Chemoinformatics: historical development of database methods

Peter Willett, University of Sheffield

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

  - “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”

  - “Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information”

  - “Chemoinformatics is the application of informatics methods to solve chemical problems.”
Emergence of chemoinformatics: I

- Chemoinformatics is not new
- 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
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”
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
- Nomenclature
  - Chemical informatics, chemical information (management/science), cheminformatics
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
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 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
  - An important component of many structure-matching procedures
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
Substructure search: I

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

Reaction databases

- How to search for structural changes occurring in a reaction?
  - 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 (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)
- An early example of an expert system (AI), as was computer-aided structure elucidation.
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
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
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
• How to use this idea?
  • Operational implementation had to wait till mid-Eighties with systems at Lederle, Pfizer/Sheffield and Upjohn
Tanimoto-based 2D similarity searching
Markush structures: I

Chemical patents are an important source of chemical information

R = 2-chlorophenyl or 2,3-dichlorophenyl
R1 = CH₃
R2 = C₂H₅
n = 2
R3 = H or CH₃
R4 = C-O-R5 or C-S-R6 or S-O-R7
R5 = H or NHCH₃ or NHCH₂CONH₂ or 2-pyridon-5-yl
R6 = NH₂ or C(=NHCN)NHCH₃
R7 = NH₂ or NHCH₃ or NH-cyclopentyl or 2-thienyl
or 8-quinolyl or 2-(4-methypiperazin-1-yl)pyrid-5-yl
This example encodes 192 specific molecules; for many patents, the number is not defined.


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.
3D substructure searching: I


- 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)
- Searching methods
  - 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

\[ a = 8.62 \pm 0.58 \text{ Angstroms} \]
\[ b = 7.08 \pm 0.56 \text{ Angstroms} \]
\[ c = 3.35 \pm 0.65 \text{ Angstroms} \]
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


• 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

• Now a standard technique for virtual screening

4PHV docked (red) into HIV protease
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
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
  - 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 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