Chim. Pays-Bas.* **111** (1992), 270–290.

[43] W.D. Ihlenfeldt, An efficient Approach toward a Flexible and General Knowledge Definition and Program Control Language System for a Synthesis Planning Program, *J. Chem. Inf. Comput. Sci.* **34** (1994), 872–880.

[44] J.B. Hendrickson and A.G. Toczko, Systematic Synthesis Design. The SYNGEN program, *Pure & Appl. Chem.* **61** (1989), 589–592.

[45] W.L. Jorgensen, E.R. Laird, A.J. Gushurst, J.M. Fleischer, S.A. Gothe, H.E. Helson, G.D. Paderes and S. Sinclair, CAMEO; a program for the logical prediction of the products of organic reactions, *Pure & Appl. Chem.* **62** (1990), 1921–1932.

[46] Z. Hippe, *Artificial Intelligence in Chemistry: Structure Elucidation and Simulation of Organic Reactions,* Amsterdam: Elsevier, 1991.

[47] J.W. Boiten, J.H. Noordik and M.B. Groen, Polyene Cyclizations. A Computer Assisted Synthesis Approach, *J. Chem. Inf. Comput. Sci.* **33** (1993), 727–735.

[48] J.H. Borkent, F. Oukes and J.H. Noordik, Chemical Reaction Searching Compared in REACCS, SYNLIB and ORAC, *J. Chem. Inf. Comput. Sci.* **28** (1988), 148–150.

[49] T.M. Miller, J.W. Boiten, M.A. Ott and J.H. Noordik, Organic Reactions Database Translation from REACCS to ORAC, *J. Chem. Inf. Comput. Sci.* **34** (1994), 653–660.

[50] J.W. Boiten, M.A. Ott and J.H. Noordik, Automated Overlap Analysis of Reaction Databases. Organic Reaction Database Translations from REACCS to ORAC, *J. Chem. Inf. Comput. Sci.* **34** (1995), 115–120.

[51] Martin A. Ott, *Computer Methods in Synthetic Analysis*, PhD thesis, Nijmegen University, 1996.

[52] Kenny B. Lipkowitz and Donald B. Boyd, *Reviews in Computational Chemistry*, 1st Volume. VCH Publishers Inc., 1990.

[53] Chemical Design Ltd. Purchased by Oxford Molecular, now part of Accelrys.

[54] MacroModel. Columbia University; now moved to Schrödinger Inc. (www.schrodinger.com).

[55] Japan Association for International Chemical Information (JAICI).

[56] Cambridge Crystallographic Data Centre. 12 Union Road, Cambridge CB2 1EZ, UK.

[57] A.H.M. Thiers and J.H. Noordik, Use of Vector Processing to search the Cambridge Structural Database, *J. Chem. Inf. Comput Sci.* **30** (1990), 19–22.

[58] F.J.J. Leusen, J.H. Noordik and H.R. Karfunkel, Racemate Resolution via Crystallization of Diastereomeric Salts: Thermodynamic Considerations and Molecular Mechanics Calculations, *Tetrahedron* **49** (1993), 5377–5396.

[59] F.J.J. Leusen, H.J. Bruins Slot, J.H. Noordik, A.D. van der Haest, H. Wijnberg and A. Bruggink, Towards a rational design of resolving agents, Part IV. Crystal packing analyses and molecular mechanics calculations for five pairs of diastereomeric salts of ephedrine and a cyclic phosphoric acid, *Recl. Trav. Chim. Pays-Bas* (*Jnl.R.Neth.Chem.Soc.*) **111** (1992), 111.

[60] F.J.J. Leusen, H.J. Bruins Slot and J.H. Noordik, Towards a rational design of resolving agents. Part III. Structural study of two pairs of diastereomeric salts of ephedrine and a cyclic phosphoric acid, *Recl. Trav. Chim. Pays Bas* (*Jnl.R.Neth.Chem.Soc.*) **109** (1990), 523.

[61] F.J.J. Leusen, *Rationalization of Racemate Resolution. A Molecular Modeling Study*, PhD thesis Nijmegen University, 1993.

[62] G. Schaftenaar, *Computational Chemistry Methods. Application to Racemate Resolution and Radical Cation Chemistry*, PhD thesis Nijmegen University, 2002.

[63] MOLDEN's acceptance and penetration is clearly demonstrated by the more than 1300 installations by the end of 2001 and by the ISI Essential Science Indicators' description of the paper describing the program as Hot Paper in Chemistry (no other paper in chemistry published in the last two years, aside from reviews, collected as many citations during the two month period November-December 2001).

[64] P. Verwer, in: *IUCR XVIII Congress and General Assembly Collected Abstracts*, C.C. Wilson, K. Shankland and T. Csoka, eds, 1999, 259.

[65] P. Verwer and F.J.J. Leusen, in: *Reviews in Computational Chemistry*, (Vol. 12), K.B. Lipkowitz and D.B. Boyd, eds, VCH Publishers, 1998, pp. 327–365.

[66] J. van der Streek, P. Verwer, P. Bennema and E. Vlieg, On The Influence of Thermal Motion on the Crystal Structuctures and Polymorphism of even n-alkanes, *Acta Cryst* **B58** (2002), 677–683.

[67] J. van der Streek, P. Verwer, R. de Gelder and F. Hollander, New evidence for beta'-p.p.+2.p Triacylglycerol Crystal Structure, *J. Am. Oil Chem. Soc.* **77** (2000), 215.

[68] J. van der Streek, *Molecular Modelling of the Crystal Structures of Long Chain Compounds*, PhD thesis Nijmegen University, 2001.

[69] P.M. Lommerse, W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Gavezotti, D.W.M. Hofmann, F.J.J. Leusen, W.T.M. Mooij, S.L. Price, B. Schweizer, M.U. Schmidt, B.P. van Eijck, P. Verwer and D.E. Williams, A test of crystal structure prediction of small organic molecules, *Acta Cryst* **B56** (2000), 697–714.

[70] W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezotti, D.W.M. Hofmann, F.J.J. Leusen, J.P.M. Lommerse, W.T.M. Mooij, S.L. Price, H. Scheraga, B. Schweizer, M.U. Schmidt, B.P. van Eijck, P. Verwer and D.E. Williams, Crystal structure prediction of small organic molecules? a second blind test, *Acta Cryst* **B58** (2002), 647–661.

[71] Th. Hahn, ed., *International Tables for Crystallography*, Vol. A, Space-group Symmetry, Reidel, Dordrecht, 1983.

[72] M.J. Ephraim, A.H.M. Thiers, A. Janner and T. Janssen, *Applying Smalltalk-80 and C++ to Crystal Symmetry Analysis Structured Programming* **14** (1993), 119–135.

[73] A.H.M. Thiers, M.J. Ephraim, T. Janssen and A. Janner, An object oriented approach towards a crystal symmetry environment, *Computer Physics Communications* **77** (1993), 167–189.