The University Of Sheffield.

Chemoinformatics: the first half century

Peter Willett

Presented at the Third Strasbourg Summer School on Chemoinformatics, 25th June 2012

Overview

- Introduction to chemoinformatics
  - What it is
  - How it has developed
- Historically important papers
  - Roughly chronological ordering
  - A personal choice
  - Many, many omissions

Chemoinformatics’ role

- The pharmaceutical industry has been one of the great success stories of scientific research in the latter half of the twentieth century
    - Range of novel drugs for important therapeutic areas
- The computer has revolutionised how the industry uses chemical (and increasingly biological) information
    - Many of these developments are within the discipline we now know as chemoinformatics
    - Focus on lead discovery and lead optimisation phases of drug discovery (also applicable to other types of specialty chemicals)

Definitions

- F.K. Brown (1998) Chemoinformatics: what it is and how does it impact drug discovery? *Annual Reports in Medicinal Chemistry*, **33**, 375-384
  - "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization"
- G. Paris (August 1999 ACS meeting), quoted by W.A. Warr at http://www.warr.com/warrzone.htm
  - "Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information"
- J. Gasteiger and T. Engels (editors) (2003). *Chemoinformatics: a textbook*. Wiley-VCH.
  - "Chemoinformatics is the application of informatics methods to solve chemical problems."

Emergence of chemoinformatics: I

- The principal driving force for current interest was the emergence during the Nineties of combinatorial synthesis and high-throughput screening
- Aims to assist more directly in the discovery of novel bioactive molecules than heretofore
  - Integration of chemical information (archival functions) and molecular modelling (small-scale correlation)
  - Development of effective and efficient tools that can exploit the chemical and biological data explosion
- Other types of -*informatics* becoming common
- But chemoionformatics is by no means new...

Emergence of chemoinformatics: II

- First appearance of the core, journal, *Journal of Chemical Documentation*, in 1961
- First book on the subject appeared in 1971
    - M.F. Lynch et al., *Computer Handling of Chemical Structure Information*
- The first two textbooks with “chemoinformatics” in the title appeared in 2003
    - A.R. Leach and V.J. Gillet, *An Introduction to Chemoinformatics*
    - J. Gasteiger and T. Engel (eds.) *Chemoinformatics*
- The first international conference on the subject was held at Noordwijkerhout in 1973, and every three years since 1987

Emergence of chemoinformatics: III

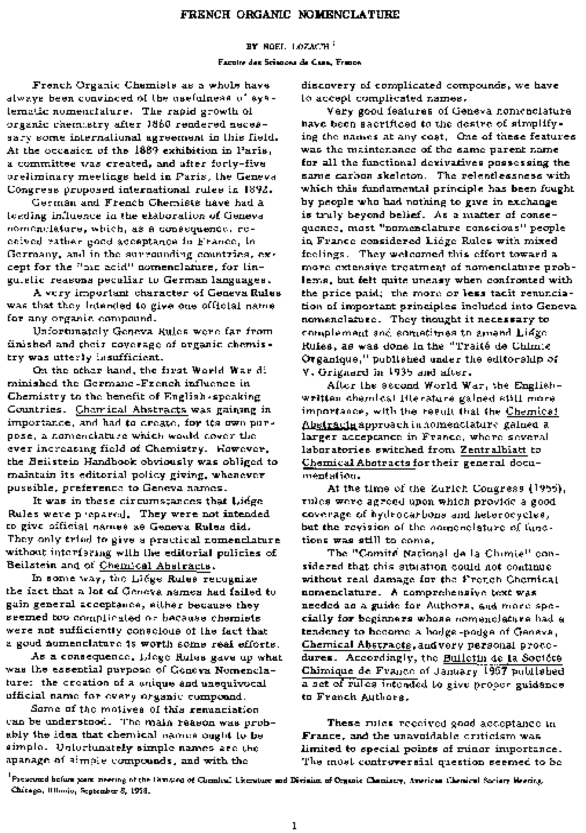
- Introduction of first full university courses in 2001
  - D.J. Wild and G. Wiggins (2006) Challenges for chemoinformatics education in drug discovery. *Drug Drug Discovery Today*, **11**, 436-439
- Nomenclature
  - Chemical informatics, chemical information (management/science), chemiinformatics, cheminformatics

    P. Willett (2008) A bibliometric analysis of chemoinformatics. *Aslib Proceedings*, **60**, 4-17
  - M. Hann and R. Green (1999) Chemoinformatics - a new name for an old problem? *Current Opinion in Chemical Biology*, **3**, 379-383

Basis in the literature

The chemical literature has been established for many years

- *Chemisches Journal* first appeared in 1778
- *Chemical Abstracts* first appeared in 1907
- Computerised systems first appeared in the early Sixties, shortly after the start of the core journal, the *Journal of Chemical Documentation*



FRENCH ORGANIC NOMENCLATURE

BY NOEL LOZACH[1]

Faculte des Sciences de Caen, France

French Organic Chemists as a whole have always been convinced of the usefulness of systematic nomenclature. The rapid growth of organic chemistry after 1860 rendered necessary some international agreement in this field. At the occasion of the 1889 exhibition in Paris, a committee was created, and after forty-five preliminary meetings held in Paris, the Geneva Congress proposed international rules in 1892.

German and French Chemists have had a leading influence in the elaboration of Geneva nomenclature, which, as a consequence, received rather good acceptance in France, in Germany, and in the surrounding countries, except for the "oic acid" nomenclature, for linguistic reasons peculiar to German languages.

A very important character of Geneva Rules was that they intended to give one official name for any organic compound.

Unfortunately Geneva Rules were far from finished and their coverage of organic chemistry was utterly insufficient.

On the other hand, the first World War diminished the Germanic-French influence in Chemistry to the benefit of English-speaking Countries. Chemical Abstracts was gaining in importance, and had to create, for its own purpose, a nomenclature which would cover the ever increasing field of Chemistry. However, the Beilstein Handbook obviously was obliged to maintain its editorial policy giving, whenever possible, preference to Geneva names.

It was in these circumstances that Liége Rules were prepared. They were not intended to give official names as Geneva Rules did. They only tried to give a practical nomenclature without interfering with the editorial policies of Beilstein and of Chemical Abstracts.

In some way, the Liége Rules recognize the fact that a lot of Geneva names had failed to gain general acceptance, either because they seemed too complicated or because chemists were not sufficiently conscious of the fact that a good nomenclature is worth some real efforts.

As a consequence, Liége Rules gave up what was the essential purpose of Geneva Nomenclature: the creation of a unique and unequivocal official name for every organic compound.

Some of the motives of this renunciation can be understood. The main reason was probably the idea that chemical names ought to be simple. Unfortunately simple names are the apanage of simple compounds, and with the discovery of complicated compounds, we have to accept complicated names.

Very good features of Geneva nomenclature have been sacrificed to the desire of simplifying the names at any cost. One of these features was the maintenance of the same parent name for all the functional derivatives possessing the same carbon skeleton. The relentlessness with which this fundamental principle has been fought by people who had nothing to give in exchange is truly beyond belief. As a matter of consequence, most "nomenclature conscious" people in France considered Liége Rules with mixed feelings. They welcomed this effort toward a more extensive treatment of nomenclature problems, but felt quite uneasy when confronted with the price paid; the more or less tacit renunciation of important principles included into Geneva nomenclature. They thought it necessary to complement and sometimes to amend Liége Rules, as was done in the "Traité de Chimie Organique," published under the editorship of V. Grignard in 1935 and after.

After the second World War, the English-written chemical literature gained still more importance, with the result that the Chemical Abstracts approach in nomenclature gained a larger acceptance in France, where several laboratories switched from Zentralblatt to Chemical Abstracts for their general documentation.

At the time of the Zurich Congress (1955), rules were agreed upon which provide a good coverage of hydrocarbons and heterocycles, but the revision of the nomenclature of functions was still to come.

The "Comité National de la Chimie" considered that this situation could not continue without real damage for the French Chemical nomenclature. A comprehensive text was needed as a guide for Authors, and more specially for beginners whose nomenclature had a tendency to become a hodge-podge of Geneva, Chemical Abstracts, and very personal procedures. Accordingly, the Bulletin de la Société Chimique de France of January 1957 published a set of rules intended to give proper guidance to French Authors.

These rules received good acceptance in France, and the unavoidable criticism was limited to special points of minor importance. The most controversial question seemed to be

[1] Presented before joint meeting of the Division of Chemical Literature and Division of Organic Chemistry, American Chemical Society Meeting, Chicago, Illinois, September 8, 1958.

1

The very first paper: N. Lozac'h (1961) *Journal of Chemical Documentation*, **1**, 1-4

Development of the journal

- Early issues of the *Journal of Chemical Documentation* largely comprised papers presented at meetings of the ACS Division of Chemical Literature
- First issue of *Journal of Chemical Information and Computer Sciences* (1975)
    - Largely comprised papers given at a National Academy of Sciences conference on databases (both textual and chemical)
- First issue of *Journal of Chemical Information and Modeling* (2005) with sections on
    - Chemical information, Computational chemistry, Computational biochemistry, Pharmaceutical modeling, and Bioinformatics

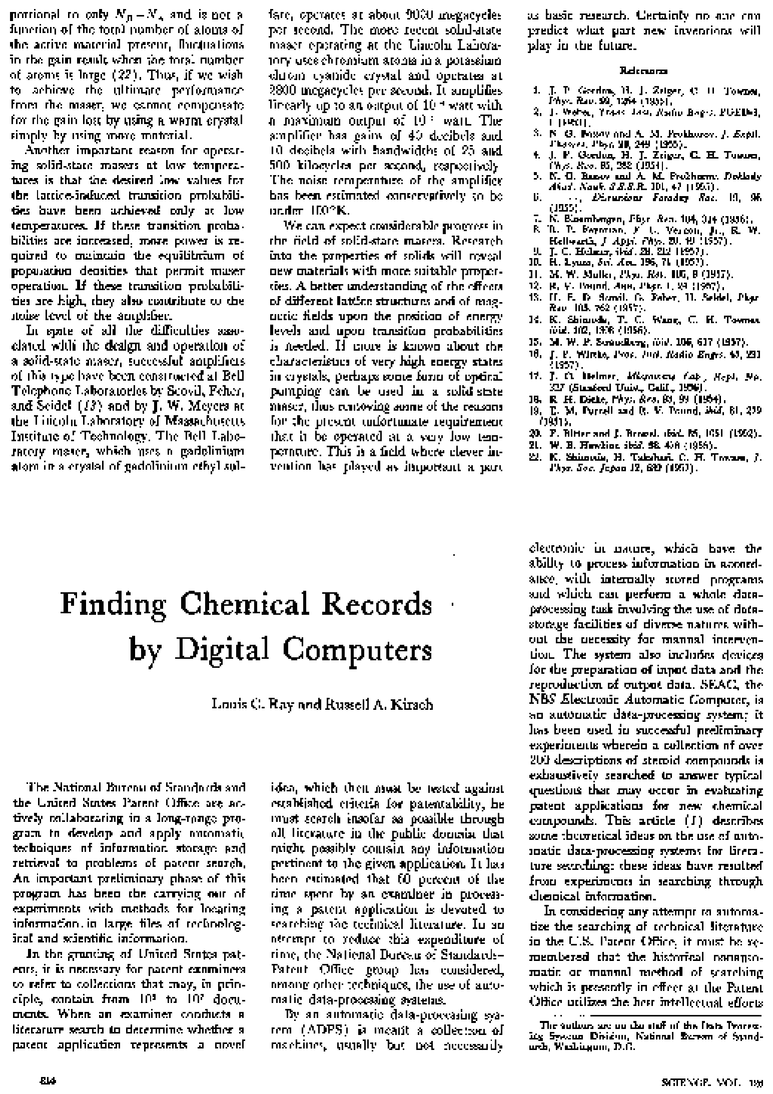
Many other journals now cover the subject, most obviously...

- *Molecular Informatics*
  - Started in 1982 as *Quantitative Structure-Activity Relationships*
- *Journal of Molecular Graphics and Modelling*
  - Started in 1983 as the *Journal of Molecular Graphics*
- *Journal of Computer-Aided Molecular Design*
  - Started in 1987
- *Journal of Cheminformatics*
  - Open access journal started in 2009

L.C. Ray and R.A. Kirsch (1957) Finding chemical records by digital computers, *Science*, **126**, 814-819

Introduced the use of graphs to represent 2D chemical structure diagrams

Applied a graph matching algorithm to a file of such representations to enable substructure searching



portional to only $N_B - N_A$ and is not a function of the total number of atoms of the active material present, fluctuations in the gain result when the total number of atoms is large (*22*). Thus, if we wish to achieve the ultimate performance from the maser, we cannot compensate for the gain lost by using a warm crystal simply by using more material.

Another important reason for operating solid-state masers at low temperatures is that the desired low values for the lattice-induced transition probabilities have been achieved only at low temperatures. If these transition probabilities are increased, more power is required to maintain the equilibrium of population densities that permit maser operation. If these transition probabilities are high, they also contribute to the noise level of the amplifier.

In spite of all the difficulties associated with the design and operation of a solid-state maser, successful amplifiers of this type have been constructed at Bell Telephone Laboratories by Scovil, Feher, and Seidel (*13*) and by J. W. Meyers at the Lincoln Laboratory of Massachusetts Institute of Technology. The Bell Laboratory maser, which uses a gadolinium atom in a crystal of gadolinium ethyl sulfate, operates at about 9000 megacycles per second. The more recent solid-state maser operating at the Lincoln Laboratory uses chromium atoms in a potassium cobalt cyanide crystal and operates at 2800 megacycles per second. It amplifies linearly up to an output of 10^{-8} watt with a maximum output of 10^{-6} watt. The amplifier has gains of 40 decibels and 10 decibels with bandwidths of 25 and 500 kilocycles per second, respectively. The noise temperature of the amplifier has been estimated conservatively to be under 150°K.

We can expect considerable progress in the field of solid-state masers. Research into the properties of solids will reveal new materials with more suitable properties. A better understanding of the effects of different lattice structures and of magnetic fields upon the position of energy levels and upon transition probabilities is needed. If more is known about the characteristics of very high energy states in crystals, perhaps some form of optical pumping can be used in a solid state maser, thus removing some of the reasons for the present unfortunate requirement that it be operated at a very low temperature. This is a field where clever invention has played as important a part as basic research. Certainly no one can predict what part new inventions will play in the future.

References

1. J. P. Gordon, H. J. Zeiger, C. H. Townes, *Phys. Rev.* 99, 1264 (1955).
2. J. Weber, *Trans. IRE, Radio Engrs.* PGED-3, 1 (1953).
3. N. G. Basov and A. M. Prokhorov, *J. Exptl. Theoret. Phys.* 28, 249 (1955).
4. J. P. Gordon, H. J. Zeiger, C. H. Townes, *Phys. Rev.* 95, 282 (1954).
5. N. G. Basov and A. M. Prokhorov, *Doklady Akad. Nauk S.S.S.R.* 101, 47 (1955).
6. ———, *Discussions Faraday Soc.* 19, 96 (1955).
7. N. Bloembergen, *Phys. Rev.* 104, 324 (1956).
8. R. P. Feynman, F. L. Vernon, Jr., R. W. Hellwarth, *J. Appl. Phys.* 28, 49 (1957).
9. J. C. Helmer, *ibid.* 28, 212 (1957).
10. H. Lyons, *Sci. Am.* 196, 71 (1957).
11. M. W. Muller, *Phys. Rev.* 106, 8 (1957).
12. R. V. Pound, *Ann. Phys.* 1, 24 (1957).
13. H. E. D. Scovil, G. Feher, H. Seidel, *Phys. Rev.* 105, 762 (1957).
14. K. Shimoda, T. C. Wang, C. H. Townes, *ibid.* 102, 1308 (1956).
15. M. W. P. Strandberg, *ibid.* 106, 617 (1957).
16. J. P. Wittke, *Proc. Inst. Radio Engrs.* 45, 291 (1957).
17. J. C. Helmer, *Microwave Lab. Rept. No.* 327 (Stanford Univ., Calif., 1956).
18. R. H. Dicke, *Phys. Rev.* 93, 99 (1954).
19. E. M. Purcell and R. V. Pound, *ibid.* 81, 279 (1951).
20. F. Bitter and J. Brossel, *ibid.* 85, 1051 (1952).
21. W. B. Hawkins, *ibid.* 98, 478 (1955).
22. K. Shimoda, H. Takahasi, C. H. Townes, *J. Phys. Soc. Japan* 12, 686 (1957).

Finding Chemical Records by Digital Computers

Louis C. Ray and Russell A. Kirsch

The National Bureau of Standards and the United States Patent Office are actively collaborating in a long-range program to develop and apply automatic techniques of information storage and retrieval to problems of patent search. An important preliminary phase of this program has been the carrying out of experiments with methods for locating information in large files of technological and scientific information.

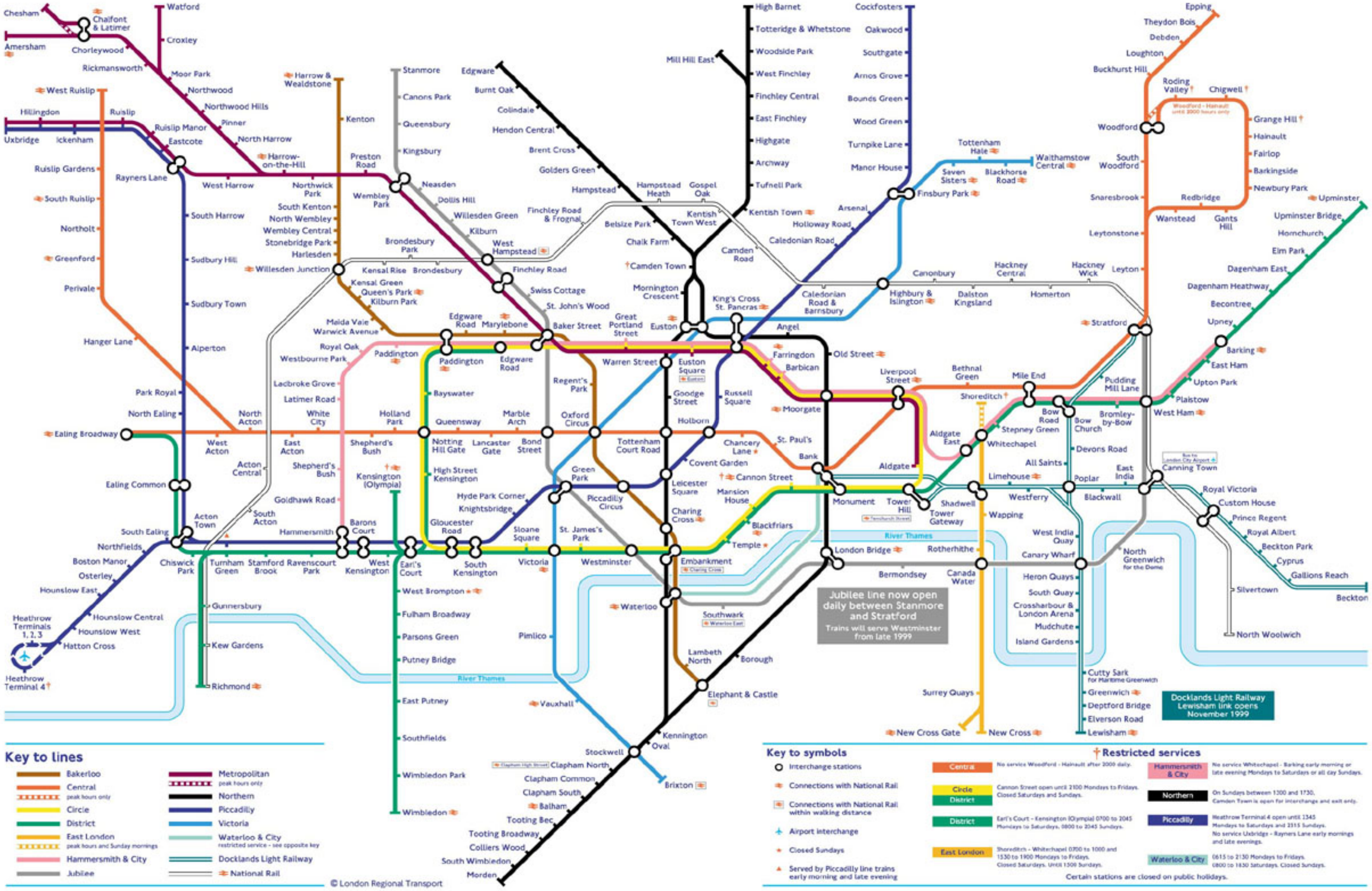
In the granting of United States patents, it is necessary for patent examiners to refer to collections that may, in principle, contain from 10^3 to 10^7 documents. When an examiner conducts a literature search to determine whether a patent application represents a novel idea, which then must be tested against established criteria for patentability, he must search insofar as possible through all literature in the public domain that might possibly contain any information pertinent to the given application. It has been estimated that 60 percent of the time spent by an examiner in processing a patent application is devoted to searching the technical literature. In an attempt to reduce this expenditure of time, the National Bureau of Standards–Patent Office group has considered, among other techniques, the use of automatic data-processing systems.

By an automatic data-processing system (ADPS) is meant a collection of machines, usually but not necessarily electronic in nature, which have the ability to process information in accordance with internally stored programs and which can perform a whole data-processing task involving the use of data-storage facilities of diverse natures without the necessity for manual intervention. The system also includes devices for the preparation of input data and the reproduction of output data. SEAC, the NBS Electronic Automatic Computer, is an automatic data-processing system; it has been used in successful preliminary experiments wherein a collection of over 200 descriptions of steroid compounds is exhaustively searched to answer typical questions that may occur in evaluating patent applications for new chemical compounds. This article (*1*) describes some theoretical ideas on the use of automatic data-processing systems for literature searching; these ideas have resulted from experiments in searching through chemical information.

In considering any attempt to automatize the searching of technical literature in the U.S. Patent Office, it must be remembered that the historical nonautomatic or manual method of searching which is presently in effect at the Patent Office utilizes the best intellectual efforts

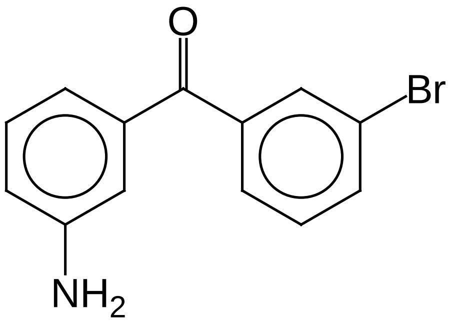
The authors are on the staff of the Data Processing Systems Division, National Bureau of Standards, Washington, D.C.

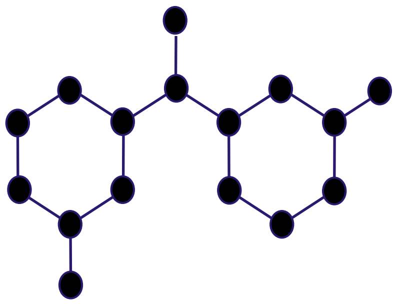
Graph theory describes sets of objects (*nodes*) and the relationships between pairs of them (*edges*)



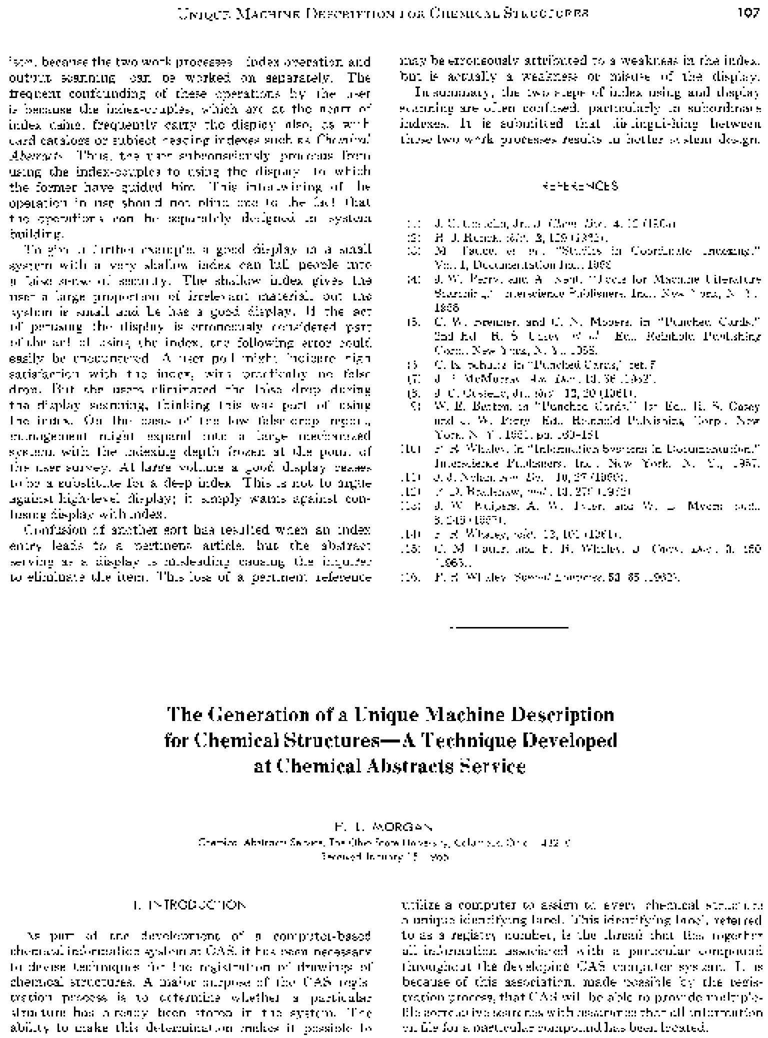
Representation of molecules by graphs

- Graph theory is applicable to any context that can be described by nodes and edges
- Can hence be used to represent and search both 2D and 3D chemical structures
- 2D chemical structure
  - Connection tables
  - Nodes correspond to atoms
  - Edges correspond to bonds
  - 2D graph describes topology
- 3D chemical structure (see later)
  - Edges correspond to distances
  - 3D graph describes geometry





- Throughout the early Sixties, Chemical Abstracts Service received very substantial funding to develop textual and chemical processing
- Heart of the new processing was the CAS Registry (now contains ca. 68M organic and inorganic molecules – see http://www.cas.org/)
- Adopted a graph-based approach: at the heart is a systematic naming procedure for chemical graphs: H. L. Morgan (1965) The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service, *Journal of Chemical Documentation* **5**, 107-113.



ism, because the two work processes—index operation and output scanning—can be worked on separately. The frequent confounding of these operations by the user is because the index-couples, which are at the heart of index using, frequently carry the display also, as with card catalogs or subject heading indexes such as *Chemical Abstracts*. Thus, the user subconsciously proceeds from using the index-couples to using the display to which the former have guided him. This intertwining of the operation in use should not blind one to the fact that the operations can be separately designed in system building.

To give a further example, a good display in a small system with a very shallow index can lull people into a false sense of security. The shallow index gives the user a large proportion of irrelevant material, but the system is small and he has a good display. If the act of perusing the display is erroneously considered part of the act of using the index, the following error could easily be encountered. A user poll might indicate high satisfaction with the index, with practically no false drops. But the users eliminated the false drop during the display scanning, thinking this was part of using the index. On the basis of the low false drop report, management might expand into a large mechanized system with the indexing depth frozen at the point of the user survey. At large volume a good display ceases to be a substitute for a deep index. This is not to argue against high-level display; it simply warns against confusing display with index.

Confusion of another sort has resulted when an index entry leads to a pertinent article, but the abstract serving as a display is misleading, causing the inquirer to eliminate the item. This loss of a pertinent reference may be erroneously attributed to a weakness in the index, but is actually a weakness or misuse of the display.

In summary, the two steps of index using and display scanning are often confused, particularly in subordinate indexes. It is submitted that distinguishing between these two work processes results in better system design.

REFERENCES

(1) J. C. Costello, Jr., J. Chem. Doc., 4, 12 (1964).
(2) H. J. [illegible], ibid., 3, 129 (1963).
(3) M. Taube, et al., "Studies in Coordinate Indexing," Vol. 1, Documentation Inc., 1953.
(4) J. W. Perry and A. Kent, "Tools for Machine Literature Searching," Interscience Publishers, Inc., New York, N. Y., 1958.
(5) C. W. [illegible] and C. N. Mooers, in "Punched Cards," 2nd Ed., R. S. Casey, et al., Ed., Reinhold Publishing Corp., New York, N. Y., 1958.
(6) C. [illegible] in "Punched Cards," ref. 5.
(7) J. [illegible] McMurray, Am. Doc., 13, 56 (1962).
(8) J. C. Costello, Jr., ibid., 12, 20 (1961).
(9) W. E. Batten, in "Punched Cards," 1st Ed., R. S. Casey and J. W. Perry, Ed., Reinhold Publishing Corp., New York, N. Y., 1951, pp. 169-181.
(10) F. R. Whaley, in "Information Systems in Documentation," Interscience Publishers, Inc., New York, N. Y., 1957.
(11) J. J. Nolan, Am. Doc., 10, 27 (1959).
(12) [illegible] Bradshaw, ibid., 13, 270 (1962).
(13) J. W. Kuipers, A. W. Tyler, and W. L. Myers, ibid., 8, 246 (1957).
(14) F. R. Whaley, ibid., 12, 101 (1961).
(15) C. M. Lucas and F. R. Whaley, J. Chem. Doc., 3, 150 (1963).
(16) F. R. Whaley, Special Libraries, 53, 85 (1962).

The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service

H. L. MORGAN
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

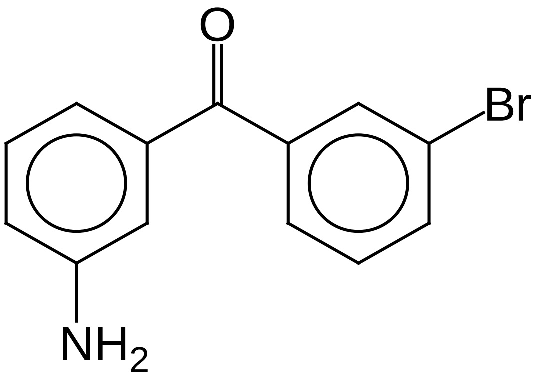
Received January 15, 1965

I. INTRODUCTION

As part of the development of a computer-based chemical information system at CAS, it has been necessary to devise techniques for the registration of drawings of chemical structures. A major purpose of the CAS registration process is to determine whether a particular structure has already been stored in the system. The ability to make this determination makes it possible to utilize a computer to assign to every chemical structure a unique identifying label. This identifying label, referred to as a registry number, is the thread that ties together all information associated with a particular compound throughout the developing CAS computer system. It is because of this association, made possible by the registration process, that CAS will be able to provide multiple-file correlative searches with assurance that all information on file for a particular compound has been located.

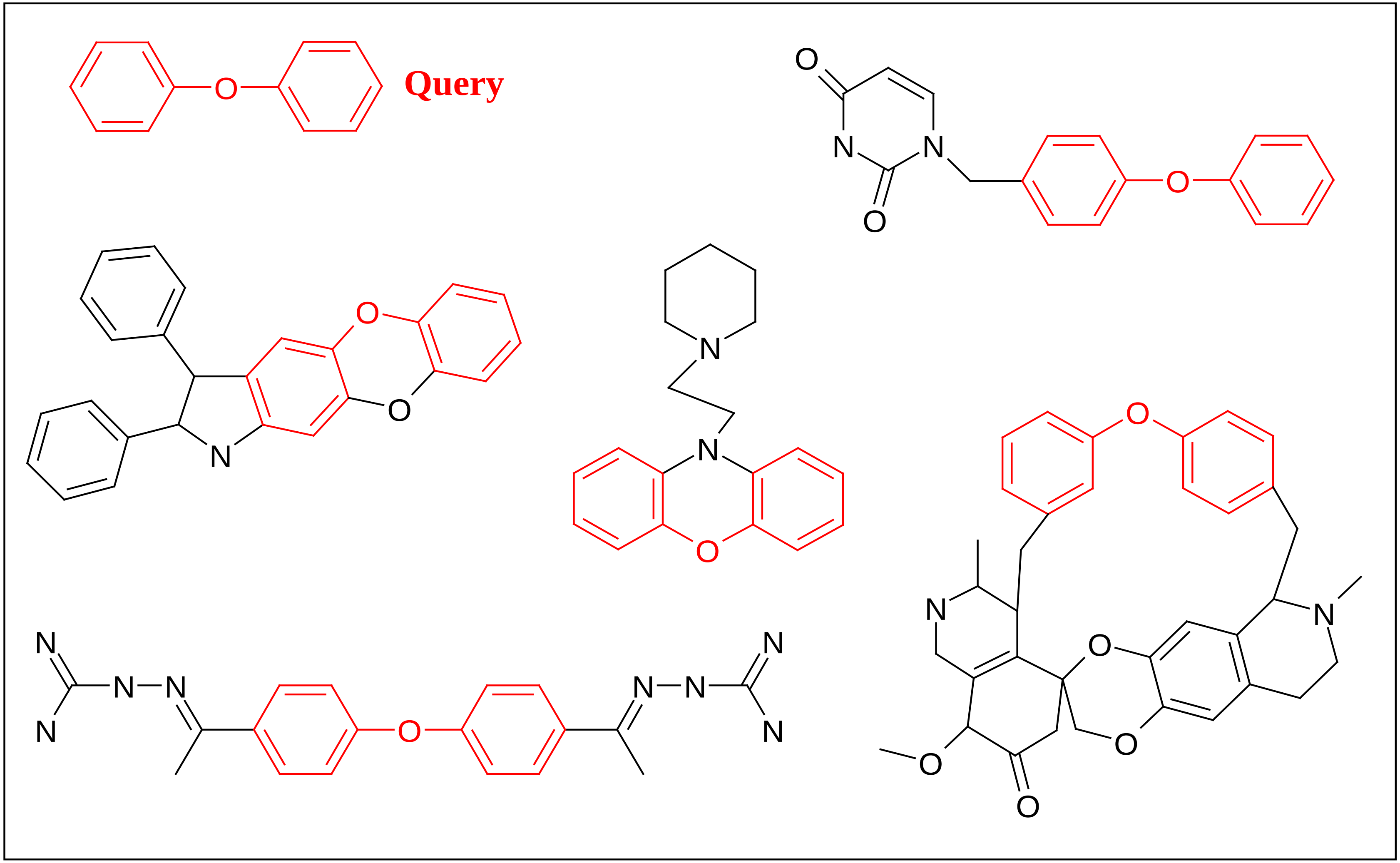
Wisweser Line Notation

- CAS work of long-term importance but most early industrial systems were based on Wiswesser Line Notation (WLN)
- Use of WLN till late Seventies
  - Cf SMILES and now InChI
- Need for conversion to connection tables
  - ICI CROSSBOW system
  - L.H. Thomson *et al*., (1967) Organic search and display using a connectivity matrix derived from Wiswesser notation, *Journal of Chemical Documentation* **7**, 204-209



ZR CVR CE

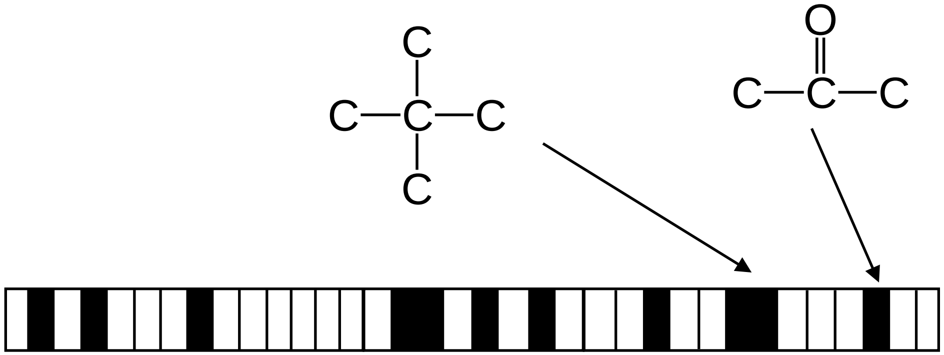
2D substructure search output



Substructure searching

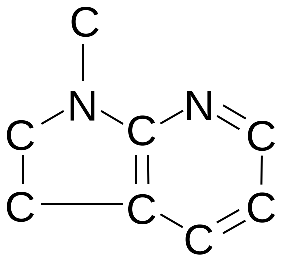
- Ability to retrieve all molecules in a database containing a user-defined substructure
  - Use of a *subgraph isomorphism algorithm*
  - Completely *effective*, but *efficiency* very low
- Standard methods such as set reduction (Sussenguth, 1965) and relaxation (Ullmann, 1976) underlie all operational substructure searching systems (both 2D and 3D)
  - Still not sufficiently fast so need for initial filter to eliminate molecules from graph processing
  - Encoding fragment screens describing query substructures and database structures in a *bit-string* or *fingerprint*

Binary vector

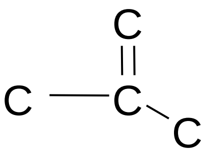


- Each bit in the bit-string (binary vector) records the presence (“1”) or absence (“0”) of a particular fragment in the molecule.
    - Typical length is a few hundred or few thousand bits
- A database structure is passed on for subgraph matching only if its bit-string contains all of the bits that have been set in the query’s bit-string
- How to select the fragments?
    - J.E. Crowe *et al*. (1970) Analysis of structural characteristics of chemical compounds in a large computer-based file. *Journal of the Chemical Society (C)* 990-996.

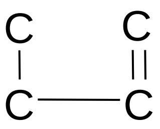
Example fragments



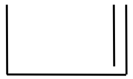
a. Augmented Atom
C rs C rd C rs C



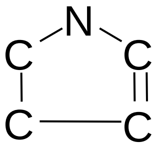
b. Atom Sequence
C rs C rs C rd C



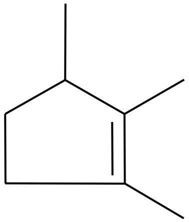
c. Bond Sequence
AA rs AA rs AA rd AA



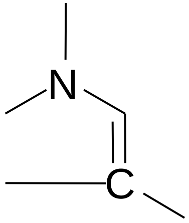
d. Ring Composition
N rs C rd C rs C rs C rs



e. Ring Fusion
XX3 XX3 XX3 XX2 XX2

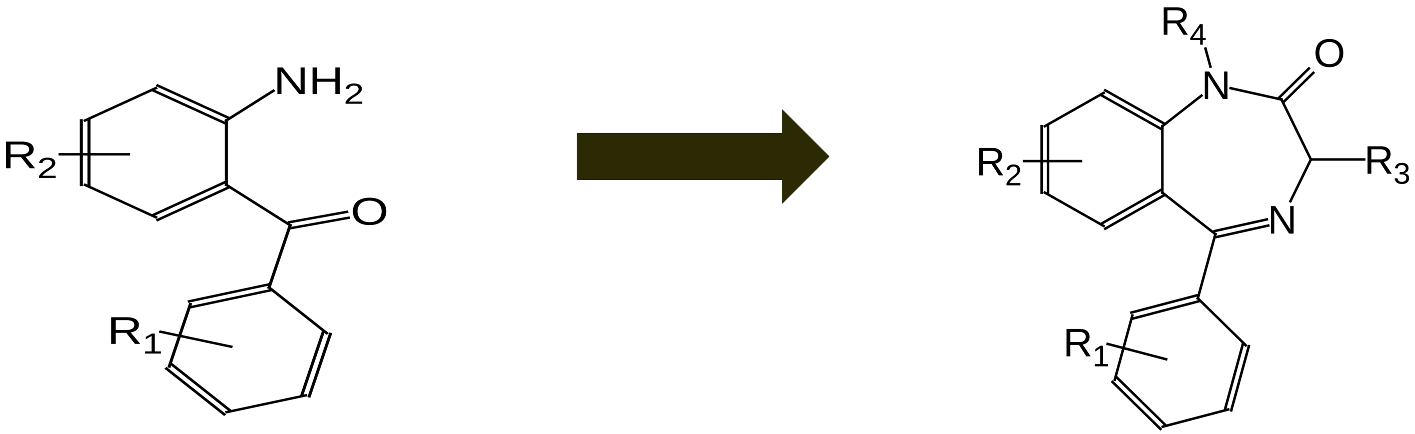


f. Atom Pair
N 0;3 - 2 - C 0;3



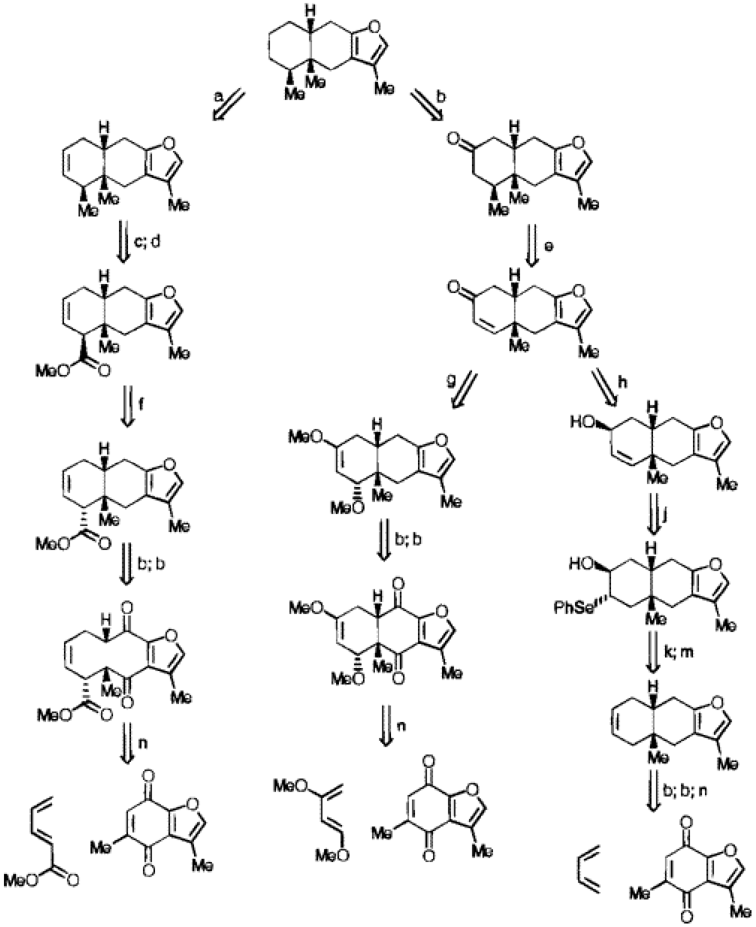
Reaction databases

- How to search for structural changes occurring in a reaction?
- G.E. Vleduts (1963) Concerning one system of classification and codification of organic reactions, *Information Storage and Retrieval* **1**, 17-146
  - Index a reaction by just those parts that have changed, the *reaction centre*, to allow searches for both changed and unchanged substructures
  - Practical realisation of his ideas not till early Eighties



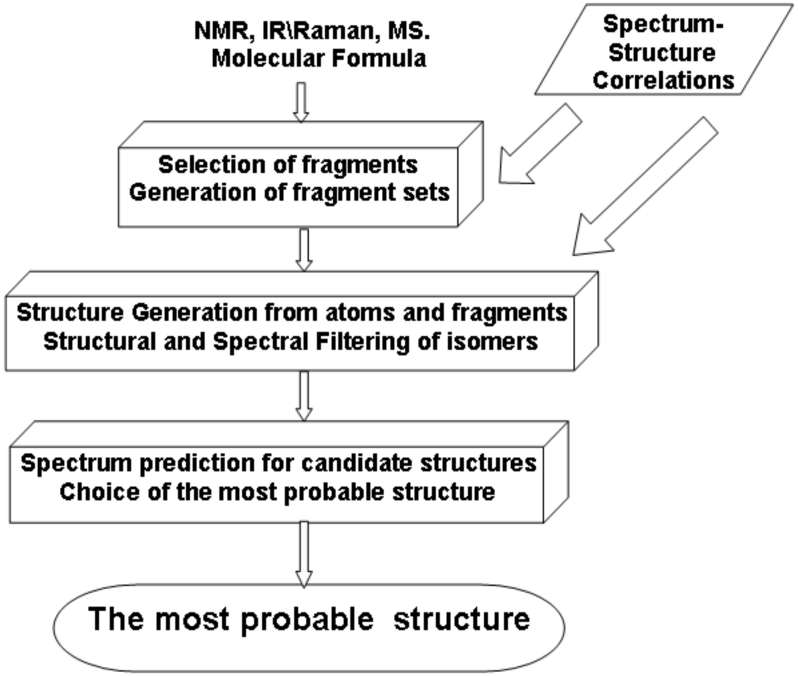
Computer-aided synthesis design

- Vleduts’ paper was also the first to suggest computer-aided synthesis design
- Potential syntheses of a target molecule using a reactions database plus appropriate inference mechanisms (“retrosynthesis”)
- An early example of an expert system
- First implemented in OCSS (subsequently LHASA) in 1969
- CASD programs can also work in the synthetic direction



Computer-aided structure elucidation

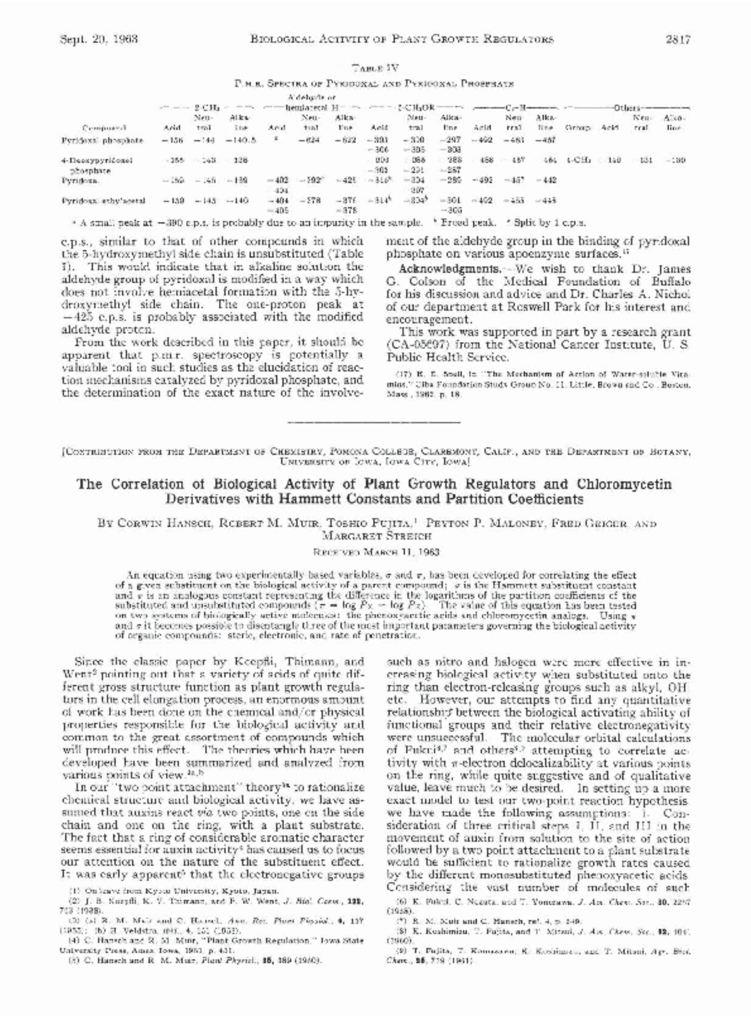
- Identification of an unknown substance based on spectral information
- DENDRAL project started in 1965: first paper J. Lederberg *et al*. (1969) Applications of artificial intelligence for chemical inference. 1. *Journal of the American Chemical Society*, **91**, 2973-2976
- First of a whole series of AI projects at Stanford (e.g., MYCIN, Prospector, XCON)


NMR, IR\Raman, MS.
Molecular Formula
Spectrum-Structure Correlations
Selection of fragments
Generation of fragment sets
Structure Generation from atoms and fragments
Structural and Spectral Filtering of isomers
Spectrum prediction for candidate structures
Choice of the most probable structure
The most probable structure

The University Of Sheffield.

Hansch analysis

- C. Hansch *et al*. (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients, *Nature* ,**194**, 178-180.
- Use of linear regression analysis to correlate physicochemical parameters with bioactivity
- The standard technique for Quantitative Structure-Activity Relationship (QSAR) studies for over two decades



Sept. 20, 1963 BIOLOGICAL ACTIVITY OF PLANT GROWTH REGULATORS 2817

TABLE IV

P.M.R. SPECTRA OF PYRIDOXAL AND PYRIDOXAL PHOSPHATE

	2-CH₃			Aldehyde or hemiacetal H			5-CH₂OR			C₆-H			Others			
Compound	Acid	Neutral	Alkaline	Acid	Neutral	Alkaline	Acid	Neutral	Alkaline	Acid	Neutral	Alkaline	Group	Acid	Neutral	Alkaline
Pyridoxal phosphate	−156	−144	−140.5	a	−624	−622	−301 −306	−310 −305	−297 −303	−492	−463	−467				
4-Deoxypyridoxal phosphate	−155	−143	−128				−303 −301	−286 −291	−288 −287	−468	−457	−464	4-CH₃	−150	−131	−130
Pyridoxal	−150	−146	−139	−402 −404	−392c	−425	−316b	−304 −307	−290	−493	−457	−442				
Pyridoxal ethylacetal	−150	−145	−140	−404 −405	−378	−376 −378	−314b	−304b	−301 −305	−492	−455	−445				

a A small peak at −390 c.p.s. is probably due to an impurity in the sample. b Freed peak. c Split by 1 c.p.s.

c.p.s., similar to that of other compounds in which the 5-hydroxymethyl side chain is unsubstituted (Table I). This would indicate that in alkaline solution the aldehyde group of pyridoxal is modified in a way which does not involve hemiacetal formation with the 5-hydroxymethyl side chain. The one-proton peak at −425 c.p.s. is probably associated with the modified aldehyde proton.

From the work described in this paper, it should be apparent that p.m.r. spectroscopy is potentially a valuable tool in such studies as the elucidation of reaction mechanisms catalyzed by pyridoxal phosphate, and the determination of the exact nature of the involvement of the aldehyde group in the binding of pyridoxal phosphate on various apoenzyme surfaces.[17]

Acknowledgments.—We wish to thank Dr. James G. Colson of the Medical Foundation of Buffalo for his discussion and advice and Dr. Charles A. Nichol of our department at Roswell Park for his interest and encouragement.

This work was supported in part by a research grant (CA-05697) from the National Cancer Institute, U. S. Public Health Service.

[17] E. E. Snell, in "The Mechanism of Action of Water-soluble Vitamins," Ciba Foundation Study Group No. 11, Little, Brown and Co., Boston, Mass., 1962, p. 18.

[CONTRIBUTION FROM THE DEPARTMENT OF CHEMISTRY, POMONA COLLEGE, CLAREMONT, CALIF., AND THE DEPARTMENT OF BOTANY, UNIVERSITY OF IOWA, IOWA CITY, IOWA]

The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients

BY CORWIN HANSCH, ROBERT M. MUIR, TOSHIO FUJITA,[1] PEYTON P. MALONEY, FRED GEIGER AND MARGARET STREICH

RECEIVED MARCH 11, 1963

An equation using two experimentally based variables, σ and π, has been developed for correlating the effect of a given substituent on the biological activity of a parent compound; σ is the Hammett substituent constant and π is an analogous constant representing the difference in the logarithms of the partition coefficients of the substituted and unsubstituted compounds ($\pi = \log P_X - \log P_H$). The value of this equation has been tested on two systems of biologically active molecules: the phenoxyacetic acids and chloromycetin analogs. Using π and σ it becomes possible to disentangle three of the most important parameters governing the biological activity of organic compounds: steric, electronic, and rate of penetration.

Since the classic paper by Koepfli, Thimann, and Went[2] pointing out that a variety of acids of quite different gross structure function as plant growth regulators in the cell elongation process, an enormous amount of work has been done on the chemical and/or physical properties responsible for the biological activity and common to the great assortment of compounds which will produce this effect. The theories which have been developed have been summarized and analyzed from various points of view.[3a,b]

In our "two point attachment" theory[3a] to rationalize chemical structure and biological activity, we have assumed that auxins react *via* two points, one on the side chain and one on the ring, with a plant substrate. The fact that a ring of considerable aromatic character seems essential for auxin activity[4] has caused us to focus our attention on the nature of the substituent effect. It was early apparent[5] that the electronegative groups such as nitro and halogen were more effective in increasing biological activity when substituted onto the ring than electron-releasing groups such as alkyl, OH, etc. However, our attempts to find any quantitative relationship between the biological activating ability of functional groups and their relative electronegativity were unsuccessful. The molecular orbital calculations of Fukui[6,7] and others[8,9] attempting to correlate activity with π-electron delocalizability at various points on the ring, while quite suggestive and of qualitative value, leave much to be desired. In setting up a more exact model to test our two-point reaction hypothesis we have made the following assumptions: 1. Consideration of three critical steps I, II, and III in the movement of auxin from solution to the site of action followed by a two point attachment to a plant substrate would be sufficient to rationalize growth rates caused by the different monosubstituted phenoxyacetic acids. Considering the vast number of molecules of such

[1] On leave from Kyoto University, Kyoto, Japan.

[2] J. B. Koepfli, K. V. Thimann, and F. W. Went, *J. Biol. Chem.*, **122**, 763 (1938).

[3] (a) R. M. Muir and C. Hansch, *Ann. Rev. Plant Physiol.*, **6**, 157 (1955); (b) B. Veldstra, *ibid.*, 4, 151 (1953).

[4] C. Hansch and R. M. Muir, "Plant Growth Regulation," Iowa State University Press, Ames, Iowa, 1961, p. 431.

[5] C. Hansch and R. M. Muir, *Plant Physiol.*, **25**, 389 (1950).

[6] K. Fukui, C. Nagata, and T. Yonezawa, *J. Am. Chem. Soc.*, **80**, 2267 (1958).

[7] R. M. Muir and C. Hansch, ref. 4, p. 249.

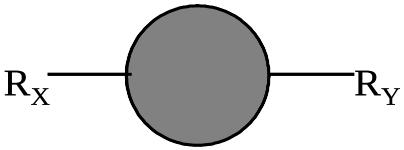
[8] K. Kushimizu, T. Fujita, and T. Mitsui, *J. Am. Chem. Soc.*, **82**, 4017 (1960).

[9] T. Fujita, T. Komatsu, K. Koshimizu, and T. Mitsui, *Agr. Biol. Chem.*, **25**, 719 (1961).

Free-Wilson analysis

S.M. Free and J.W. Wilson (1964) A mathematical contribution to structure-activity studies, *Journal of Medicinal Chemistry*, **7**, 395-399

Use of structural, rather than physicochemical, variables in the regression , these denoting the presence of substituents on a common scaffold



R_X = X1, X2,….
R_Y = Y1, Y2, ….

	Indicator variables					
Molecule ID	X1	X2	…..	Y1	Y2	…..
1	1	0	…..	1	0	…..
2	0	1	…..	1	0	…..
3	1	0	…..	0	1	…..
4	0	1	…..	0	1	…..
…..	…..	…..	…..	…..	…..	…..

$$\mathrm{y} = \mathrm{C} + a_1\mathrm{X1} + a_2\mathrm{X2} + a_3\mathrm{X3} + ...b_1\mathrm{Y1} + b_2\mathrm{Y2} + b_3\mathrm{Y3} + ...$$

$$\mathrm{y} = \mathrm{C} + \sum_{i=1}^{n} a_i\mathrm{X}i + \sum_{j=1}^{m} b_j\mathrm{Y}j$$

Substructural analysis

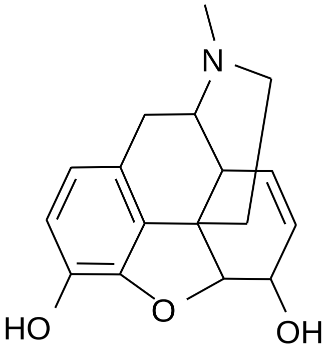
- R.D. Cramer *et al*. (1974) Substructural analysis. A novel approach to the problem of drug design, *Journal of Medicinal Chemistry*, **17**, 533-535
- Extension of Free-Wilson ideas to encompass
  - Structurally diverse molecules
  - Qualitative activity data
- Calculation of fragment weights
  - Probability that a molecule containing that fragment will be active, e.g., $N_{act}/(N_{act}+N_{inanct})$
- Used in US government anti-cancer programme in Eighties, but then in abeyance for many years till “re-discovery” as naive Bayesian classifier
  - First application of machine learning in virtual screening

Moving on

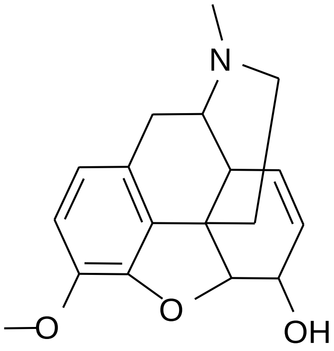
- Throughout the Seventies, chemical search systems (mainly WLN-based) became widely available across the pharmaceutical industry
- Extensive use of QSAR methods in lead optimisation
- Computer hardware/software limitations meant processing slow
- Things did not change much till the late-Seventies/early-Eighties, with the advent of MDL (now Accelrys) and CAS Online

Similarity searching

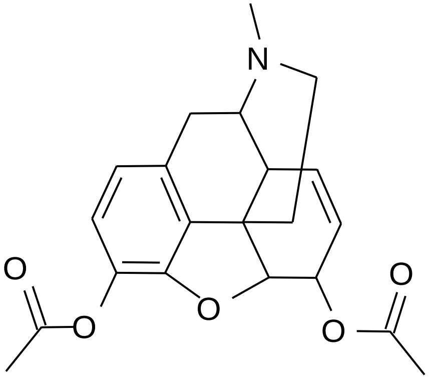
- Substructure searching very powerful but requires a clear view of the types of structures of interest
- Given a *target* (or *reference*) structure find molecules in a database that are most similar to it (“give me ten more like this”)
- The *similar property principle* states that structurally similar molecules tend to have similar properties (cf *neighbourhood principle*)



Morphine



Codeine

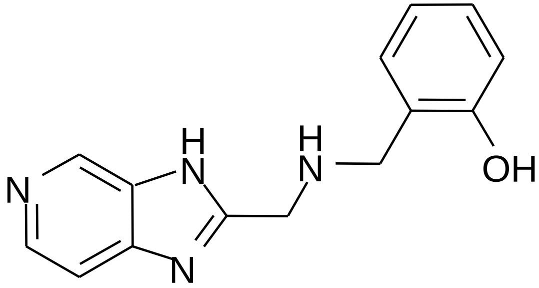


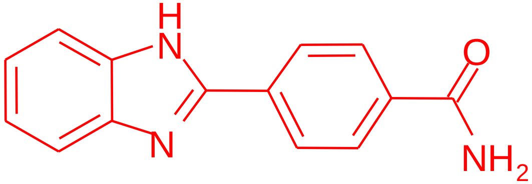
Heroin

How to define chemical similarity?

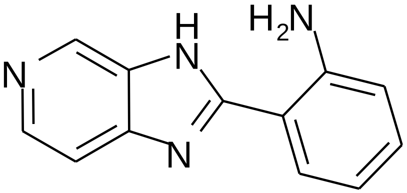
- Most obvious way is use of a maximum common subgraph isomorphism procedure but far too time-consuming for database-scale applications
- Use of fingerprint comparisons
  - G.W. Adamson and J.A. Bush (1973) A method for the automatic classification of chemical structures, *Information Storage and Retrieval*, **9**, 561-568
- How to use this idea?
  - Operational implementations from mid-Eighties with systems at Lederle, Pfizer/Sheffield and Upjohn
- Still the most widely used approach, despite inherent simplicity

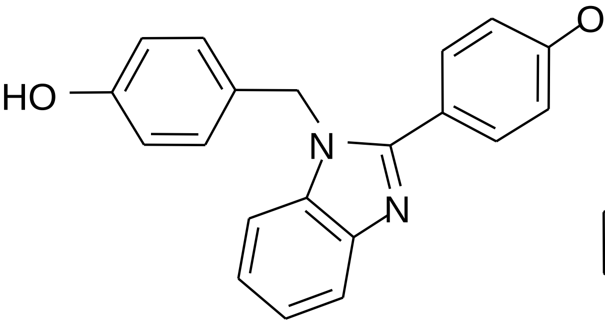
Tanimoto-based 2D similarity searching

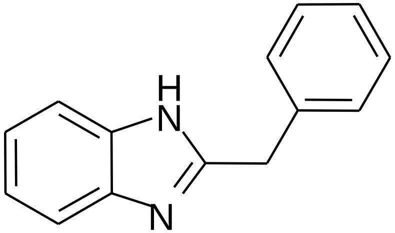


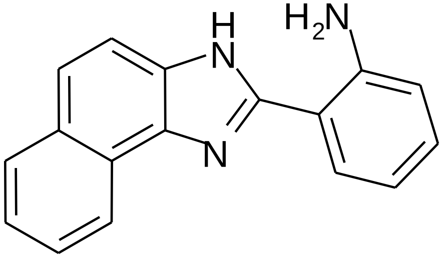


Query



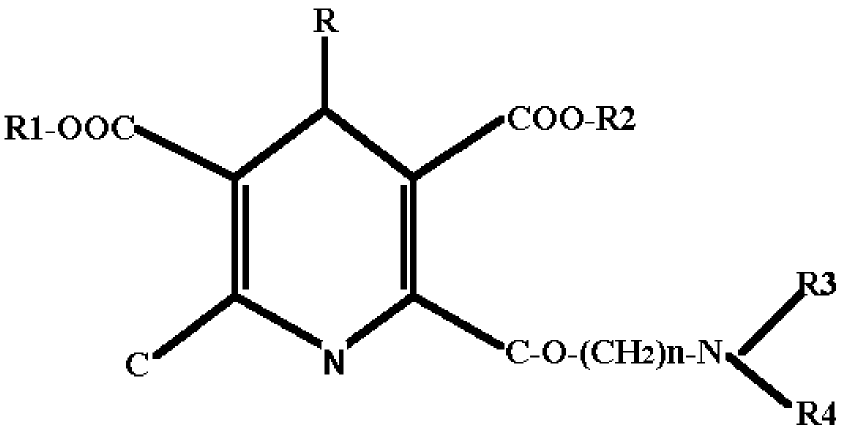






Markush structures: I

Chemical patents are an important source of chemical information



R = 2-chlorophenyl or 2,3-dichlorophenyl

R1 = CH_3

R2 = C_2H_5

N = 2

R3 = H or CH_3

R4 = C-O-R5 or C-S-R6 or S-O-R7

R5 = H or $NHCH_3$ or $NHCH_2CONH_2$ or 2-pyridon-5-yl

R6 = NH_2 or $C(=NHCN)NHCH_3$

R7 = NH_2 or $NHCH_3$ or NH-cyclopentyl or 2-thienyl

or 8-quinolyl or 2-(4-methypiperazin-1-yl)pyrid-5-yl

Markush structures: II

- This example encodes 192 specific molecules; for many patents, the number is not defined
- M.F. Lynch *et al*. (1981) Computer storage and retrieval of generic chemical structures in patents, Part 1. *Journal of Chemical Information and Computer Sciences,* **21**, 148-150.
- Extension of fingerprint and graph matching methods for specifics
- Work in collaboration with Derwent and CAS, resulting in the operational systems Markush DARC (now Merged Markush Service MMS) and MARPAT

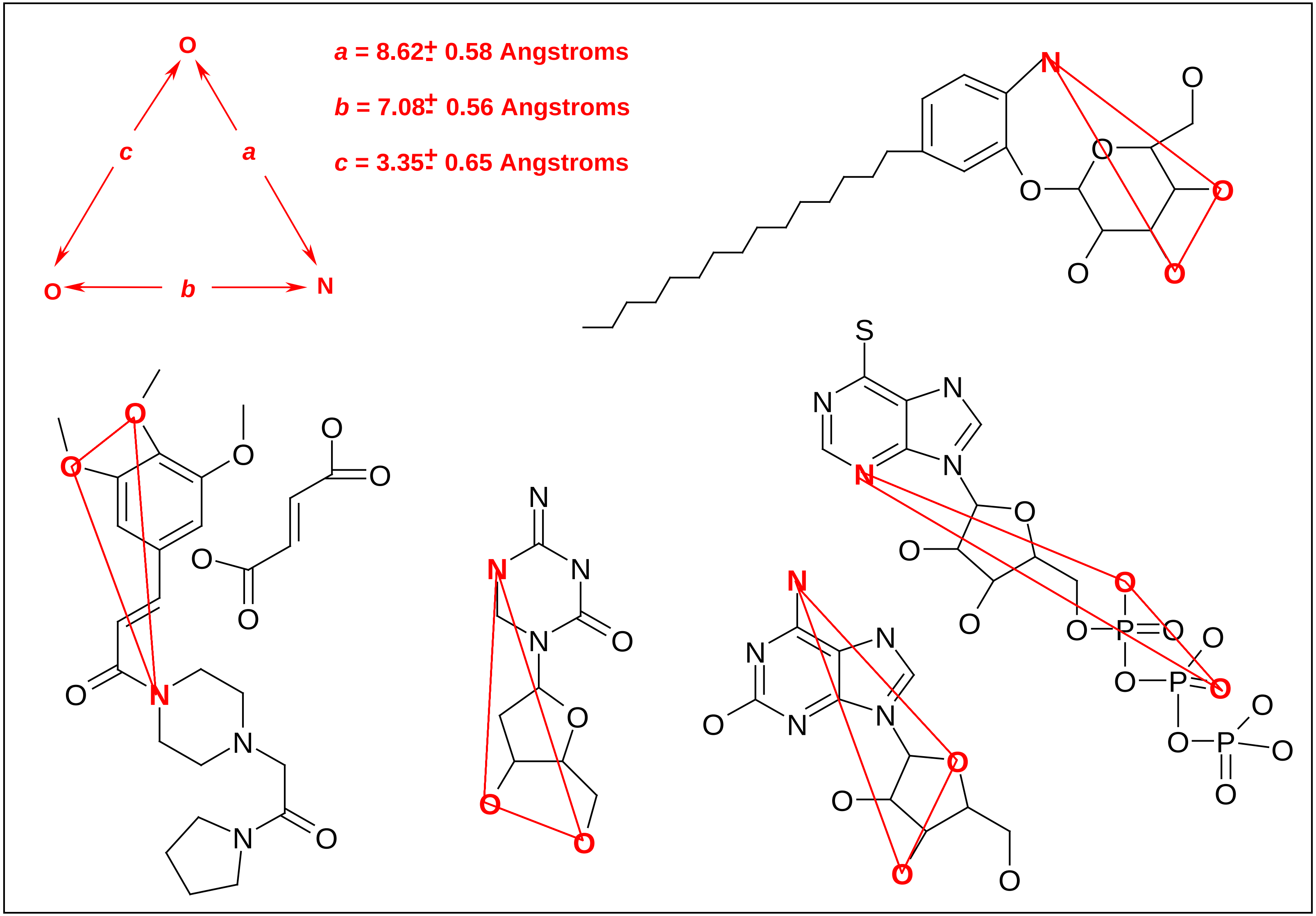
3D substructure searching: I

- P. Gund (1977) Three-dimensional pharmacophoric pattern searching, *Progress in Molecular and Subcellular Biology*, **5**,117-143
- Recognition that the nodes and edges of a graph could represent the atoms and inter-atomic distances (where ‘atom’ may include pharmacophore points, e.g., lone pairs) of a 3D molecule
- But ideas not taken up for a decade:
  - Lack of structural data (except for the Cambridge Structural Database)
  - There was no obvious way of carrying out a search efficiently

3D substructure searching: II

- Intense interest from mid/late Eighties as both problems addressed
- Approximate 3D coordinates from structure-generation programs
  - CONCORD (Pearlman group at Austin, Texas)
  - CORINA (Gasteiger group at Erlangen)
- Fingerprint- and graph-based searching methods
  - S.E. Jakes and P. Willett (1986) Pharmacophoric pattern-matching in files of 3-D chemical structures - selection of interatomic distance screens, *Journal of Molecular Graphics,* **4**, 12-20
  - Basis of first systems at Pfizer and Lederle. Later extensions to encompass conformational flexibility, with industrial systems widely available from the mid-Nineties.

3D substructure search output: searching for pharmacophores



a = 8.62 ± 0.58 Angstroms

b = 7.08 ± 0.56 Angstroms

c = 3.35 ± 0.65 Angstroms

Pharmacophore mapping

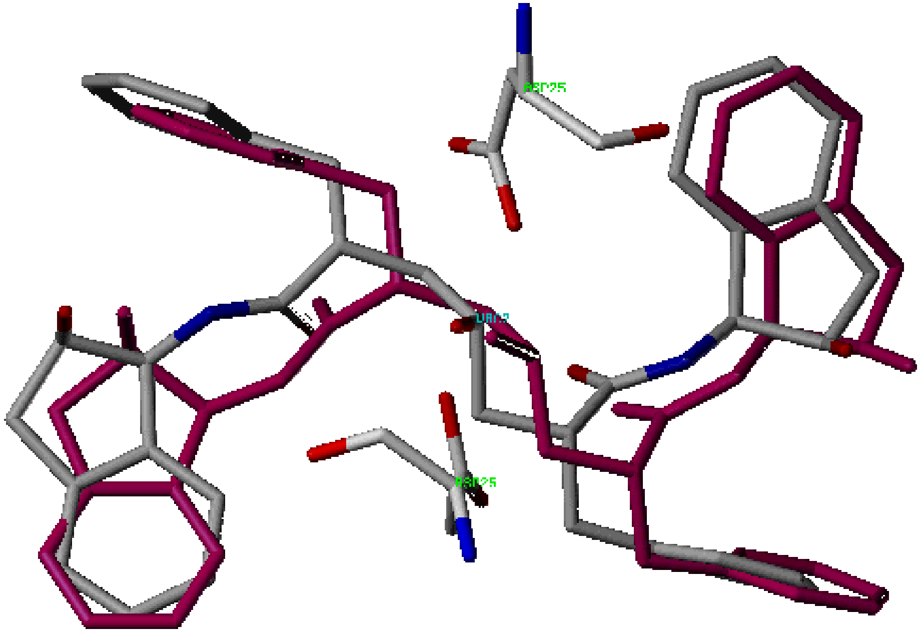
- Given a set of bioactive molecules: what is the common pharmacophore?
- Specify the points
  - G.R. Marshall *et al*. (1979) The conformational parameter in drug design: the active analogue approach in computer-assisted drug design, *ACS Symposium Series*, **112**, 205-226.
- No constraints
  - Y.C. Martin *et al*. (1993) A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists, *Journal of Computer-Aided Molecular Design*, **7**, 83-102.
- Inclusion of flexibility
  - G. Jones *et al*. (1995) A genetic algorithm for flexible molecular overlay and pharmacophore detection, *Journal of Computer-Aided Molecular Design*, **9**, 532-549.

Ligand docking: I

- Fitting a molecule into a binding site
  - “Lock and key” model
- Two-part problem
  - Search algorithm to investigate possible poses
  - Scoring function to prioritise poses/molecules
- I.D. Kuntz *et al*. (1982) A geometric approach to macromolecule-ligand interactions, *Journal of Molecular Biology*, **161**, 269-288
- The DOCK program for fitting an individual molecule into an active site

Ligand docking: II

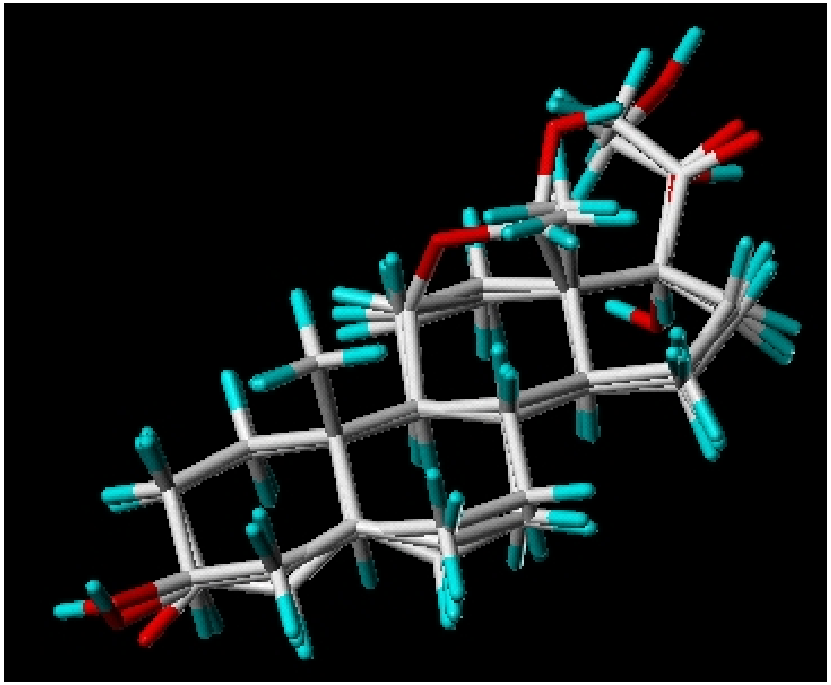
- Extensions for
  - Scanning an entire database, taking each molecule in turn
  - Including ligand flexibility: G. Jones et al. (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation". *Journal of Molecular Biology*, **245**, 43-53.
- Now a standard technique for virtual screening



4PHV docked (red) into HIV protease

Comparative Molecular Field Analysis (CoMFA)

- R.D. Cramer *et al*. (1988) Comparative Molecular-Field Analysis (CoMFA). 1. *Journal of the American Chemical Society,* **110**, 5959-5967.
- Extension of Free-Wilson to 3D
  - Align a set of molecules, and place them in a 3D grid
  - Treat the computed interactions at each grid point as a variable for PLS analysis
- Now the standard QSAR tool



Molecular diversity analysis: I

- Technological developments in the early Nineties led to a data explosion in the volumes of chemical and biological data
- Many more compounds **could** be made: which **should** be made?
- Need for tools to:
  - Quantify diversity
  - Select molecules so as to maximise diversity (NB not a new question, e.g., early Pfizer/Upjohn clustering work)

Molecular diversity analysis: II

- Huge range of papers, focussing on fingerprint-based similarity approaches
  - E.J. Martin *et al*. (1995) Measuring diversity - experimental-design of combinatorial libraries for drug discovery, *Journal of Medicinal Chemistry,* **38**, 1431-1436.
  - R.D. Brown and Y.C. Martin (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, *Journal of Chemical Information and Computer Sciences,* **36**, 572-584.

Diversity alone is not enough

- It soon became clear that many of the molecules being generated had poor ADME characteristics
- ADME traditionally studied during optimisation
    - “Fail fast” paradigm implies that such molecules should be filtered out as early as possible
- C.A. Lipinski *et al*. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews,* **23**, 3-25
    - Criteria for oral activity: ideally, not more than 5 donors or 10 acceptors, MW under 500 and logP under 5
- Idea of drugability or drug-likeness

Conclusions

- Chemoinformatics integrates long-established research into structure searching and bioactivity prediction
- Ever-increasing demands on the pharmaceutical industry will make it even more important in the future, e.g.
  - ADMETox prediction, Chemogenomics, and Virtual screening
- Histories
  - W.L. Chen (2006) “Chemoinformatics: past, present and future” *Journal of Chemical Information and Modeling*, **46**, 2230-2255
  - J. Gasteiger (2006) “Chemoinformatics: a new field with a long tradition” *Analytical and Bioanalytical Chemistry*, **384**, 57-64
  - A.G. Maldonado *et al*. (2006) “Molecular similarity and diversity in chemoinformatics: from theory to applications” *Molecular Diversity*, **10**, 39-79
  - P. Willett (2008) “From chemical documentation to chemoinformatics: fifty years of chemical information science.” *Journal of Information Science*, **34**, 477-499