



The
University
Of
Sheffield.

Chemoinformatics and information management

Peter Willett, University of Sheffield, UK



Overview

- What is chemoinformatics and why is it necessary
- Managing structural information
- Typical facilities in chemoinformatics software
- Examples of current research



Drug discovery: I

- Drug discovery is a vastly complex, multi-disciplinary task that can extend over two decades
- The total cost for the discovery and development of a novel therapeutic agent is now ca. \$1.5B
- Even so, only about 1 in 3 cover the R&D costs
 - But when they can do the pay-offs can be massive: Lipitor in 2006 made \$12.5B (cf MS Windows and Boeing 747)
- Patent cover is 20 years from initial announcement
 - Time is money so need to find potential drugs (and to reject non-drugs) much faster (and similarly for agrochemicals)



Drug discovery: II

- Chemoinformatics is one way of increasing the cost effectiveness of drug discovery
- Initial work in chemoinformatics as early as the Sixties: current interest because of developments in
 - Combinatorial chemistry
 - High throughput screening (HTS)
 - Change from sequential to massively parallel processing
- Resulting explosion in the amounts of data available in drug-discovery programmes, and an increased interest in computational methods
 - Focus on chemical structure diagram, cf development of other types of *-informatics* specialisms

Definitions

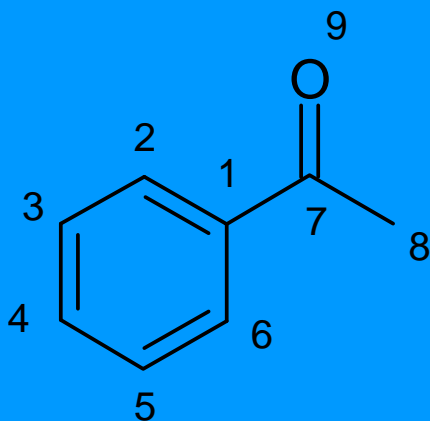
- F.K. Brown (1998). *Annual Reports in Medicinal Chemistry*, 33, 375-384
 - “The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into 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.”



Representation of molecules

- Need for a machine-readable representation
 - 1D – computed/experimental global properties
 - 2D – the chemical structure diagram
 - 3D – atomic coordinate data
- 1D representations handled using conventional DBMS software
- Need to manipulate 2D and 3D data

Connection tables

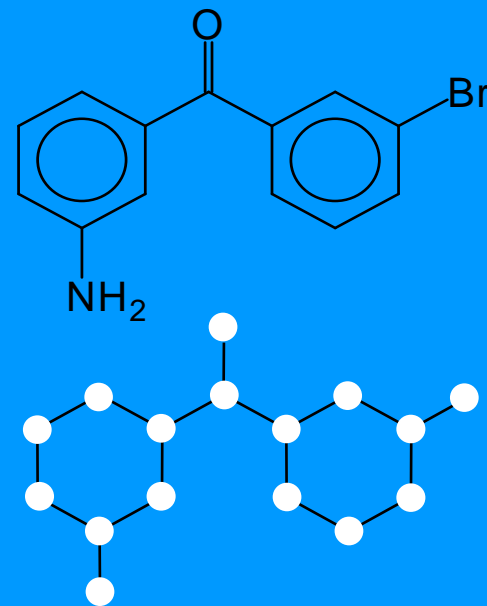


1	C	2	2	6	1	7	1
2	C	1	2	3	1		
3	C	2	1	4	2		
4	C	3	2	5	1		
5	C	4	1	6	2		
6	C	1	1	5	2		
7	C	1	1	8	1	9	2
8	C	7	1				
9	O	7	2				

- An unambiguous representation of a 2D chemical structure diagram
- A connection table is a graph, the underlying data structure in chemoinformatics

Graph theory and chemistry

- Graph theory
 - Branch of mathematics that describes sets of objects, called *nodes* and the relationships between them, called *edges*
- A 2D connection table is a graph:
 - Nodes correspond to atoms
 - Edges correspond to bonds
- Graph matching algorithms
 - Search chemical databases
- Generation of other representations



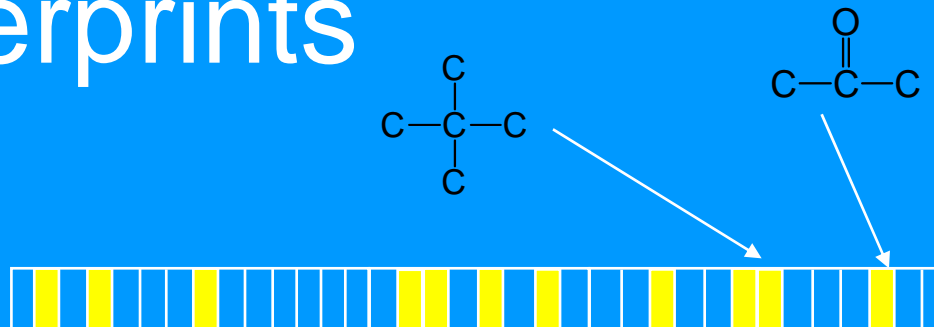


Types of search

- Exact structure search (hashed connection table with graph isomorphism for collision handling)
- Substructure search (subgraph isomorphism)
 - cf partial or boolean matching in text
- Similarity searching (maximal common subgraph isomorphism (or simpler))
 - cf best match search or web searching
- Graph matching algorithms are effective
 - But time is factorial with the number of nodes
 - Need for efficient heuristics



Fingerprints



- A fingerprint (or fragment bit-string) is a binary vector encoding the presence (“1”) or absence (“0”) of fragment substructures in a molecule
- Each bit in the fingerprint represents one molecular fragment. Typical length is ~1000 bits
- An approximate representation, but one that can be processed very efficiently and hence often used as a precursor to graph matching

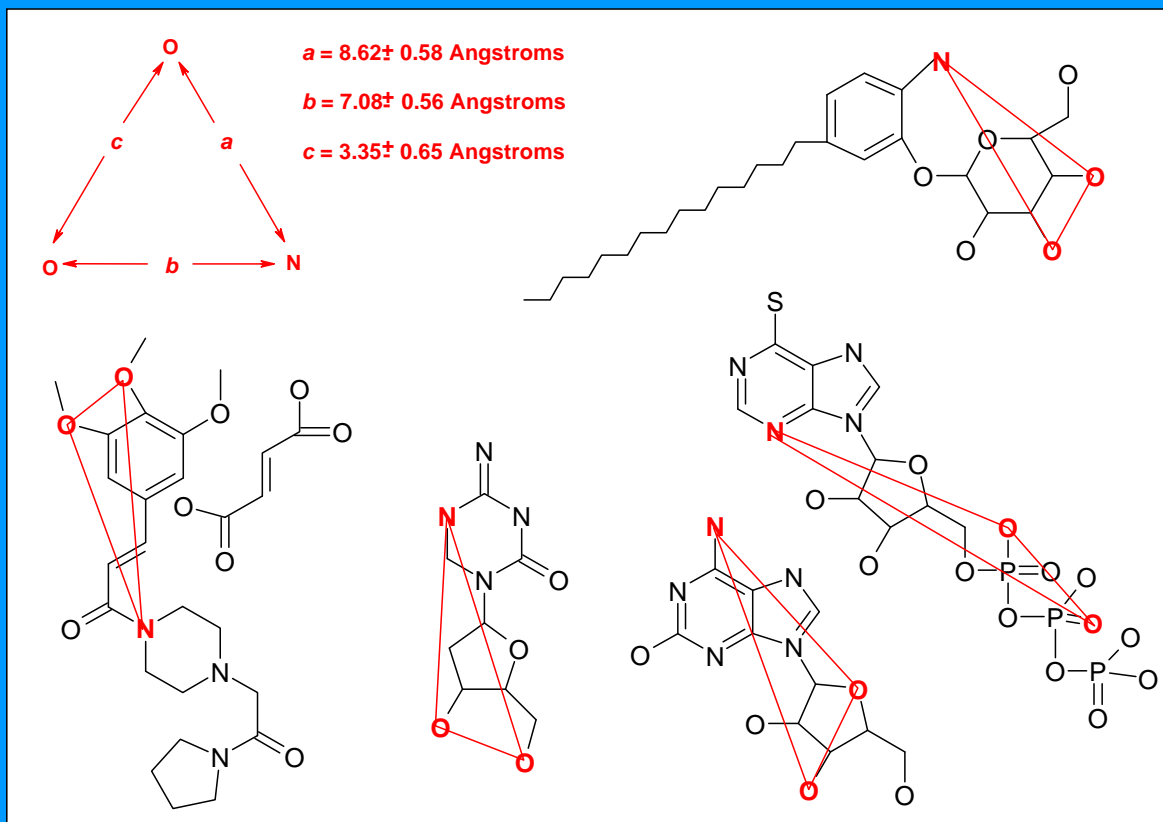


Chemoinformatics facilities

- Database searching as described previously
 - Structure and substructure searching originally
 - Similarity searching from mid-Eighties
 - 3D substructure searching from mid-Nineties (first rigid then flexible)
- Applications
 - Database clustering
 - Molecular diversity analysis
 - Drug-likeness
 - Virtual screening
 - Ligand-based
 - Structure-based

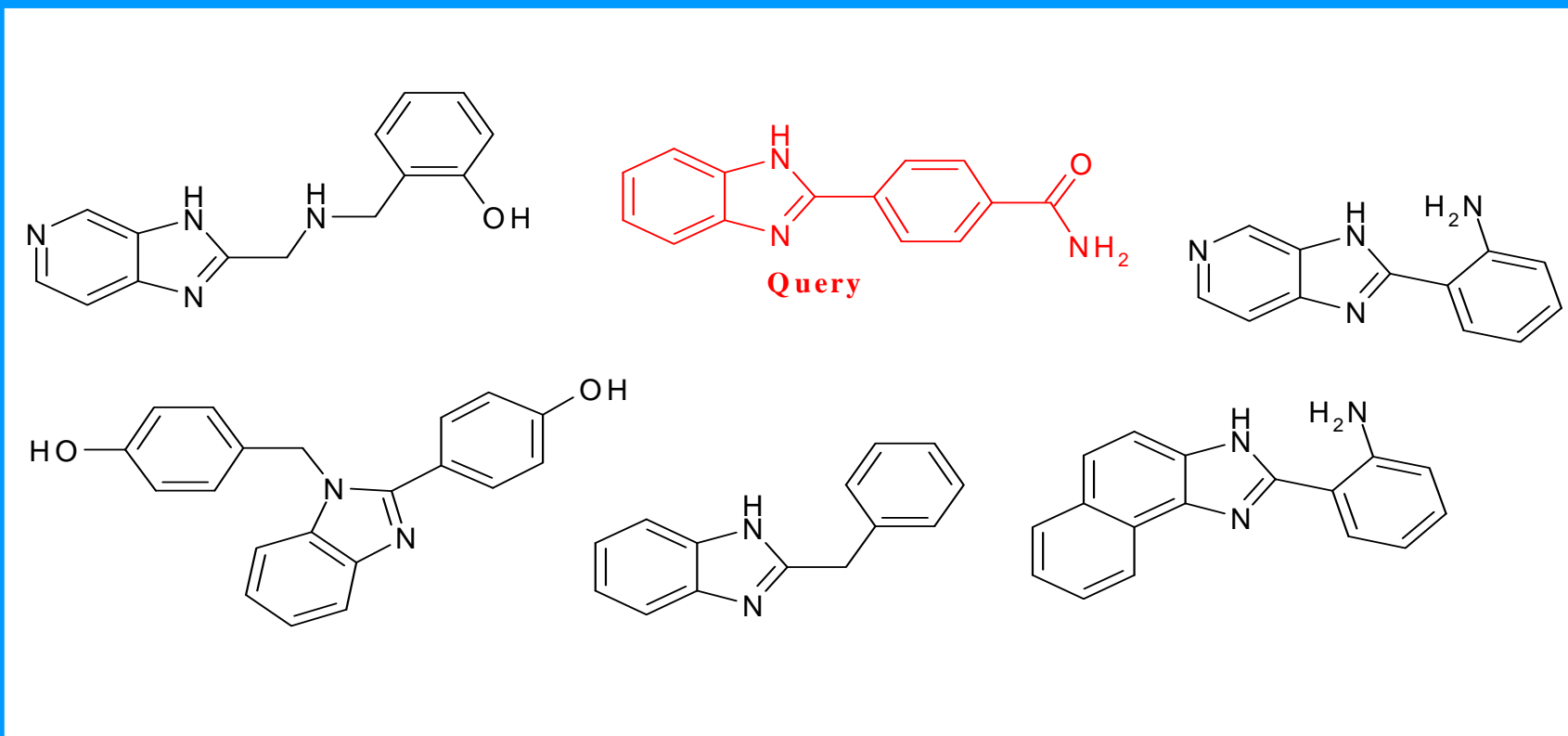
3D substructure searching

- Generation of pharmacophore patterns
- Use of MOGA and hyperstructure approaches



Similarity searching using 2D fingerprints

Use of data fusion methods to enhance performance,
combining information from multiple searches





Molecular modelling and QSAR

- Use of computational chemistry to obtain the structures and properties of small molecules
 - Quantum mechanics
 - Molecular dynamics
 - Molecular modelling
- Statistical correlation of structure (however described) with physical, chemical and biological properties
 - Initially biological activity (QSAR)
 - Now pharmacokinetics and toxicity (ADMET)



Integration with database searching

- Related, but largely separate, research areas for many years
 - Simple search operations on very large numbers of molecules
 - Increasingly complex operations on smaller and smaller (normally homogeneous) datasets
 - Substructural analysis as an early, notable exception
- The future lies in the integration of these two approaches, applying more sophisticated methods on larger datasets
 - Docking now well established
 - Property calculations at a database level
 - ADMET



General references

J. Gasteiger (ed.), *Handbook of Chemoinformatics* (Wiley-VCH, Weinheim, 2003).

W.L. Chen, Chemoinformatics: past, present and future, *Journal of Chemical Information and Modeling* 46 (2006) 2230-2255.

D.J. Wild and G.D. Wiggins, Challenges for chemoinformatics education in drug discovery, *Drug Discovery Today* 11 (2006) 436-439.

A.R. Leach and V.J. Gillet, *An Introduction to Chemoinformatics* (Kluwer, Dordrecht, 2nd sediton, 2007).

P. Willett, A bibliometric analysis of chemoinformatics, *Aslib Proceedings* 60 (2008) 4-17.