The University Of Sheffield.

Similarity methods for ligand-based virtual screening

Peter Willett, University of Sheffield

Computers in Scientific Discovery 5, 22nd July 2010

The University Of Sheffield.

Overview

- Molecular similarity and its use in virtual screening
- Use of fragment weighting schemes
- Comparison of fusion rules

The University Of Sheffield.

Chemoinformatics

- 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
  - Agrochemicals and other fine-chemicals
- Chemoinformatics has played an increasingly important role in these developments
  - Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information" (Greg Paris, quoted at http://www.warr.com/warrzone.htm)
  - Particular focus on the manipulation of information about chemical structures (2D or 3D)
- Virtual screening now a key area of study

The University Of Sheffield.

Virtual screening

- Ranking the molecules in a database in order of decreasing probability of activity
  - Focus interest on just those at the top of the ranking
- Range of methods available, varying in the types of information available
  - Use of structure-based methods when an X-ray structure for the biological target is available
  - Use of ligand-based methods when no such information is available

    Database searching a common approach

The University Of Sheffield.

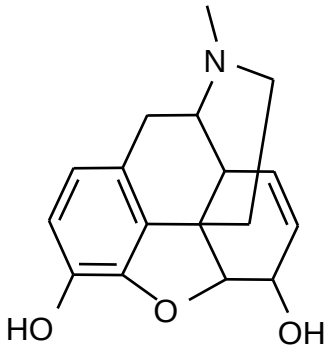
Searching chemical databases

- Three main types of search
  - Structure search
    - "Find me information about this molecule"
  - Substructure search
    - "Find me molecules that contain this partial structure"
  - Similarity search
    - "Find me molecules like this molecule"

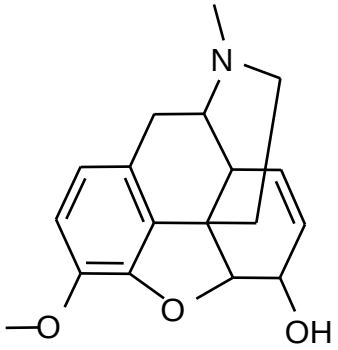
The University Of Sheffield.

Similarity searching

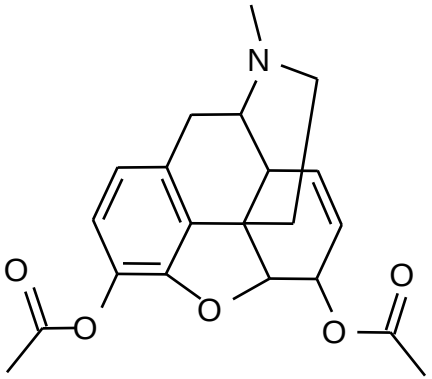
- Substructure searching very powerful but requires a clear view of the types of structures of interest
- Given a *reference* structure find molecules in a database that are most similar to it ("give me ten more like this")
- The *similar property principle* states that structurally similar molecules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



Morphine



Codeine



Heroin

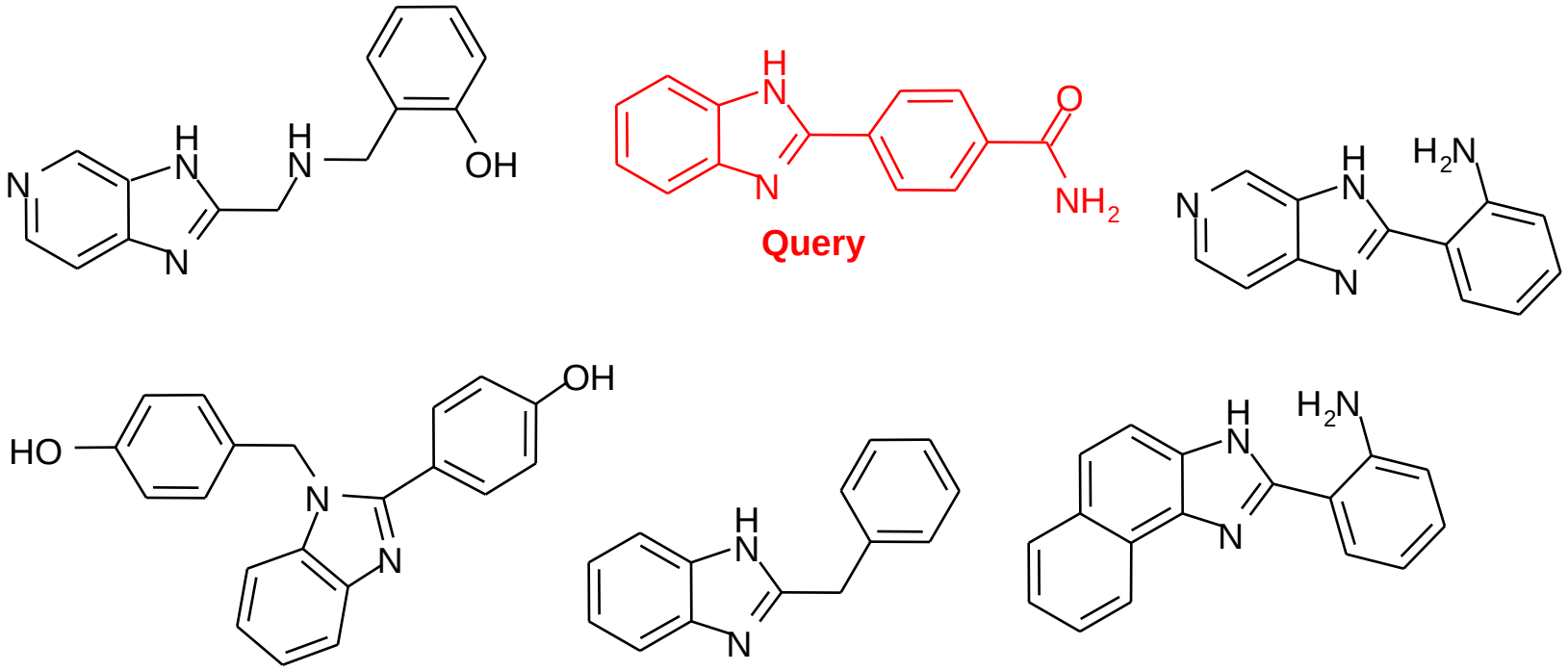
The University Of Sheffield.

How to define chemical similarity?

- Need for a similarity measure
  - A structure representation
  - A weighting scheme
  - A similarity coefficient
- Very many different similarity measures: the most common uses 2D fingerprints and the Tanimoto coefficient
  - First suggested in early Seventies but operational implementations not till mid-Eighties

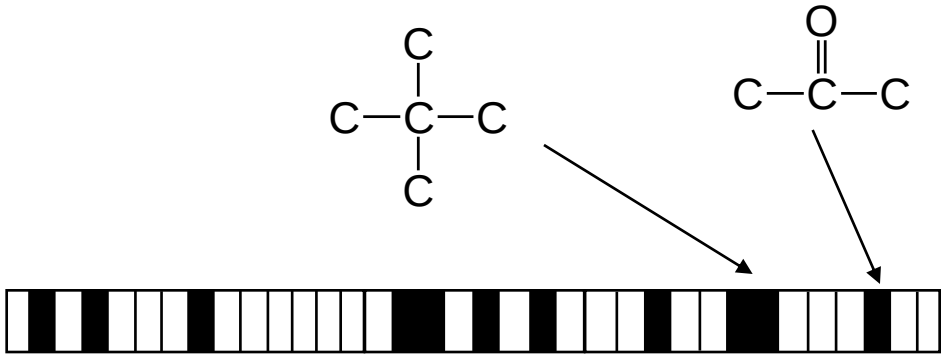
The University Of Sheffield.

Similarity searching with 2D fingerprints and the Tanimoto coefficient



The University Of Sheffield.

Fingerprints



- A simple, but approximate, representation that encodes the presence of fragment substructures 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
- Two fingerprints are regarded as similar if they have many common bits set

The University Of Sheffield.

Tanimoto coefficient for binary bit strings

$$S_{RD} = \frac{C}{R + D - C}$$

- *C* bits set in common between Reference and Database structures
- *R* bits set in Reference structure
- *D* bits set in Database structure
- S_{RD} equal to one (or zero) corresponds to identical fingerprints (or no bits in common)
- More complex form for use with non-binary data, e.g., when one has non-binary fragment weights
- Many other similarity coefficients exist, e.g. cosine coefficient, Euclidean distance, Tversky index

The University Of Sheffield.

Experimental details

- Use of MDDR (ca. 102K structures) and WOMBAT (ca. 130K structures) databases
  - Sets of molecules with known biological activities
  - Molecules represented by various types of fingerprint
- Simulated virtual screening using an active as the reference structure
  - How many of the top-ranked molecules from a similarity search are also active?

The University Of Sheffield.

Use of fingerprint weighting

- Binary fingerprints work well, but can we do better, given additional information?
- Use of frequency information
  - Focus for this work
- Use of activity information
  - Powerful machine learning methods, but need to have many actives and inactives

The University Of Sheffield.

Types of frequency information

- Frequency within a molecule
  - If two molecules have multiple occurrences of a fragment in common then more similar than if just a single occurrence in common
- Frequency within a database
  - If two molecules share a very rare fragment then more similar than if share a very common fragment

Weighting in textual information retrieval

- Weighting of keywords in textual IR
  - Both types of weighting improve performance as compared to simple binary weighting
- Is this also the case in similarity-based virtual screening?
  - Previous studies on small-scale and equivocal results

The University Of Sheffield.

Weighting in chemoinformatics: I

$$S_{RD} = \frac{\sum_{i=1}^{k} R_i D_i}{\sum_{i=1}^{k} R_i^2 + \sum_{i=1}^{k} D_i^2 - \sum_{i=1}^{k} R_i D_i}$$

Experiments show that

- Use of occurrence, rather than incidence, data is generally useful
- Best results using the square root of the occurrence frequencies in both the reference and database structures

The University Of Sheffield.

Weighting in chemoinformatics: II

- For a fragment occurring in *T* of the *N* molecules in a database use the inverse frequency weight log(*N*/*T*)
- Experiments show that:
  - If the actives are closely related then this weight enhances performance over unweighted searching.
  - If the actives are structurally diverse sets then unweighted searching is superior

The University Of Sheffield.

Data fusion

- Originally developed for signal processing but an entirely general approach:
  - Improved performance can be obtained by combining evidence from several different sources
- When used for similarity searching, combine multiple rankings of a database to give a single, fused ranking
  - Similarity fusion

    A single reference structure with multiple similarity measures (e.g., different fingerprints or different similarity coefficients)
  - Group fusion

    A single similarity measure but multiple reference structures
- How to combine different rankings?

The University Of Sheffield.

Fusion rules

- Given multiple input rankings, a fusion rule outputs a single, combined ranking
  - The rankings can be either the computed similarity values or the resulting rank positions
- Previous work has identified use of:
  - CombMAX for similarity data
  - CombSUM for rank data
  - Many others can be used (15 in all here)

The University Of Sheffield.

Fusion rules for the *x*-th database structure

- CombMax = $\max\{S_1(x), S_2(x)..S_i(x)..S_n(x)\}$
  - Also CombMIN
- CombSum = $\sum S_i(x)$
  - Also CombMED and other averages
- CombRKP = $\sum(1/R_i(x))$
  - Can only be used with rank data

The University Of Sheffield.

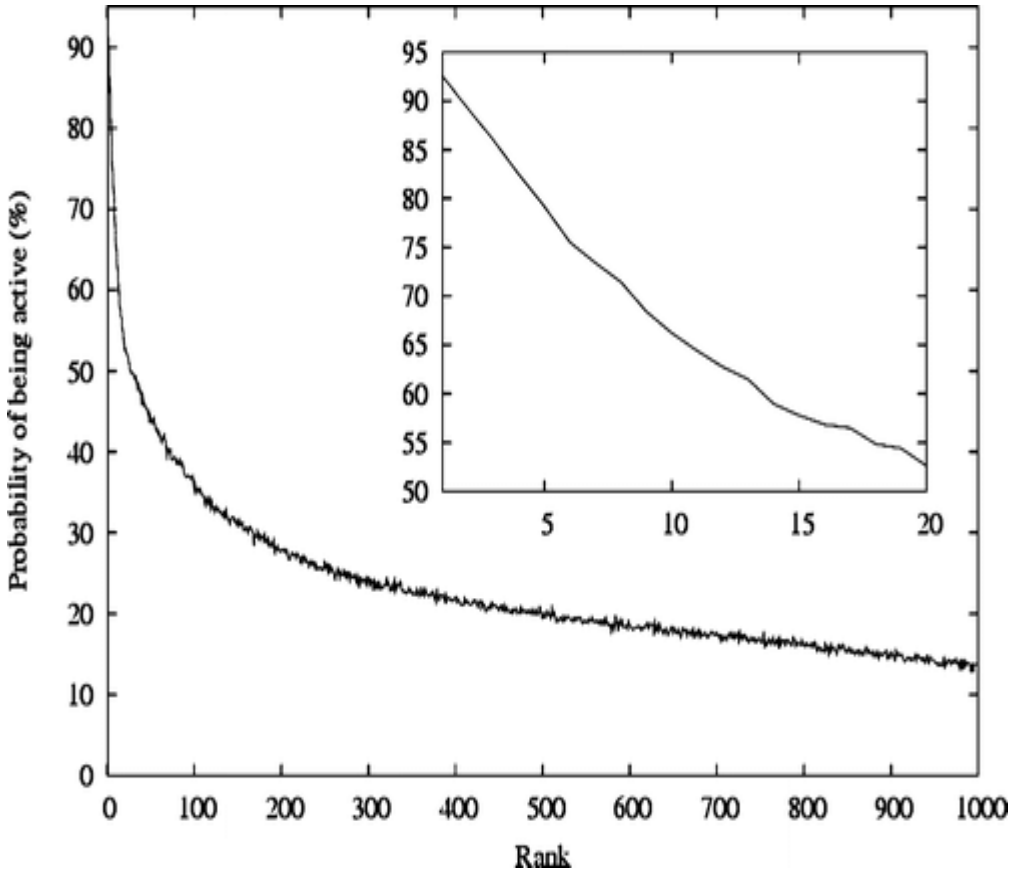
Experimental details

- Searches carried out using
  - Similarity fusion and group fusion
  - Various percentages of the ranked database
  - Different fusion rules
- Results show conclusively that:
  - Use just the top 1-5% of each ranked list
  - Use the CombRKP fusion rule

The University Of Sheffield.

Use of CombRKP: I

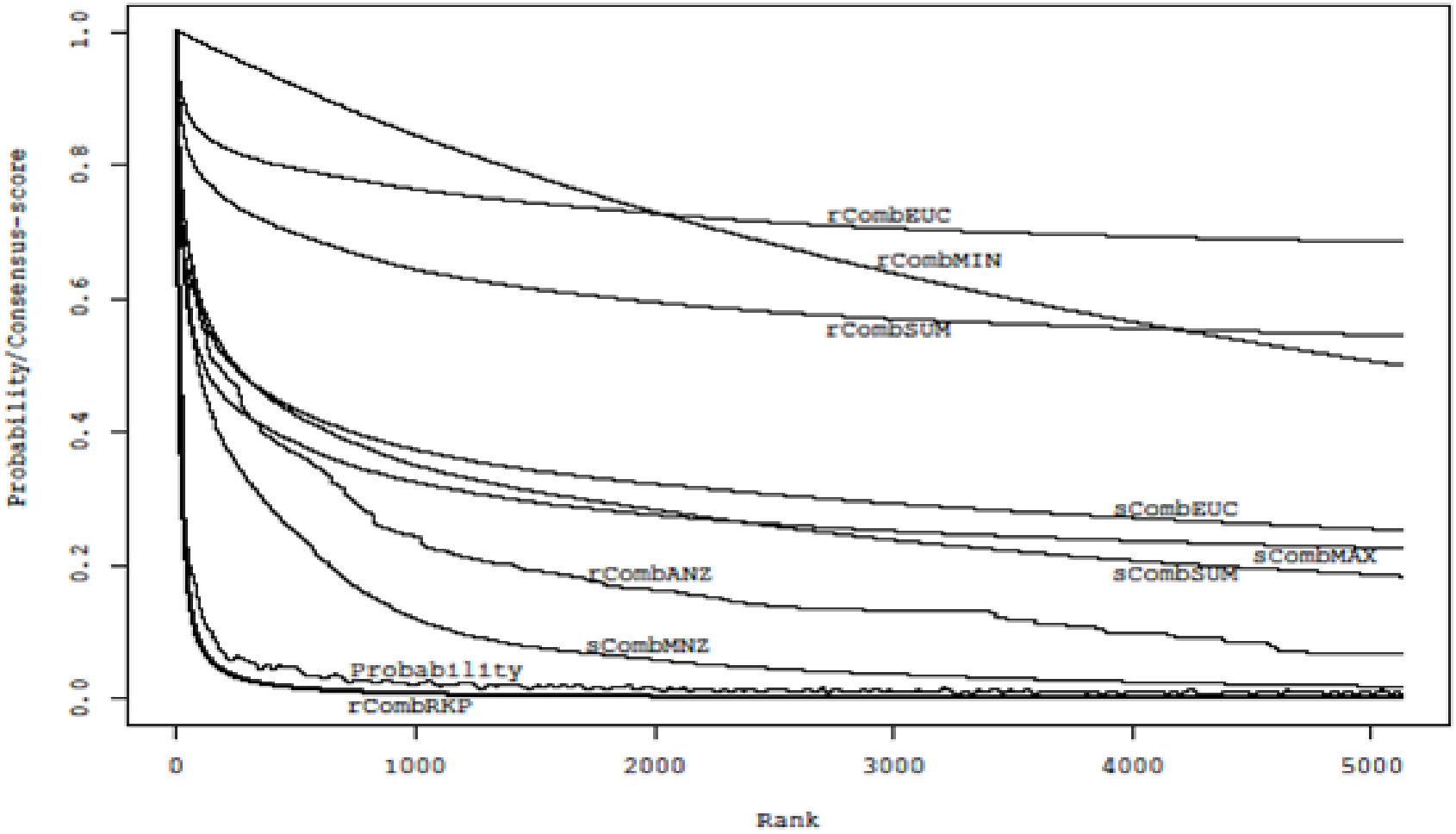
Virtual screening seeks to rank molecules in decreasing order of probability of activity: MDDR searches (*J. Med. Chem.*, **2005**, *48*, 7049) show a hyperbola-like plot



The University Of Sheffield.

Use of CombRKP: II

Probability of activity approximated by (1/Rank), and hence CombRKP likely to perform well



The University Of Sheffield.

Conclusions

- Similarity-based virtual screening using fingerprints well-established
- Can enhance screening effectiveness by:
  - Using fragment occurrence data
  - Combining the rankings from multiple searches using the CombRKP fusion rule

Acknowledgments

- Shereen Arif
- John Holliday
- Christoph Mueller
- Nurul Malim