Chemical Informatics

Introduction of Computers in Chemical Structure Information Systems,
or What Is Not Recorded in the Annals

Michael F. Lynch

Abstract

This paper describes the course of research on the classical elements of representation, storage and retrieval of chemical structures and chemical reactions from the viewpoint of a research group much involved in chemical information science for over three decades. It is weighted heavily towards the 1960s and 1970s, when the chemical information scene as we know it today began to take shape. It concentrates on the science and on the people who originated it rather than on organizations and information services. The research also led to a reinterpretation of Claude Shannon's information theory and to a means of compressing data, which in turn was generalized as the Ziv-Lempel technique, now widely used.

First, I wish to thank the organizers of this conference, Mary Ellen Bowden and W. Boyd Rayward, for their very kind invitation to me to participate. The topic of the conference is very dear to my heart. I was privileged to know and work with many of the personalities of the time, whose imaginative initiatives transformed chemical information handling in so many ways. The University of Sheffield played an important role, then as now. We offered the first course in this area in 1968, published the first textbook (Lynch, Harrison, Town, & Ash, 1971), and started one of the first master's courses in chemoinformatics (Schofield, Willett, & Wiggins, 2001). I want to thank Chemical Abstracts Service (CAS), which employed me and enabled me to learn much of the business of chemical information at close quarters. I was a young lad, still wet behind the ears as far as any knowledge of chemical information and computer sciences was concerned. CAS allowed me to make my mistakes at their expense and to gain experience in the chemoinformatics area. I am also grateful to America for having given me the opportunity to learn the "can-do" culture.

There is a delightful book published by the German Chemical Society titled *Was nicht in den Annalen steht,* or *What is not recorded in the Annals* (Hausen, 1969). It is an informal account of what chemists got up to in earlier years and includes an anecdote about Robert Bunsen, discoverer of strontium, who was a bachelor and ate at the same restaurant each day. He came to suspect that the cook was recycling the bones of his chicken dinner into the next day's soup, and so he added some strontium salt to the leftover bones and tested his soup next day with a platinum wire and flame for the presence of the characteristic crimson color—and confirmed his fears. Another tale relates to the fact that German glassware was better than French at one time, but a high import duty was charged at the French border. One French chemist overcame this by having corks inserted and the flasks labeled "German air—for scientific research." No duty was payable on air!

This paper is intended to be a bit like that book. It is weighted heavily towards the decades of the 1960s and 1970s, when the chemical information scene as we know it today began to take shape. I concentrate mainly on the "classical" elements of the storage and retrieval of chemical structures and chemical reactions, with only a sidewards glance at some other innovative areas. I concentrate on the science and the people who originated it and on strategic contributions, rather than describing a broader front without detail, which would reduce to a mere timeline. I hope that Eugene Garfield will not mind

if I paraphrase the title of his earlier paper with an Irishism—"Here we sit side by side with those on whose shoulders we stand." The line through this paper is the line the research itself took, as one idea followed another. There was always an interplay between research in the chemical structure area and the textual areas, an instance of a warp and weft being formed by the two. This interplay has always been a strong element of the Sheffield work, both in my group and in that of Peter Willett.

There are many admirable reviews of research and practice in this area for this time period, as well as a wide range of excellent textbooks (Tate, 1967; Lynch, 1968; Lynch et al., 1971; Davis & Rush, 1974; Wipke, Heller, Feldmann, & Hyde, 1974; Ash & Hyde, 1975; Rush, 1976; Wipke & Howe, 1977; Rush, 1985; Ash, Chubb, Ward, Welford, & Willett, 1985; Willett, 1986; Allen & Lynch, 1989; Lipscomb, Lynch, & Willett, 1990; Bawden & Mitchell, 1990; Ash, Warr, & Willett, 1991; Warr & Suhr, 1992).

A series of conferences on chemical structures and computers at Noordwijkerhout in the Netherlands, the first in 1973 funded by NATO (North Atlantic Treaty Organization) and organized by the Chemical Notation Association, was instrumental in bringing together a disparate group of researchers for high-caliber discussions. (The most recent of these conferences was in 2002.) Subsequent meetings were organized by an international consortium of scientific societies led by the Chemical Structure Association. The International Conferences on Computers in Chemical Research and Education (ICCCRE) from 1971 onward also contributed strongly to international communication.

First, I will relate the background to the issues as they stood in 1961 when I joined CAS as a member of the Research Department with G. Malcolm Dyson. There were great stirrings in science information at that time because of *Sputnik,* the challenge to the United States from the Soviet Union in October 1957. *Sputnik*'s beep-beep tones took the world totally by surprise. When the dust had settled, it became apparent that the Soviets had published their intentions in the open literature, but the science information system in the West was in disarray. The system had not been considered sufficiently important nor was it well enough funded to keep up with the vast increases in the numbers of scientists employed and publishing in the postwar period. There was said to be a cocktail called Sputnik, one part vodka and three parts sour grapes.

Coverage of the sources of chemical information was thus incomplete and unsystematic. By 1960 the indexes to *Chemical Abstracts* (*CA*) were two to three years late in production. The cumulative indexes appeared at intervals of ten years. In addition, access to chemical structures was highly problematic. Chemists need to determine whether a specific substance is known and to find its synthesis or other provenance, its reactions, and its properties. Further, chemists need to interrelate substances and their properties by means of their common substructures. Discussions in organic chemistry invoke myriad examples of structural relationships, for example, in such generic expressions as "five-membered nitrogen heterocycles," or "p,p′-dihalodiphenyl ketones." Thus, Paul Craig of Smith Kline and French reckoned in 1962 that only 70 percent of all known phenothiazines could be found under that heading in the *CA* subject indexes, since many other chemical features can take precedence over the phenothiazine ring system in the naming process. The other 30 percent of entries on phenothiazines were scattered through the alphabet. Robert Fugmann, speaking at this conference, is a prime exponent of means of expressing these relations and a designer of the famous GREMAS (Generische Recherchieren Mittels Magnetband [Generic searching by means of magnetic tape]) system, which dates back to the 1950s (Fugmann, Braun, & Vaupel, 1963).

Suddenly, after *Sputnik,* funding for information science and technology research became readily available on both sides of the Atlantic. In the context of chemical information the role of CAS in mastering access to chemical information was pivotal. Malcolm Dyson had become director of research there in the late 1950s. One of Dyson's many early innovations at CAS was to apply Hans Peter Luhn's KWIC (keyword-in-context) index concept to provide a stopgap—*Chemical Titles*—between the appearance of the abstracts in *CA* and the availability of the full subject and formula indexes (Dyson, 1961). It was the CAS response to *Current Abstracts of Chemistry—Index Chemicus* that Gene Garfield's Institute for Scientific Information (ISI) had introduced in 1960 at the urgent request of the pharmaceutical industry.

In 1946 Dyson began to develop a chemical cipher or notation system, the purpose of which was to simplify and overcome the many inadequacies and inconsistencies of traditional chemical nomenclature. A form of the Dyson cypher was adopted by the International Union of Pure and Applied Chemistry (IUPAC) as the

international standard in 1961, Dyson himself having been a member of the appropriate IUPAC committee (International Union of Pure and Applied Chemistry, 1961). Dyson was exciting to work with. Despite having been gassed in the trenches in World War I so that his health was never good, his commitment was total. Our best work was done with him at lunches in the Jai Lai restaurant on Olentangy River Road, using the paper mats as work sheets.

The CAS Registry System: Chemical Line Notation Considerations

I was first hired to develop a novel computer-produced abstracting publication keyed particularly to the needs of the pharmaceutical industry—*Chemical–Biological Activities*, or *CBAC*—along lines mapped out by Dyson. It was aimed at the same industry sector as *Current Abstracts of Chemistry—Index Chemicus* and was to be the first product supported by the CAS Chemical Structure Registry System. Hence I was drawn into the design of the registry system and, central to that, to the representation to be used for this single-entry file for chemical substances. In Dyson's view the representation was to be the IUPAC notation. Each substance on entry into the registry system would receive a registry number, and the substance would be known thereafter by its name, notation, and registry number.

In 1963 I produced a few sample pages of *CBAC* and showed it at an ACS Division of Chemical Literature Meeting in Cincinnati, Ohio. Aaron Addelston, a chemist whose passion was accuracy in chemical information sources, examined a copy very closely and discovered many errors I had made in the description of substances in the IUPAC notation (Addelston & Goldsmith, 1966). These were mostly errors in applying the rules for ensuring a unique notation rather than erroneous description of the substance. Two themes are linked in this experience: first, the issue of the choice of representation for chemical structures and, second, the automation of publications at CAS.

Dyson's views were dominated by his advocacy of the IUPAC notation. His interests, particularly regarding structures, tended to be retrospective, despite his being technically highly innovative. Fred Tate was appointed assistant director and acting editor of CAS in 1961. His outstanding intellect enabled him to provide CAS with singular leadership in those years. With only the example of the automation of publication of *Index Medicus* by Charles Austin (1964) (under the overall leadership of Brad Rogers) as a model, Tate largely determined the future shape of the systems we know today. He realized that the only way to provide up-to-date information was to automate the production process and in time to deal with the backlog of existing information. The most important information, current information, was then available for use with whatever retrieval methods might be available, while the backlog diminished in importance over time. The automation of the publication process, and with it the development of the registry system, was the most extensive undertaking CAS has seen then or since. Sadly, Tate did not live to see the completion of his ambitious schemes, which were driven through by Dale Baker.

Another factor was rightly prominent in Tate's thinking—the complexity of producing the abstracts and index entries and most particularly the costs of naming the myriad chemical substances included. A name had to be generated for each substance included in the indexes. Given the complexities of the CAS nomenclature system, this process called for the employment of many Ph.D.-level chemists. Each substance included had to be named, whether or not it had already appeared in the CAS system: there was no cost-effective way to discover if it had already been named. Any measures that would reduce the high cost and speed up the process would be invaluable. The registry system promised such improvement: once a substance name, if available, was keyboarded, the system could indicate whether the substance had already been named according to CAS and retrieve that name for inspection. Otherwise, it would indicate that the substance was novel and send it to be input as a structure and named. This measure proved a substantial cost-reduction and a contribution to reducing time delays (Leiter, Morgan, & Stobaugh, 1965; Dittmar, Stobaugh, & Watson, 1976). Leslie Blankenship and Val Metanomski (2002) and Chuck Davis (2004), speaking at this conference, tell us more about this aspect of the registry system.

Fred Tate and Malcolm Dyson also had an unusual kind of confidence in the future: although hardware and software were limited at that time, their power to tackle the challenges facing them would grow. It reminds me of a situation in the middle of the nineteenth century. In Europe the major continental powers, including France and Germany, were busily building railroads across their territories. Strategic military power was of course the major factor. The old Austrian Empire had a problem in that its Mediterranean fleet was based at

Trieste on the Adriatic. In 1848 Kaiser Franz Josef and his ministers began to build a railroad south over the Alps, across the Semmering Pass, that involved gradients greater than anything the locomotives of the day could master. They did this in the supreme confidence that, when the route was finished, the locomotives would be sufficiently powerful. Six years later it turned out to be so. So too with Tate and Dyson.

It is worth reviewing what resources were available to these pioneers (Rush, 1985). The only item of data-processing equipment that CAS had in late 1961 was a punched-card reader that drove an electric typewriter—just so the IUPAC notations, which required upper- and lower-case letters and sub- and superscript numerals—could be printed. When CAS acquired a computer, it was an IBM 1401 with an internal memory of 8K of six-bit bytes.

Many arguments focused on the relative merits of the IUPAC and Wiswesser notations. Both were useful means of inputting run-of-the-mill two-dimensional chemical structures at a time when computer graphics capabilities were very low. This scarcely changed until the Apple II personal computer came on the scene. Neither IUPAC nor Wiswesser was free from definitional problems. Their designers were ignorant of the rigor of graph theory or other such artificial language definition schemes as context-free grammars. The relevance of notations declined as graphics input and display methods improved, although SMILES (Simplified Molecular Input Line Entry System), introduced by D. Weininger (1988), has achieved a measure of popularity.

The CAS Registry System: Connection Table Considerations

To return to the matter of a representation for chemical structures: despite my strong loyalty to Malcolm Dyson and his efforts on behalf of the IUPAC notation, I could no longer support the idea of its use in production if it was so error prone. After the debacle with the sample issue of *CBAC* at the ACS meeting, I had investigated methods of translation from IUPAC notations into connection tables—with the automatic generation of canonical notations in mind. This was not far removed from Garfield's seminal doctoral work on generation of molecular formulas from chemical names, which opened up novel possibilities (1961, 1962). Ed Cossum (computer manager in the R&D department), Harry Morgan (programmer and mathematician), and I began to explore other options. With Dyson we looked at the random matrix, that is, connection tables, for substructure search (Dyson, Cossum, Lynch, & Morgan, 1963a, 1963b, 1963c; Cossum, Krakiwsky, & Lynch, 1965). G. W. Wheland had been the first to show how this could be done (1949); Calvin Mooers (1951) and Hans Peter Luhn (1955) later suggested much the same thing. L. C. Ray and R. A. Kirsch (1957) demonstrated an iterative substructure search using connection tables to represent chemical structures. Ernst Meyer at BASF had also introduced a form of connection table, which to our eyes at the time seemed highly redundant and space consuming (Meyer & Wenke, 1962). Jacques-Emile Dubois had also been working with connection tables at the University of Paris in the 1950s, which led to Système DARC (Documentation et Automatisation des Recherches de Correlations [Documentation and Automated Research of Correlations]) (Dubois & Viellard, 1968a, 1968b, 1968c).

Around that time Ken Zabriskie from DuPont replaced Dyson as research director. Soon after that Zabriskie and I traveled to Wilmington, Delaware, to receive an offer from DuPont of an algorithm developed by Dave Gluck (1965) to generate a canonical form of connection table. This was important, as a canonical or unique form gives an easy way of identifying whether or not a structure is already known. We investigated this, but a counterexample raised by Sylvan Eisman blew it out of the water: he showed that you could generate two different representations for the same molecule, depending on the initial numbering.

A conference on chemical structure representations was held at Airlie House in Virginia in March 1964. It was memorable in that Malcolm Dyson and Bill Wiswesser, who were usually at loggerheads with one another, stood together for the first time, defending chemical notations against the onslaught of the connection table. I was party to a conversation in which Harry Morgan and Calvin Mooers discussed the issue of a canonical connection table. Mooers, in the light of the counterexample raised against the work of Dave Gluck, suggested a permutational approach to the design of an algorithm, which would overcome the objections. Harry Morgan (1965) then devised the method that bears his name and that has been a keystone in chemoinformatics since that time.

A little more needs to be said about line notations. Thus far I have dealt only with the IUPAC notation. The method of choice in the United States and the United Kingdom at that time and later was the Wiswesser

notation, not least because it used only the symbols of the standard line printer (Smith, 1968). Wiswesser notation was widely used in industry as well as at ISI, where it was to form the basis for providing structural information to subscribers to *Current Abstracts of Chemistry—Index Chemicus,* in part through the use of the HAIC (hetero-atom-in-context) index (Davis & Rush, 1974).

Extensions to Structure Representations

A little later Jim Rush, Tony Petrarca, and I developed methods of extending the Morgan algorithm to handle single tetrahedral atom stereochemistry, which amounted basically to manipulating a Fischer projection of a molecule (Petrarca, Lynch, & Rush, 1965). The methods were later extended to other stereochemistry situations by A. E. Petrarca and J. E. Rush (1969) and even later implemented by W. T. Wipke and T. M. Dyott (1974) as SEMA (stereochemically extended Morgan algorithm).

Another meeting at about the same time was highly significant for me. Its purpose was to showcase the joint work of the National Bureau of Standards and the U.S. Patent and Trademarks Office on storing and retrieving generic chemical structures—"Markush" structures. These include radicals that may include infinite sets. The researchers laid out their ambitious plans, only to have Julian Bigelow, a Princeton math professor, comment, "infinite series, finite search algorithm—no solution." This restrained me from tackling that particular problem for many years until I was able to interest E. J. Krishnamurthy of Bangalore, India, to advise on possible solutions, including just those context-free grammars for language definition that the designers of early line notations had lacked.

A development that was to have the most profound influence on better understanding the role and function of three-dimensional molecules and particularly of proteins was the founding in 1965 of the Cambridge Crystallographic Data Centre by Olga Kennard at Cambridge University. This center played a leading role in storing three-dimensional crystallographic data on structures and in developing methods for three-dimensional search, in part in collaboration with Peter Willett. By recording crystallographic data from journals and other sources, full stereochemical data are provided and three-dimensional and topological values generated (Kennard, Watson, & Town, 1972; Allen et al., 1979; Allen & Lynch, 1989).

Also in 1965 Joshua Lederberg and Ed Feigenbaum introduced DENDRAL (DENDRitic ALgorithm) to interpret the mass spectra of individual molecules. This imaginative approach was part of the birth of the artificial intelligence movement, much of which, sadly, was to be grossly overhyped (Lederberg, 1990; Gray, 1986).

Input and Display

At the earliest stages of the CAS Registry System connection tables were input by hand. Ascher Opler and Norma Baird's use of a light pen for graphical entry of chemical structures on a visual display unit had already shown the way ahead (1959). Chemical typewriters were soon on the scene (Feldman, Holland, & Jacobus, 1963; Mullen, 1966), followed eventually by graphics programs. Ernst Meyer (1965) had earlier scanned input of drawn structures on a special machine.

Another valuable means of input again reflected Garfield's early initiative with name analysis. Thus, in 1965 Gerald Vander Stouw and I put in hand the development of automatic generation of connection tables from CAS chemical names, which today still assists in structure input and validation and was indispensable in scanning in *CA* formula index entries from earlier years and translating them into connection tables (Vander Stouw, Naznitsky, & Rush, 1967; Vander Stouw, Elliott, & Isenberg, 1974).

Few contributed more to graphics display than Ernest Hyde and Lucille Thomson (1968) in their design of the CROSSBOW System at ICI Pharmaceuticals, later to be further developed at CAS (Dittmar, Mockus, & Couvreur, 1977). In addition to the great energy Hyde put into his work, he and his wife, Barbara, found time to run a home for the care of the physically and mentally handicapped. Later still William Town's company, Hampden Data Services (HDS) (now managed by Peter Nichols), excelled in the characteristics of its PC-based structure input system, PSIDOM, which was later used in the Scientific and Technical Network (STN).

Chemical Reactions

Chemical reactions are no less important than the structures themselves. Early approaches used names of the discoverers of reactions, such as Claisen condensation and Hoesch reaction. Ernst Meyer (1965) described methods for isolating the changes involved in reactions, linked to his method of inputting graphic structures. Another very early contribution was the doctoral thesis by Uwe Pape of the University of Braunschweig, who examined the possibilities of interaction of simple

molecules by arbitrary breaking and remaking of bonds. While this was done on a formalistic basis only, it was a remarkable early piece of work, carried out, moreover, on a German Zuse Z22 machine, which was one of the series that first operated in 1941 (Pape, 1968).

In 1963 George Vladutz, then working at VINITI (All-Union Institute of Scientific and Technical Information) in Moscow, had published a seminal paper in which he suggested approaches to the automatic identification of "skeletal reaction schemes" in general and symbolic terms in the records of reactants and products in organic reactions. Vladutz later explained to me the reason for the symbolic treatment: theoretical research was the norm at VINITI, and actual computer resources to demonstrate practicality were available only for the few months in the run up to an international conference in the U.S.S.R. The paper triggered my interest, and I began to look at the possibilities. The first piece of independent work that I did at Sheffield University dealt with the *CA* subject indexes. One aspect of this suggested a method that might be applied to isolating the skeletal reaction schemes. A series of papers was published, the first titled "Automatic Detection of Structural Similarities among Chemical Compounds," but we made little real progress (Armitage & Lynch, 1967a; Armitage, Crowe, Evans, Lynch, & McGuirk, 1967; Harrison & Lynch, 1970; Clinging & Lynch, 1973, 1974; Lynch, Nunn, & Radcliffe, 1978).

Vladutz and I had corresponded indirectly in the meantime, and my first research associate in Sheffield, Janet Armitage (Ash), had visited him at VINITI in Moscow. He informed me in 1975 that he had obtained an exit visa from the Soviet Union and was already in Italy. Later he told me that he had applied for an exit visa for himself and his family when Henry Kissinger first visited Moscow and was granted it on the occasion of Kissinger's second visit. In between, they were treated as pariahs by the Soviet authorities. I was able to secure a research fellowship from the British Library Research and Development Department to enable him to join me at Sheffield for nine months before he joined the staff of ISI in Philadelphia. Vladutz, who grew up in a Hungarian-speaking enclave of Romania, was the epitome of the completely cultured European, at home in perhaps ten languages, profoundly knowledgeable about art and music, and a delight to be with. He also claimed descent from Vlad the Impaler, otherwise known as Count Dracula.

That year Peter Willett was a student in our master's program. He expressed an interest in looking at the question of reaction analysis again. If it had not been for George's presence, I would have discouraged him from considering the idea, but out of this work came the algorithm that was the most notable achievement in Willett's Ph.D. thesis and that is widely used as the basis for reaction analysis techniques today (Vladutz, 1977; Willett, 1978; Lynch & Willett, 1978; McGregor & Willett, 1981). The work was based on Wiswesser line notation records of substances involved in reactions identified in *Current Abstracts of Chemistry—Index Chemicus*. The database was compiled by pairing off the Wiswesser line notations of reactant and product molecules and translating them into connection tables.

Screens for Substructure Search

Even if the early approaches to analyzing chemical reactions were fruitless, they did suggest a possible solution to an immediate and outstanding issue, that of searching databases of chemical structures for embedded substructures. When L. C. Ray and R. A. Kirsch tested the iterative trial-and-error search mechanism, they made it clear that screens to reduce the extent of atom-by-atom searching—a hugely expensive process—were absolutely essential to ensure that structures that could not possibly meet the criteria for the query would be excluded.

I will introduce the Sheffield strategy to devising screens by quoting Claude Shannon's *Encyclopaedia Britannica* article (1971). He mentioned the game of Twenty Questions and posited the view that for a million items, one could achieve almost unique selection by having twenty binary search features that were equifrequent and not statistically associated (Shannon, 1948). However, the problem with chemical structures is that the distribution of features is highly uneven—a low-entropy state, in Shannon's terms. Thus, the difference in frequencies of atoms in molecules over the first ten atom types is 1,000 to 1. Iodine atoms occur a thousand times less frequently than carbon atoms, hence a screen for an iodine atom would be very powerful. But how frequently would a substructure query require an iodine atom? Once in a thousand times?

Once again my work on *CA* subject indexes provided an unusual insight. The subject index entries then consisted of a heading and one or more entries. I noticed that the entries were longer when there were many entries under a particular heading and shorter when there were few entries. This applied even when entries related to the same document: the form of the entry was context sensitive. In other words, in order to discriminate better between entries when there was a

large population, the indexers increased the specificity of the subject description (Lynch, 1966; Armitage & Lynch, 1967b, 1968).

This was a time when there was a great deal of discussion of the role of information theory in indexing. Thus, Calvin Mooers's Zatocoding system used superimposed punches in edge-notched cards (1947). The number of punches was inversely related to the frequency with which a term was used. For the most frequent headings there was a unique punch, while for the less frequent headings several randomly assigned punch positions were allocated. This interpretation of information theory was conventional and akin to Morse code, in which the most frequent characters, such as E, T, A, I, O, and N, are allocated short codes of dits and dahs. Data-compression algorithms such as that of David Huffman take the same line (1952), as Ron Kline has indicated at this conference (2004). But in the *CA* subject indexes we were seeing a quite different approach: what is frequent is described in detail, and what is infrequent is described in outline.

We argued that using screens in a serial search system required an optimization of the resolving power of each screen; the screens should ideally be equifrequent, that is, the ideal was to transform a low-entropy, highly hyperbolic distribution into a high-entropy rectangular distribution. This approach allowed us to develop screens for substructure searching. Using a random sample of substances from the registry system kindly provided by CAS, we were able to analyze the frequencies with which different features occurred. We were also able to demonstrate that substructure queries posed as text queries in current awareness searches were quite similar in distributional characteristics from those we identified in the sample database. We identified atom-centered and bond-centered fragments (coining such expressions as *augmented atoms, coordinated atoms,* and *bonded atoms,* and the same for *bond-centered fragments*) and developed methods to determine at what level of detail particular fragments needed to be indicated in order to achieve roughly equivalent search value for each screen. Alfred Feldman and Louis Hodes later extended this by making the analysis and fragment selection processes fully automatic (Lynch, Orton, & Town, 1969; Crowe, Lynch, & Town, 1970, and subsequent papers; Adamson, Cowell, Lynch, McLure, Town, & Yapp, 1973; Adamson, Clinch, & Lynch, 1973; Adamson, Bush, McLure, & Lynch, 1974; Lynch, 1974; Feldman & Hodes, 1975; Lynch, 1977).

On this basis we developed a scheme or dictionary of structural features appropriate for searches of serial files; evaluation showed their power in serial searching. At the first international conference held at Noordwijkerhout in 1973, I talked to Heinz Kaindl, who represented the cooperative chemical information service for BASIC, the Swiss-based pharmaceutical industry, namely, CIBA-Geigy, Hoffman–La Roche, and Sandoz. BASIC was already subscribing to the CAS Registry System files, and it needed a better substructure search system than that which CAS had provided. BASIC implemented a system based largely on our screens, and it worked to BASIC's satisfaction (Schenck & Wegmüller, 1976; Graf, Kaindl, Kniess, Schmidt, & Warszawski, 1982).

In France, Jacques-Emile Dubois had already been working on file structures for substructure search on direct access rather than serial systems—the DARC system, using the notion of the FREL (**Fr**agment of an **E**nvironment **L**imited in X) (Dubois & Viellard, 1968a, 1968b, 1968c; Dubois, 1973, 1974; Attias, 1983). This was an astonishing accomplishment: the DARC system operated on a single mainframe rather than on an array of minicomputers as used later by CAS. As part of international arrangements between countries, CAS was providing registry system files to various countries. The French mounted a system that rapidly became successful with industry around the globe. Shortly thereafter, CAS was placed under pressure by industry to mount a public substructure search system. The BASIC group in Basel offered CAS its system, and with some extensions—such as sequences of connectivity values as further search screens—CAS implemented a system based largely on the Sheffield work (Dittmar, Stobaugh, & Watson, 1976).

As an aside I will mention that we later extended this principle to examine screens for use in text searching. Similar hyperbolic distributions occur in text, and we began to study these in terms of the frequencies of strings of characters—for example, "-ation_of_-", which is more frequent in English text than the rarest characters like Q and X—first as screens in serial searching of text databases. My research colleague Howard Petrie observed that one could use the method for a quite different application: text compression. We were able to achieve a reduction of 50 percent in storage, in work that was language dependent—that is, a code book was needed for each different language. At a conference of the Association of Computing Machinery's Special Interest Group in Information Retrieval and the British Computer Society in Cambridge in the early 1970s, I described our work to an Israeli information scientist.

The principle was later generalized and made adaptive by the Israeli scientists Jacob Ziv and Abraham A. Lempel as the Ziv-Lempel algorithm, and it is also known as PKZIP or WINZIP on your PC. Thus, we brought about a major paradigm change in the way information theory applications in data compression are viewed (Clare, Cook, & Lynch, 1972; Barton, Lynch, Petrie, & Snell, 1974; Lynch, 1977; Ziv & Lempel, 1977).

Chemical Reaction Design

A development that greatly engaged the imagination of chemists was introduced by Elias Corey and Todd Wipke (1969). The OCSS-LHASA (Organic Chemical Synthesis Simulation—Logic and Heuristics) system was designed to suggest starting materials and reactions to produce a given target molecule. In Wipke's SECS (Simulation and Evaluation of Chemical Synthesis) system this process involved many highly accomplished industrial chemists in encoding the reaction types in which they had particular skills. With Stuart Marson and Steve Peacock, Wipke founded Molecular Design Limited in 1977, whose MACCS (Molecular Access System) product was the first to have a major impact on industry, followed by products from a range of other companies. Other contributors to this area included Jim Hendrickson (1971), Peter Johnson at Leeds University (Johnson & Cook, 1985), and Ivar Ugi and collaborators, especially Johnny Gasteiger, now at the University of Erlangen-Nuremberg (Ugi et al., 1979). Ugi and his associates introduced a very different formalism, following in the path of J. Dugundji's bond-electron formalism, which owed little to chemists' experience (Dugundji & Ugi, 1973). This important sphere warrants a history to itself.

Conclusion

I can think of no better way of rounding out this paper and providing an appraisal and forward-looking view of the technology to which the early work I have described here has led than by providing a summary of the chemoinformatics elements of the master's in chemoinformatics program at the Department of Information Studies at Sheffield, recently introduced by Peter Willett. This program is one of three, the others being at the University of Manchester Institute of Science and Technology (Helen Schofield) and at Indiana University (Gary Wiggins) (Schofield, Willett, & Wiggins, 2001).

The Sheffield University course, which is funded by the U.K. Engineering and Physical Sciences Research Council and supported by collaboration with industry, aims to provide a broad understanding of the computational techniques available for processing databases of chemical and biological structural information. These techniques include representing and searching two-dimensional and three-dimensional chemical structures, similarity searching, chemical patent searching, structure-activity relationships, combinatorial library design and molecular diversity, structure-based drug design, and representing and searching biological data. The course also provides practical experience in using a variety of commercially available systems, so that by the end of the module students will have learned the techniques available for representing and searching databases of chemical structures, a critical awareness of the techniques used to design novel bioactive compounds in the pharmaceutical and agrochemical industries, the techniques available for representing and searching biological sequences and structures, and how to apply a range of different software tools to support the discovery of novel bioactive compounds.

Acknowledgments

I thank the Ministry of Education of the Land North-Rhine Westphalia, Germany, for a studentship to study for a year at the Technische Hochschule in Aachen in 1951–52; the Royal Commission for the Exhibition of 1851 for a Commonwealth Postdoctoral Fellowship to study with Vladimir Prelog at the Eidgenössische Technische Hochschule, Zurich, 1957–59; Wilfrid L. Saunders for welcoming me to the Postgraduate School of Librarianship in 1965 and for constant support and wise counsel; Peter Willett for permission to use extracts from the description of the master's degree in chemoinformatics at Sheffield, as well as for his long-standing support and encouragement; Janet Ash for her long-standing collaboration and for providing very helpful comments on this paper; Charles E. Davis and James E. Rush for providing very constructive comments on this paper; Val W. Metanomski for his eagle-eyed scrutiny of the text and his valuable comments thereon; and Patricia Wieland for her meticulous editing.

References

Adamson, G. W., Bush, J. A., McLure, A. H. W., & Lynch, M. F. (1974). An evaluation of a substructure search screen system based on bond-centred fragments. *Journal of Chemical Documentation*, *14*, 44–48.

Adamson, G. W., Clinch, V. A., & Lynch, M. F. (1973). Relationship between query and data-base microstructure in general substructure search systems. *Journal of Chemical Documentation*, *13,* 133–136.

Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., & Yapp, A. M. (1973). Strategic considerations in the design of a screening system for substructure searches of chemical structure files. *Journal of Chemical Documentation*, *13*, 153–157.

Addelston, A., & Goldsmith, U. J. (1966). Procedures for correcting errors in chemical literature. *Journal of Chemical Documentation*, *6*, 126–129.

Allen, F. H., et al. (1979). The Cambridge Crystallographic Data Centre: Computer-based search, retrieval, analysis and display of information. *Acta Crystallographia*, *B35*, 2331–2339.

Allen, F. H., & Lynch, M. F. (1989). The storage and retrieval of chemical structures. *Chemistry in Britain*, 25, 1101–1108.

Armitage, J. E., Crowe, J. E., Evans, P. N., Lynch, M. F., & McGuirk, J. A. (1967). Documentation of chemical reactions by computer analysis of structural changes. *Journal of Chemical Documentation*, *7*, 209–215.

Armitage, J. E., & Lynch, M. F. (1967a). Automatic detection of structural similarities among chemical compounds. *Journal of the Chemical Society, C*, 521–528.

Armitage J. E., & Lynch, M. F. (1967b). Articulation in the generation of subject indexes by computer. *Journal of Documentation*, *7,* 170–178.

Armitage, J. E., & Lynch, M. F. (1968). Some structural characteristics of articulated subject indexes. *Information Storage and Retrieval*, *4*, 101–111.

Ash, J. E., Chubb, P. A., Ward, S. E., Welford, S. M., & Willett, P. (1985). *Communication, storage and retrieval of chemical information*. Chichester, U.K.: Ellis Horwood.

Ash, J. E., & Hyde, E. (1975). *Chemical information systems*. Chichester, U.K.: Ellis Horwood.

Ash, J. E., Warr, W. A., & Willett, P. (Eds.). (1991). *Chemical structure systems: Computational techniques for representation, searching and processing of structural information*. Chichester, U.K.: Ellis Horwood.

Attias, R. (1983). DARC substructure search system: A new approach to chemical information. *Journal of Chemical Information and Computer Sciences, 23*, 102–108.

Austin, C. J. (1964). The MEDLARS system. *Datamation*, *10*(12), 28–31.

Barton, I . J., Lynch, M. F., Petrie J. H., & Snell, M. J. (1974). Variable-length character string analyses of three databases, and their application for file compression. In *Proceedings of the 1st Informatics Conference* (pp. 54–162). London: Aslib.

Bawden, D., & Mitchell, E. (Eds.). (1990). *Chemical information systems—beyond the structure diagram.* Chichester, U.K.: Ellis Horwood.

Blankenship, L., & Metanomski, V. W. (2002, November 15–17). The evolution of *Chemical Abstracts:* 95 years of responding to chemists' needs. Paper presented at the 2002 Conference on the History and Heritage of Scientific and Technical Information Systems, Philadelphia, PA.

Clare, A. C., Cook, E. M., & Lynch, M. F. (1972). The identification of variable-length equifrequent character strings in a natural language database. *Computer Journal*, *15*(3), 252–262.

Clinging, R., & Lynch, M. F. (1973). Production of printed indexes of chemical reactions. I: Analysis of functional group interconversions. *Journal of Chemical Documentation*, *13*, 98–102.

Clinging, R., & Lynch, M. F. (1974). Production of printed indexes of chemical reactions. II: Analysis of reactions involving ring formation, cleavage and interconversion. *Journal of Chemical Documentation*, *14,* 69–71.

Corey, E. J., & Wipke, W. T. (1969). Computer-assisted design of complex organic syntheses. *Science* (Washington), *166*(3902), 178–192.

Cossum, W. E., Krakiwsky, M. L., & Lynch, M. F. (1965). Advances in automatic chemical substructure searching techniques. *Journal of Chemical Documentation*, *5*, 33–35.

Craig, P. N. (1962). The documentation research chemist. *Journal of Chemical Documentation*, *2*, 169–171.

Crowe, J. E., Lynch, M. F., & Town, W. G. (1970). Analysis of structural characteristics of chemical compounds in a large computer-based file. I: Non-cyclic fragments. *Journal of the Chemical Society, C*, 990–996.

Davis, C. E. (2004). Indexing and editing at *Chemical Abstracts* before the registry system. In W. B. Rayward & M. E. Bowden (Eds.), *Proceedings of the 2002 Conference on the History and Heritage of Scientific and Technical Information Systems* (pp. 182–189). Medford, NJ: Information Today.

Davis, C. H., & Rush, J. E. (1974). *Information retrieval and documentation in chemistry*. Westport, CT: Greenwood Press.

Dittmar, P. G., Mockus, J., & Couvreur, K. M. (1977). An algorithmic computer graphics program for generating chemical structure diagrams. *Journal of Chemical Information and Computer Sciences*, *17*, 186–192.

Dittmar, P. G., Stobaugh, R. E., & Watson, C. E. (1976). The Chemical Abstracts Service Chemical Registry System. I: General design. *Journal of Chemical Information and Computer Sciences*, *16*, 11–121.

Dubois, J.-E. (1973). French national policy for chemical information and the DARC system as a potential tool of this policy. *Journal of Chemical Documentation*, *13*, 8–13.

Dubois, J.-E. (1974). DARC system in chemistry. In W. T. Wipke, S. R. Heller, R. J. Feldmann, & E. Hyde (Eds.), *Computer representation and manipulation of chemical information.* New York: John Wiley.

Dubois, J.-E., & Viellard, H. (1968a). Théorie de génération-description. I: Principes généraux [Theory of generation-description. I: General principles]. *Bulletin de la Société Chimique de France*, 900–904.

Dubois, J.-E., & Viellard, H. (1968b). Théorie de génération-description. II: Etablissement du descripteur uniligne d'un segment A_i-B_{ij}: le DEL [Theory of generation

description. II: Establishment of a linear description of a segment A_i-B_{ij}: the DEL]. *Bulletin de la Société Chimique de France*, 905–912.

Dubois, J.-E., & Viellard, H. (1968c) Théorie de génération-description. III: Description générale des structures par le DEL [Theory of generation-description. III: General description of structures by the DEL]. *Bulletin de la Société Chimique de France*, 913–919.

Dugundji, J., & Ugi, I. (1973). Algebraic model of constitutional chemistry as a basis for chemical computer programs. *Topics in Current Chemistry*, *39*, 19–64.

Dyson, G. M. (1946). *A new notation of organic chemistry.* London: Royal Institute of Chemistry Lecture, Chemical Society and Society of Chemical Industry.

Dyson, G. M. (1961). Current research at *Chemical Abstracts. Journal of Chemical Documentation*, *1*(2), 26.

Dyson, G. M., Cossum, W. E., Lynch, M. F., & Morgan, H. L. (1963a). Mechanical manipulation of chemical structures. *Information Storage and Retrieval*, *1*, 69–99.

Dyson, G. M., Cossum, W. E., Lynch, M. F., & Morgan, H. L. (1963b). Mechanical searching of chemical substructures. In H. P. Luhn (Ed.), *Automation and scientific communication* (pp. 15–19). Chicago: American Documentation Institute.

Dyson, G. M., Cossum, W. E., Lynch, M. F., & Morgan, H. L. (1963c). Mechanical encipherment of chemical ring structures from the random matrix. In H. P. Luhn (Ed.), *Automation and Scientific Communication* (pp. 79–82). Chicago: American Documentation Institute.

Feldman, A., & Hodes, L. (1975). An efficient design for chemical structure searching. I: The screens. *Journal of Chemical Information and Computer Sciences*, *15*, 147–152.

Feldman, A., Holland, D. B., & Jacobus, D. P. (1963). The automatic encoding of chemical structures. *Journal of Chemical Documentation*, *3,* 187–189.

Fugmann, R., Braun, W., & Vaupel, W. (1963). GREMAS—Ein Weg zur Klassifikation und Dokumentation in der organischen Chemie [GREMAS—A method for classification and documentation in organic chemistry]. *Nachrichten für Dokumentation*, *14*, 179–190.

Garfield, E. (1961). Chemico-linguistics: Computer translation of chemical nomenclature. *Nature* (London), *4798,* 192.

Garfield, E. (1962). An algorithm for translating chemical names to molecular formulas. *Journal of Chemical Documentation*, *2*(4), 177–179.

Gluck, D. J. (1965). A chemical structure storage and search system developed at Du Pont. *Journal of Chemical Documentation, 5(*2), 43–51.

Graf, W., Kaindl, H. K., Kniess, H., Schmidt, B., & Warszawski, R. (1982). Substructure retrieval by means of the BASIC Fragment Search Dictionary based on the Chemical Abstracts Service Registry III System. *Journal of Chemical Information and Computer Sciences*, *19,* 51–55.

Gray, N. A. B. (1986). *Computer-assisted structure elucidation*. New York: Wiley.

Harrison, J. M., & Lynch, M. F. (1970). Computer analysis of chemical reactions for storage and retrieval. *Journal of the Chemical Society*, *C*, 2082–2087.

Hausen, J. (1969). *Was nicht in den Annalen steht* [What is not recorded in the annals]. Weinheim/Bergstrasse: Verlag Chemie.

Hendrickson, J. B. (1971). A systematic characterization of structures and reactions for use in organic synthesis. *Journal of the American Chemical Society*, *93*, 6847–6862.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineering*, *40*, 1098–1101.

Hyde, E., & Thomson, L. D. (1968). Structure display. *Journal of Chemical Documentation*, *8*(3), 138–146.

International Union of Pure and Applied Chemistry, Commission on Codification, Ciphering, and Punched Card Techniques. (1961). *Rules for IUPAC notation of organic compounds*. New York: John Wiley.

Johnson, A. P., & Cook, A. P. (1985). Automatic keyword generation for reaction searching. In P. Willett (Ed.), *Modern approaches to chemical reaction searching*. London: Gower.

Kennard, O., Watson, D. G., & Town, W. G. (1972). Cambridge Crystallographic Data Centre. I: Bibliographic file. *Journal of Chemical Documentation, 12*, 14–19.

Kline, R. (2004). What is information theory a theory of? Boundary work among information theorists and information scientists in the United States and Britain during the early cold war. In W. B. Rayward & M. E. Bowden (Eds.), *Proceedings of the 2002 Conference on the History and Heritage of Scientific and Technical Information Systems* (pp. 15–28). Medford, NJ: Information Today.

Lederberg, J. (1990). How DENDRAL was conceived and born. In B. I. Blum & K. Duncan (Eds.), *A history of medical informatics* (pp. 14–44). New York: Association for Computing Machinery Press.

Leiter, D. P., Morgan, H. L., & Stobaugh, R. L. (1965). Installation and operation of a registry for chemical compounds. *Journal of Chemical Documentation*, *5,* 238–242.

Lipscomb, K. J., Lynch, M. F., & Willett, P. (1990). Chemical structure processing. In M. E.Williams (Ed.), *Annual reviews of information science and technology* (pp. 189–238). Medford, NJ: Information Today.

Luhn, H. P. (1955). *A serial notation for describing the topology of multi-dimensional branched structures (nodal index for branched structures*). New York: IBM Corporation.

Lynch, M. F. (1966). Subject indexes and automatic retrieval of information. *Journal of Documentation*, *22*, 167–185.

Lynch, M. F. (1968). Storage and retrieval of information on chemical structures. *Endeavour, 27,* 68–73.

Lynch, M. F. (1974). Screening large chemical files. In J. Ash & E. Hyde (Eds.), *Chemical information systems* (pp. 177–194). Chichester, U.K.: Ellis Horwood.

Lynch, M. F. (1977). Variety generation—a re-interpretation of Shannon's mathematical theory of communication, and its implications for information science. *Journal of the American Society for Information Science*, *28*(1), 19–25.

Lynch, M. F., Harrison, J. M., Town, W. G., & Ash, J. E. (1971). *Computer handling of chemical structure information*. London: Macdonald.

Lynch, M. F., Nunn, P. R., & Radcliffe, J. (1978). Production of printed indexes of chemical reactions using Wiswesser line notations. *Journal of Chemical Information and Computer Sciences*, *18*, 94–96.

Lynch, M. F., Orton, J., & Town, W. G. (1969). Organisation of large collections of chemical structures for computer searching. *Journal of the Chemical Society, C,* 1732–1736.

Lynch, M. F., & Willett, P. (1978). The automatic identification of chemical reaction sites. *Journal of Chemical Information and Computer Sciences, 18*, 154–159.

McGregor, J. J., & Willett, P. (1981). Use of a maximal common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *Journal of Chemical Information and Computer Sciences*, *21*, 137–140.

Meyer, E. (1965). Encoding of organic-chemical structural formulas and reactions by machine. In H. P. Luhn (Ed.), *Automation and scientific documentation* (p. 131). Washington, DC: American Documentation Institute.

Meyer, E., & Wenke, K. (1962). Ein System zur topologischen Verschlüsselung organisch-chemischer Strukturformeln für die mechanisierte Dokumentation [A system for topological coding of organic chemical structures for mechanized documentation]. *Nachrichten für Dokumentation*, *13*(1), 13–19.

Mooers, C. N. (1947). Putting probability to work in coding punched cards: Zatocoding. *Zator Technical Bulletin, 10* (Cambridge, MA: Zator Company).

Mooers, C. N. (1951). Ciphering structural formulas—the Zatopleg system. *Zator Technical Bulletin, 59* (Cambridge, MA: Zator Company).

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *Journal of Chemical Documentation, 5*(2), 107–113.

Mullen, J. M. (1966). Atom-by-atom typewriter input for computerized storage and retrieval of chemical structures. *Journal of Chemical Documentation, 7,* 88–93.

Opler, A., & Baird, N. (1959). Display of chemical structural formulas as digital computer output. *American Documentation*, *10*, 59–63.

Pape, U. (1968). The transformation and analysis of networks by means of computer algorithms. *International Journal of Computer Mathematics*, *3,* 75–110.

Petrarca, A. E., Lynch, M. F., & Rush, J. E. (1965). Methods of computer generation of unique structural representations of stereoisomers. *Journal of Chemical Documentation, 7*(3), 154–165.

Petrarca, A. E., & Rush, J. E. (1969). Methods for computer generation of unique configurational descriptors for stereoisomeric square planar and octahedral complexes. *Journal of Chemical Documentation*, *9*(1), 32–37.

Ray, L. C., & Kirsch, R. A. (1957). Finding chemical records by computer. *Science*, *126*(3227), 814–819.

Rush, J. E. (1976). Status of notation and topological systems and potential future trends. *Journal of Chemical Information and Computer Science, 16,* 202–210.

Rush, J. E. (1985). Computer hardware and software in chemical information processing. *Journal of Chemical Information and Computer Sciences*, *25*, 140–149.

Schenck, H. R., & Wegmüller, F. (1976). Substructure retrieval by means of the BASIC Fragment Search Dictionary based on the Chemical Abstracts Service Registry II System. *Journal of Chemical Information and Computer Sciences, 16,* 153–161.

Schofield, H., Willett, P., & Wiggins, G. (2001). Recent developments in chemoinformatics education. *Drug Discovery Today, 6,* 931–934.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423, 623–656.

Shannon, C. E. (1971). *Information theory* (vol. 12, p. 247b). Chicago, IL.: Encyclopaedia Britannica.

Smith, E. G. (1968). *The Wiswesser line-formula chemical notation.* New York: McGraw-Hill.

Tate, F. A. (1967). Handling chemical compounds in information systems. In C. Cuadro (Ed.), *Annual review of information science and technology* (pp. 285–309). Washington, DC: American Documentation Institute.

Ugi, I., et al. (1979). New applications for computers in chemistry. *Angewandte Chemie, International Edition in English*, *18*, 111–123.

Vander Stouw, G. G., Elliott, P. M., & Isenberg, A. C. (1974). Automated conversion of chemical substance names into atom-bond connection tables. *Journal of Chemical Documentation, 14*(3), 185–193.

Vander Stouw, G. G., Naznitsky, I., & Rush, J. E. (1967). Procedures for converting systematic names of organic compounds into atom-bond connection tables. *Journal of Chemical Documentation, 7*(3), 165–169.

Vladutz, G. E. (1963). Concerning one system of classification and codification of organic reactions. *Information Storage and Retrieval, 1,* 117–146.

Vladutz, G. E. (1977). *Development of a combined WLN/CTR*

multilevel approach to the algorithmic analysis of chemical reactions in view of their automatic indexing. London: British Library Research and Development Department.

Warr, W. A., & Suhr, C. (1992). *Chemical information management.* Weinheim: VCH Verlagsgesellschaft.

Weininger, D. (1988). SMILES, a chemical language and information system. 1: Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences, 28*(1), 31–36.

Wheland, G. W. (1949). *Advanced organic chemistry.* New York: John Wiley.

Willett, P. (1978). Computer analysis of chemical reaction information for storage and retrieval. Unpublished doctoral dissertation, University of Sheffield.

Willett, P. (Ed.). (1986). *Modern approaches to chemical reaction searching.* Aldershot, U.K.: Gower.

Wipke, W. T., & Dyott, T. M. (1974). Stereochemically unique naming algorithm. *Journal of the American Chemical Society, 96,* 4,834–4,842.

Wipke, W. T., Heller S. R., Feldmann, R. J., & Hyde, E. (Eds.). (1974). *Computer representation and manipulation of chemical information.* New York: John Wiley.

Wipke, W. T., & Howe, W. J. (Eds.). (1977). *Computer-assisted organic synthesis* (ACS Symposium Series, *61*). Washington DC: American Chemical Society.

Ziv, J., & Lempel, A. A. (1977). Universal algorithm for sequential data compression. *IEEE Transactions on Information Theory, 23*(3), 337–343.