The University Of Sheffield.

# Chemoinformatics: historical development of database methods

Peter Willett, University of Sheffield

Presented at “Celebrating the History of Chemical Information” 29th November 2010

The University Of Sheffield.

# Overview

- Introduction to chemoinformatics
  - What it is
  - How it has developed
- Historically important papers
  - A personal choice
  - Roughly chronological ordering
  - Focus on searching, with many omissions (QSAR, modelling)

The University Of Sheffield.

# 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."

The University Of Sheffield.

# Emergence of chemoinformatics: I

- Chemoinformatics is not new
  - M. Hann and R. Green (1999) Chemoinformatics - a new name for an old problem? *Current Opinion in Chemical Biology*, **3**, 379-383
  - J. Gasteiger (2006) Chemoinformatics: a new field with a long tradition. *Analytical and Bioanalytical Chemistry*, **384**, 57-64.
- Current interest driven by the data explosion resulting from the introduction of combinatorial synthesis and high-throughput screening in the Nineties
- Focus on chemical structures (both 2D and 3D)
  - Cf bioinformatics and GIS

The University Of Sheffield.

# Emergence of chemoinformatics: II

- First appearance of the core journal, *Journal of Chemical Documentation*, in 1961
  - 1975 *Journal of Chemical Information and Computer Sciences*
  - 2005 *Journal of Chemical Information and Modeling*
- First book on the subject appeared in 1971
  - M.F. Lynch et al., *Computer Handling of Chemical Structure Information*
- 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*
  - Currently 13 such books in Amazon, and another 5 with “cheminformatics”

The University Of Sheffield.

# Emergence of chemoinformatics: III

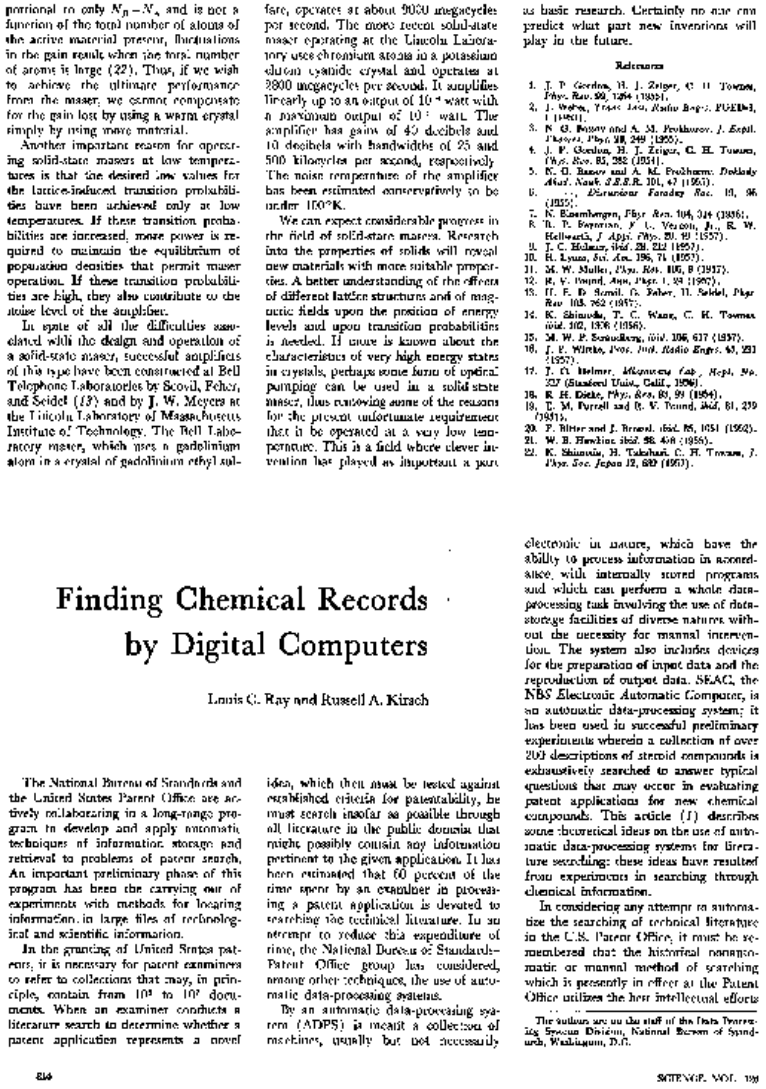
- The first international conference at Noordwijkerhout in 1973, and every three years since 1987
  - Sheffield conferences and regular sessions at ACS national meetings
- Introduction of first full university courses in 2001
  - D.J. Wild and G. Wiggins (2006) Challenges for chemoinformatics education in drug discovery. *Drug Discovery Today*, **11**, 436-439
- Nomenclature
  - Chemical informatics, chemical information (management/science), cheminformatics

The University Of Sheffield.

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 $N_2 - N_1$ 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 sul-

fate, 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 $10^{-9}$ watt with a maximum output of $10^{-5}$ 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 100°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 endowing 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 an 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. Inst. Radio Engrs. PGED-3*, 1 (1953).
3. N. G. Basov and A. M. Prokhorov, *J. Expl. 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. L. Sproull, *J. Appl. Phys.* **27**, 89 (1956).
9. J. C. Helmer, *ibid.* **28**, 212 (1957).
10. H. Lyons, *Sci. Am.* **196**, 74 (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., Hepl. No. 237* (Stanford Univ., Calif., 1956).
18. R. H. Dicke, *Phys. Rev.* **89**, 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**, 408 (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.

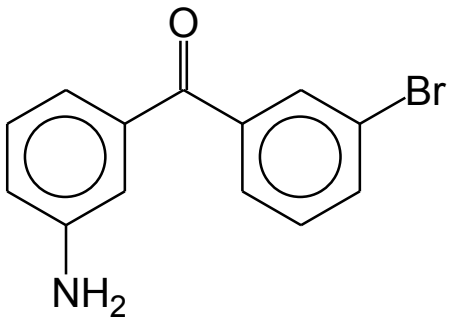
In considering any attempt to automatize the searching of technical literature in the U.S. Patent Office, it must be remembered that the historical nonautomatic or manual method of searching which is presently in effect at the Patent Office utilizes the best intellectual efforts

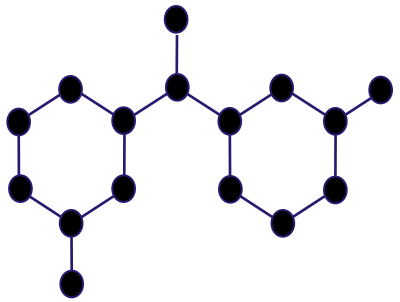

The authors are on the staff of the Data Processing Systems Division, National Bureau of Standards, Washington, D.C.

The University Of Sheffield.

# 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
  - 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



The University Of Sheffield.

# The Morgan algorithm

- Throughout the early Sixties, Chemical Abstracts Service received very substantial funding to develop methods for textual and chemical processing
- Principal result was the CAS Registry System (now contains ca. 55M molecules)
- A graph-based approach based on the Morgan algorithm for systematic naming of chemical graphs
  - H.L. Morgan (1965) Generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service, *Journal of Chemical Documentation*, **5**, 107-113
  - An important component of many structure-matching procedures

The University Of Sheffield.

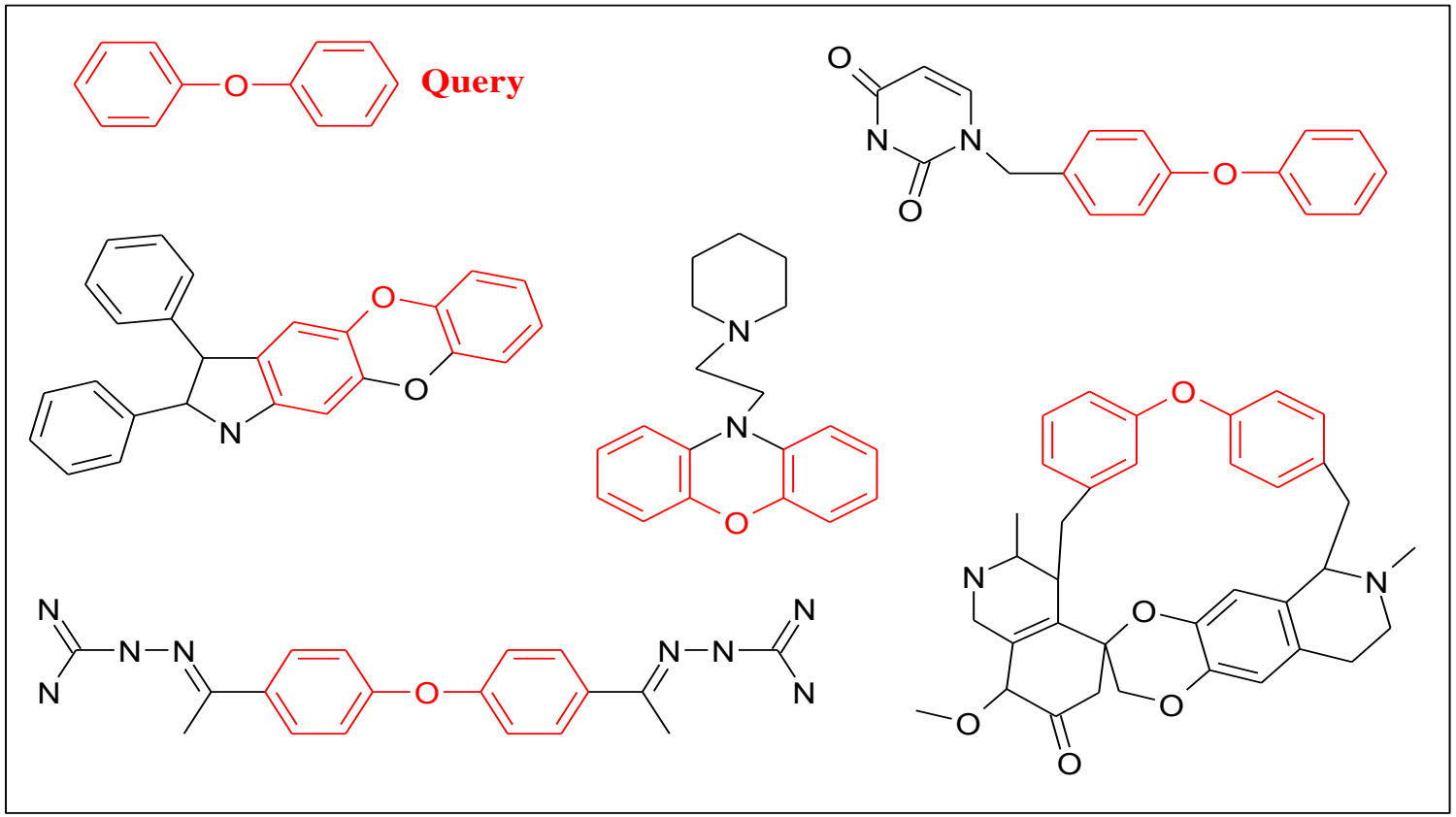
# Wiswesswer Line Notation (WLN)

- Alphanumeric string describing a molecule's topology, albeit implicitly
- Complex coding rules, but the basis for most industrial systems (and printed indices) though out the Sixties and early Seventies
- Need to make information explicit for structure display and precise substructure searching first studied in the CROSSBOW project
  - E. Hyde *et al*. (1967) Conversion of Wiswesser notation to a connectivity matrix for organic compounds, *Journal of Chemical Documentation,* **7**, 200-204

The University Of Sheffield.

# Substructure search: I

Ability to retrieve all molecules in a database containing a user-defined substructure

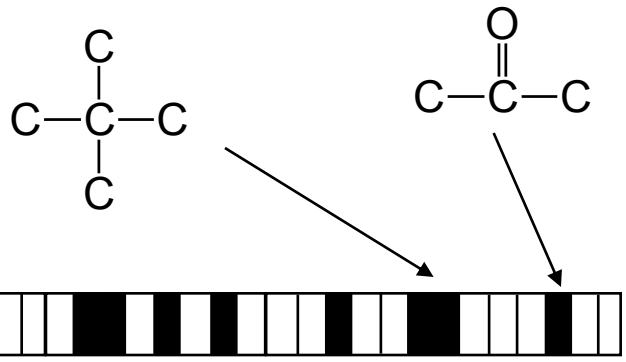


The University Of Sheffield.

# Substructure search: II

- Graph isomorphism algorithm to look for complete structures: check for identity
- Subgraph isomorphism algorithm to look for partial structures: check for inclusion
  - 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*
  - Cf keywords indexing textual documents

The University Of Sheffield.



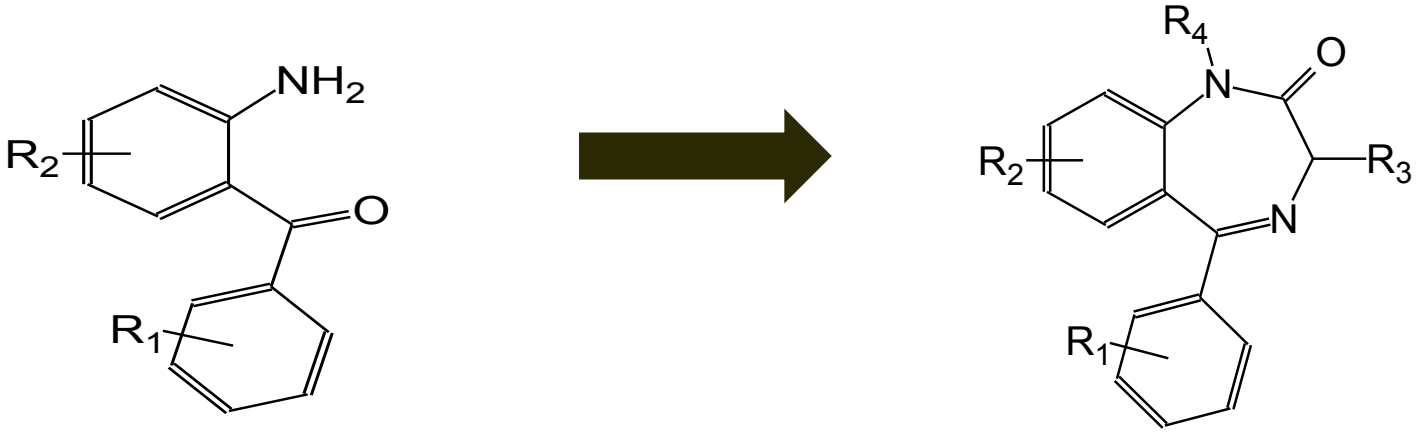

# Binary vector

- 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.

The University Of Sheffield.

# Reaction databases

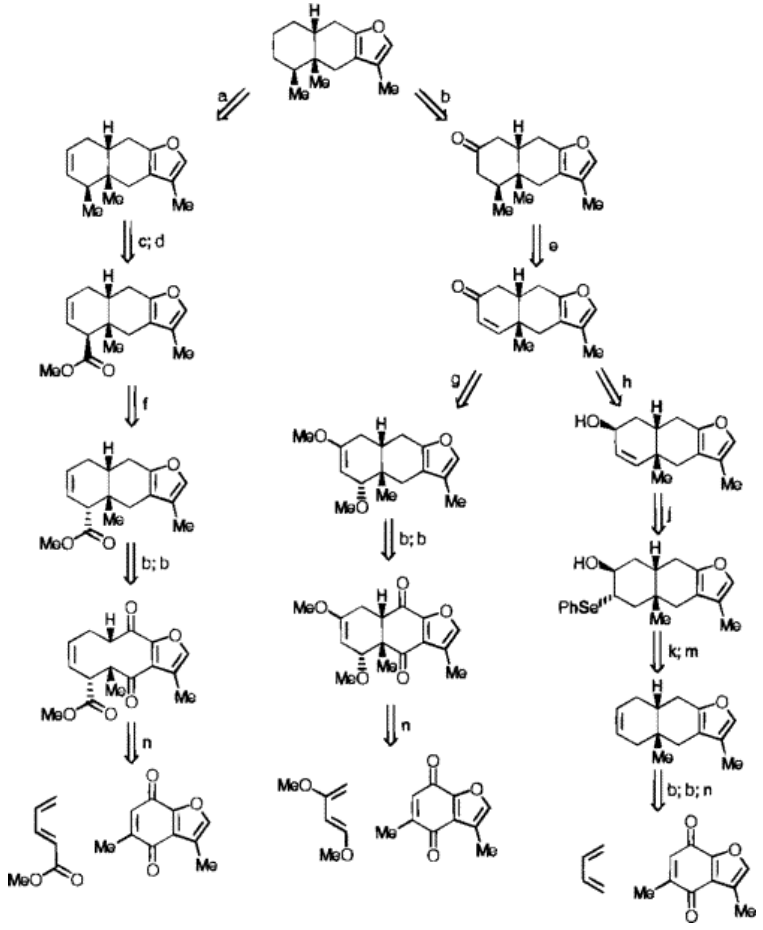
- How to search for structural changes occurring in a reaction?
- G.E. Vleduts (1963) Concerning one system of classification and codification of organic reactions, *Information Storage and Retrieval* **1**, 17-146
  - Index a reaction by just those parts that have changed, the *reaction centre*, to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



The University Of Sheffield.

# Computer-aided synthesis design (CASD)

- Vleduts also the first to suggest computer-aided synthesis design
- "Retrosynthesis": Potential syntheses of a target molecule using a reactions database plus appropriate inference mechanisms
  - CASD programs can also work in the synthetic direction
- First implemented in OCSS (subsequently LHASA)
  - E.J. Corey and W.T. Wipke, (1969) Computer-assisted design of complex organic syntheses, *Science*, **166**, 178-193
- An early example of an expert system (AI), as was computer-aided structure elucidation.



The University Of Sheffield.

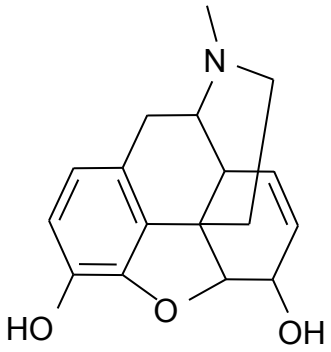
# Moving on

- Throughout the Seventies, chemical search systems (mainly based on Wiswesser Line Notation) became widely available across the pharmaceutical industry
- Computer hardware/software limitations meant processing slow
- Things did not change much till the late-Seventies/early-Eighties, e.g., advent of MDL and CAS Online
- Then new wave of developments

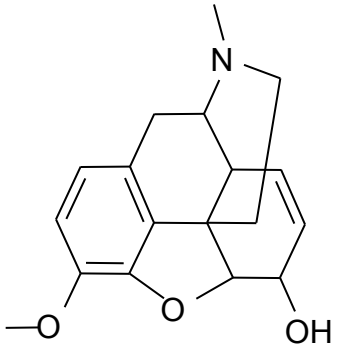
The University Of Sheffield.

# Similarity searching

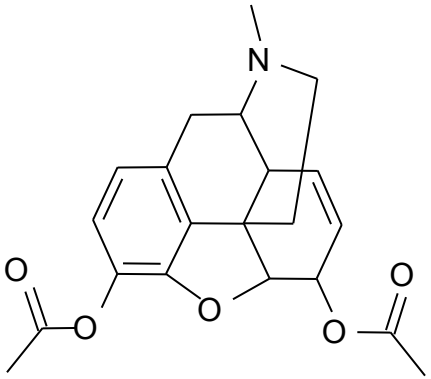
- Substructure searching very powerful but requires a clear view of the types of structures of interest
- Given a *target* (or *reference*) structure find molecules in a database that are most similar to it ("give me ten more like this")
- Rational is the *similar property principle*, which states that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




Morphine




Codeine




Heroin

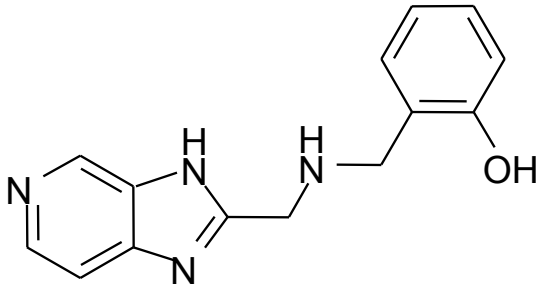
The University Of Sheffield.

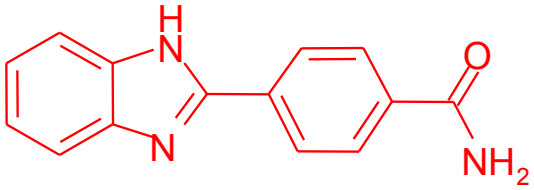
# How to define chemical similarity?

- 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 implementation had to wait till mid-Eighties with systems at Lederle, Pfizer/Sheffield and Upjohn

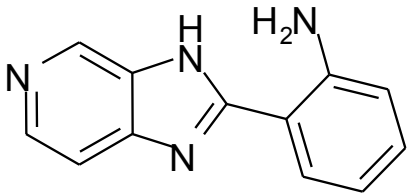
The University Of Sheffield.

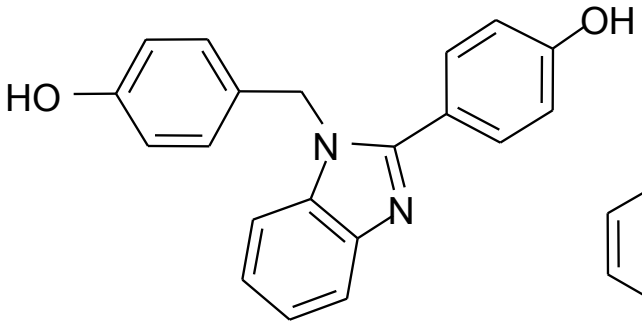
# Tanimoto-based 2D similarity searching

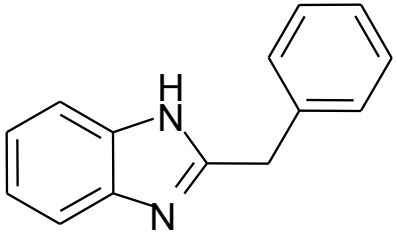


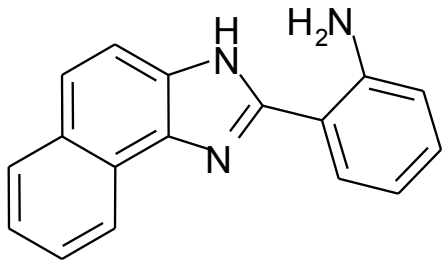




**Query**



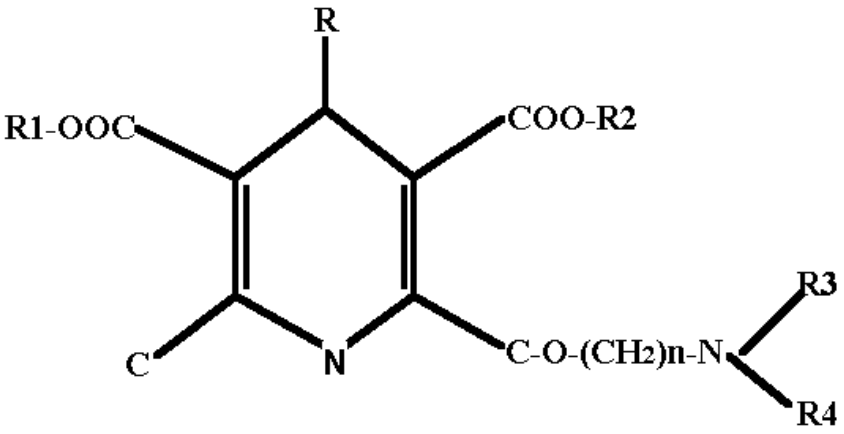



# Markush structures: I

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

The University Of Sheffield.

# 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 MMS) and MARPAT

The University Of Sheffield.

# 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

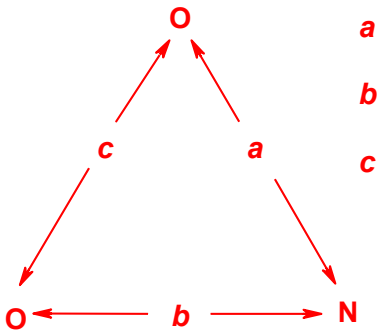
The University Of Sheffield.

# 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)
- 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.

The University Of Sheffield.

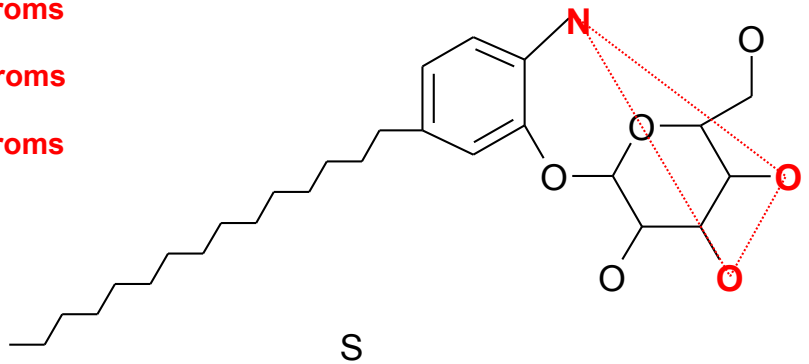
# 3D substructure search output: searching for pharmacophores

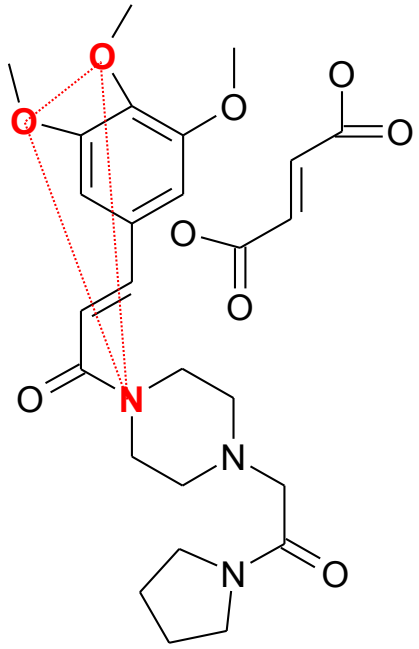


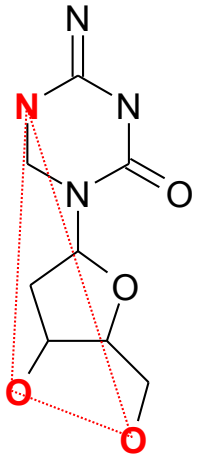

*a* = 8.62± 0.58 Angstroms

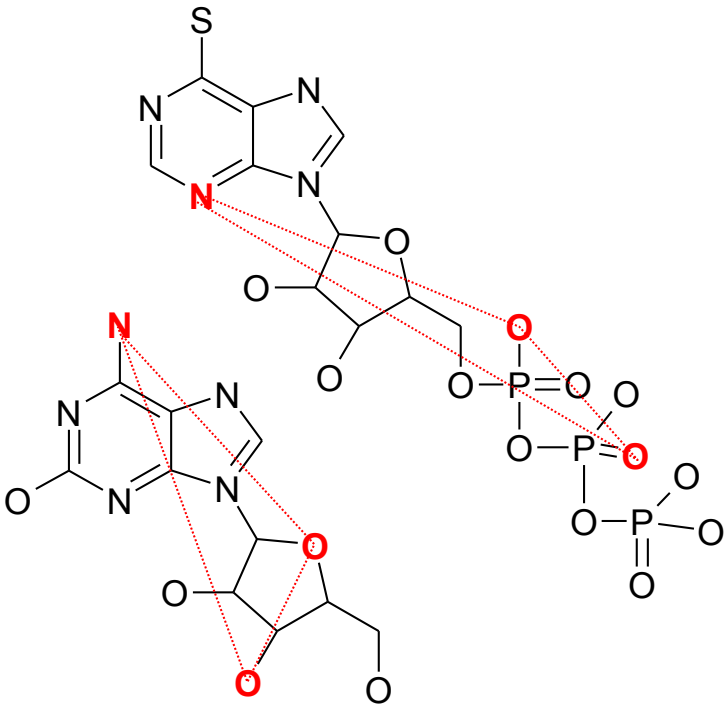
*b* = 7.08± 0.56 Angstroms

*c* = 3.35± 0.65 Angstroms









The University Of Sheffield.

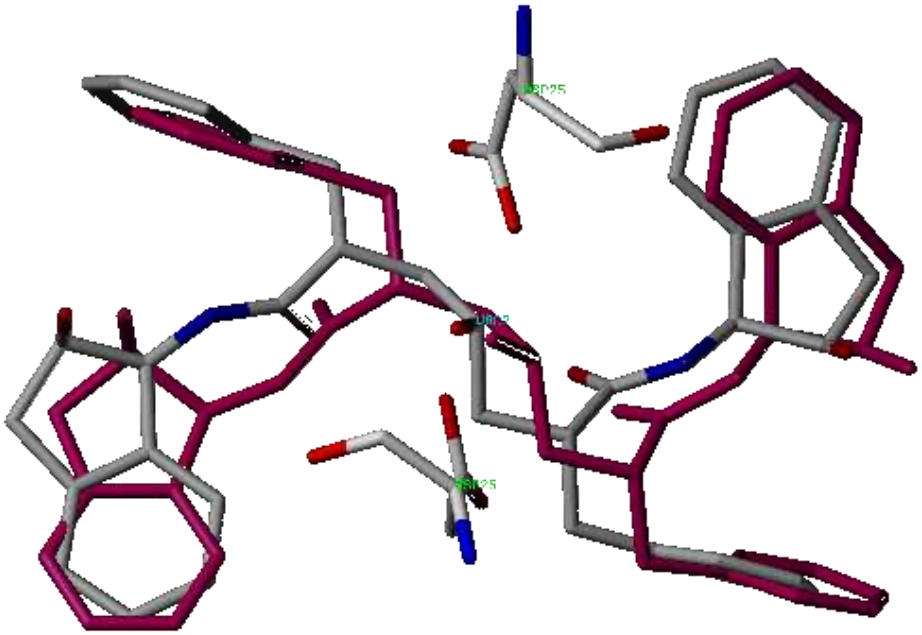
# Ligand docking: I

- 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

The University Of Sheffield.

# 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

The University Of Sheffield.

# Molecular diversity analysis

- Technological developments in the early Nineties meant that many more compounds could be made
  - Which should be made?
- Need for tools to quantify diversity and to select molecules so as to maximise diversity
- 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.

The University Of Sheffield.

# 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

The University Of Sheffield.

# Conclusions

- Chemoinformatics is NOT new
- What is new is the widespread recognition of its importance, and this will increase further given the current challenges facing the pharmaceutical industry
- Histories
  - W.L. Chen (2006), Chemoinformatics: past, present and future. *Journal of Chemical Information and Modeling*, **46**, 2230-2255
  - P. Willett (2008) From chemical documentation to chemoinformatics: fifty years of chemical information science. *Journal of Information Science*, **34**, 477-499
  - R. Al Jishi and P. Willett (2010) The *Journal of Chemical Documentation* and the *Journal of Chemical Information and Computer Sciences*: Publication and citation statistics. *Journal of Chemical Information and Modeling*, **50**, 1915-1923