



The
University
Of
Sheffield.

The Calculation of Molecular Similarity: Principles and Practice

Peter Willett, University of Sheffield

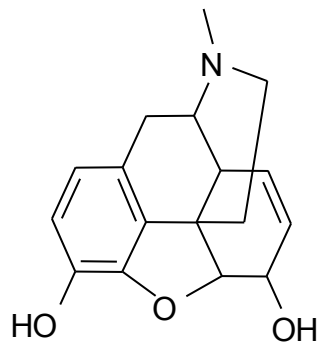
For details, see the full paper in the
Summer School issue of *Molecular Informatics*

Overview

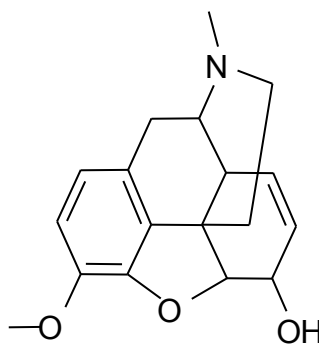
- Principles
 - Why is molecular similarity important?
 - Components of a similarity measure
 - Molecular descriptors
 - Weighting schemes
 - Similarity coefficients
- Practice
 - Similarity searching
 - Cluster analysis and molecular diversity analysis
 - Recent Sheffield applications

Why is molecular similarity important?

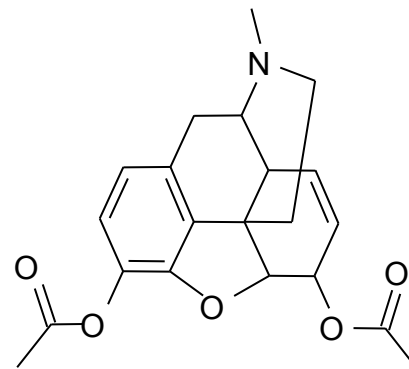
- Much of chemistry is based on structural analogies, and would be very difficult if this were not the case
- More formally, the *similar property principle* states that structurally similar molecules tend to have similar properties



Morphine



Codeine



Heroin

Quantification of similarity

- Note that there are many exceptions to the principle but it is an excellent rule-of-thumb in the absence of more detailed knowledge
- Focus here on chemical similarity, but increasing interest in biological similarity
- People's judgements of similarity are inherently subjective, so need to provide a quantitative basis, a *similarity measure*, for assessing the degree of resemblance
- There is no single measure of similarity



Which two are most similar?



Banana



Orange



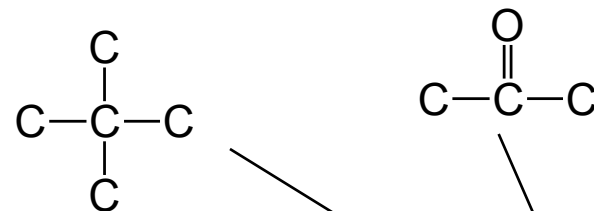
Basketball

Components of a similarity measure

- Molecular descriptors
 - Numerical values assigned to structures
 - 1D properties: MW, logP, PSA etc
 - 2D properties: fingerprints, topological indices, maximum common substructures
 - 3D properties: molecular fields, shape
- Weighting scheme
 - Used to ensure equal (or non-equal) contributions from all parts of the descriptor
- Similarity coefficient
 - A quantitative measure of similarity between two sets of molecular descriptors

Molecular descriptors

- The most intuitive approach is to identify the overlap between the graphs representing a pair of molecules
 - Such maximum common subgraph isomorphism methods are very slow
- Use of 2D fingerprints originally developed for substructure searching as an alternative
 - Binary vectors (or bit-strings) encoding chemical substructures (or fragments)
 - Currently, the standard way of computing molecular similarity (e.g., similarity searching, clustering and diversity analysis)



Binary vector



- Each bit records the presence (“1”) or absence (“0”) of a fragment in the molecule
- Two main ways of creating a fingerprint
 - Dictionary approaches (one-to-one mapping of fragments to bits)
 - Hashing approaches (many-to-many mapping of fragments to bits)
- It is assumed that two fingerprints with many bits in common represent similar parent molecules
- Clearly a very crude measure but surprisingly effective across a wide range of applications

Weighting schemes

- Weighted fingerprints associate a degree of relative importance with each bit in a fingerprint
 - Number of occurrences of a fragment in a molecule
 - Number of occurrences of a fragment in an entire database
- The former approach appears to be more useful, and can be more effective than binary fingerprints
- Much less studied to date than descriptors and coefficients

Similarity coefficients

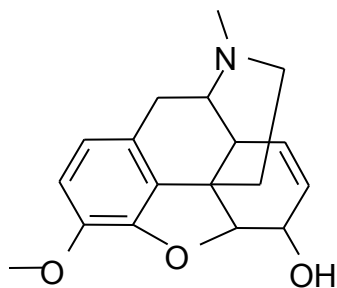
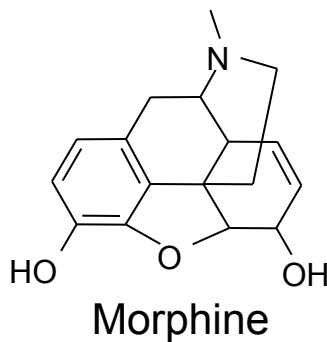
- Tanimoto coefficient for two molecules A and B

$$SIM_{AB} = \frac{c}{a + b - c}$$

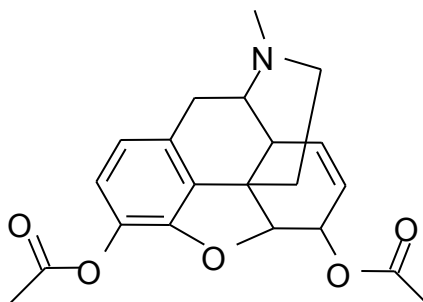
- c bits set in common in the two fingerprints
- a and b bits set in the fingerprints for A and B
- Much more complex form for use with non-binary data, e.g., physicochemical property vectors
- Many, many other types of similarity coefficient exist (e.g., cosine coefficient, Euclidean distance, Tversky index) but fingerprint/Tanimoto measures are the standard

2D encodes just the topologies of molecules

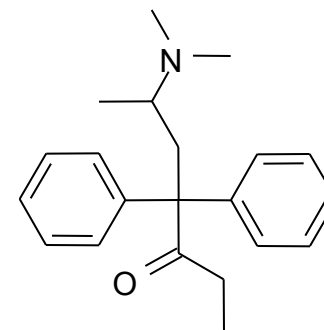
Daylight fingerprints;
Tanimoto similarities



0.99 similar
Codeine

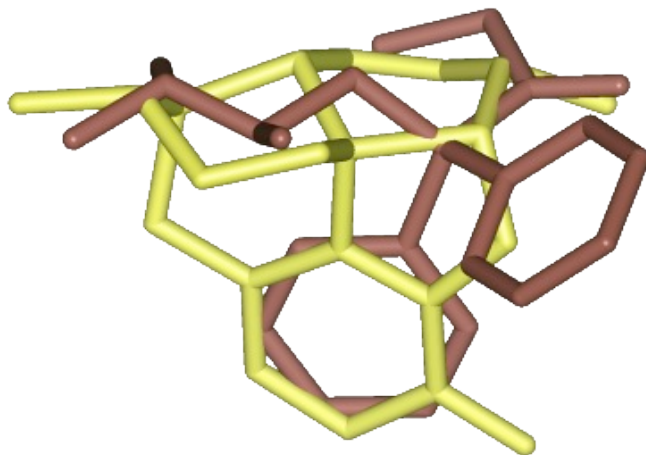
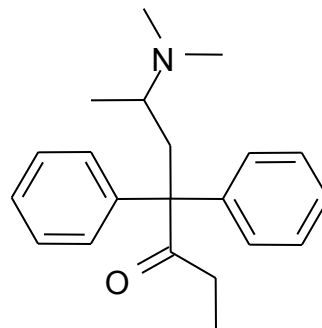
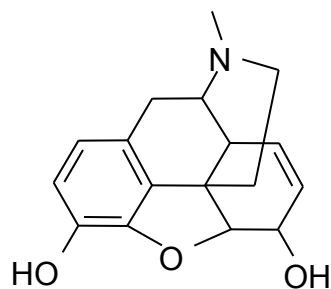


0.95 similar
Heroin



0.20 similar
Methadone

Morphine and methadone



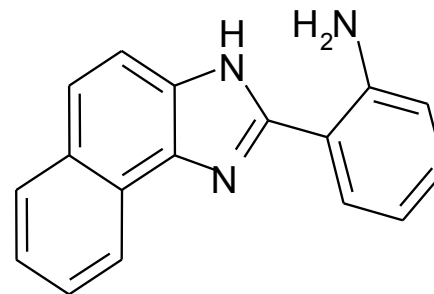
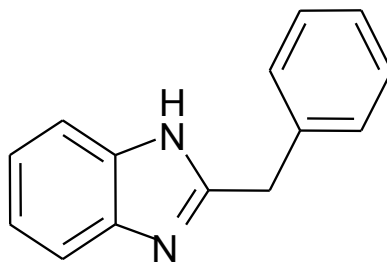
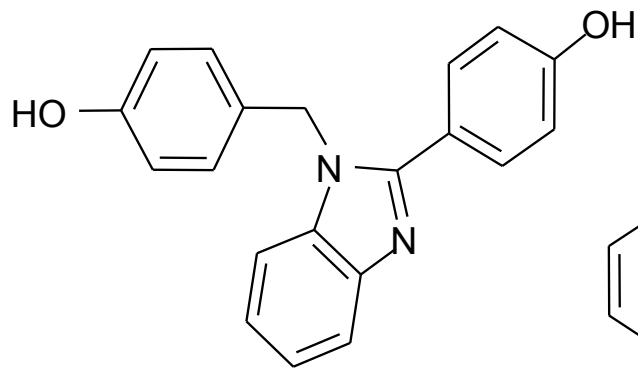
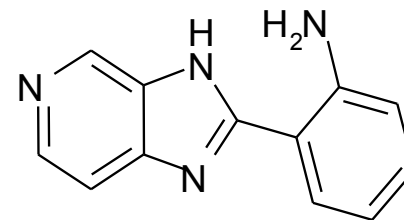
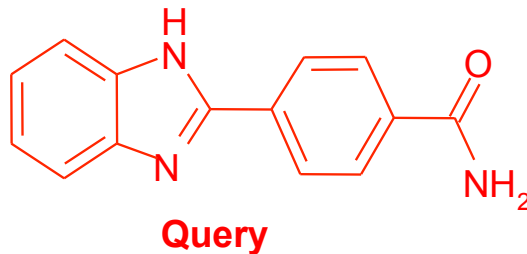
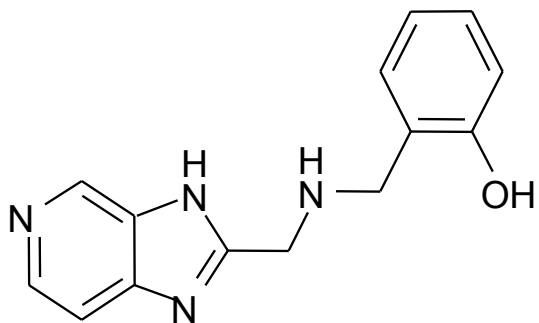
3D similarity measures

- Would expect that 3D descriptors would provide a more detailed characterisation of a molecule than a simple 2D fingerprint
- Wide range of descriptors now under investigation, e.g.
 - Distance-based 3D fingerprints
 - Overlay of 3D shapes or electrostatic fields
- No consensus as yet as to the most generally effective approach
 - Need for conformational analysis
 - Computationally demanding

Similarity searching

- Given a *target* (or *reference*) structure find molecules in a database that are most similar to it (“give me ten more like this”)
 - Compare the target structure with each database structure and measure the similarity
 - Sort the database in order of decreasing similarity
 - Display the top-ranked structures (“nearest neighbours”) to the searcher
 - Use of interesting structures (however defined) for further searches, bioactivity testing or whatever

Fingerprint/Tanimoto-based 2D similarity searching



Similarity searching

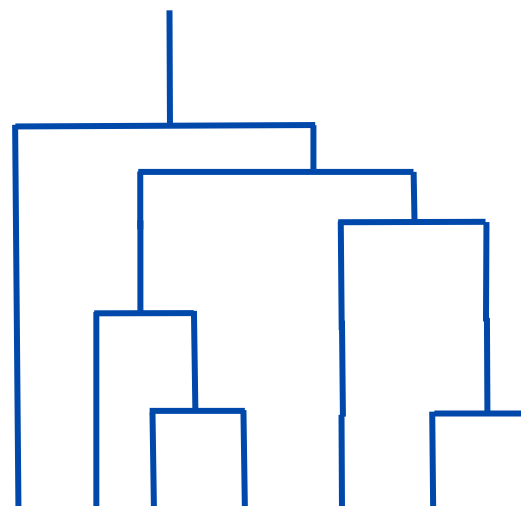
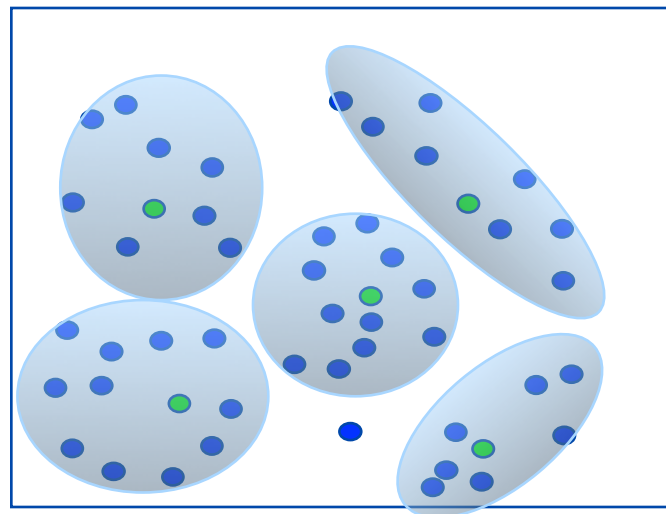
- Originally developed as a complement to substructure searching
 - No need for a detailed pharmacophore
 - Control over volume of output
- Rapidly adopted since both efficient and effective, and basic ideas extended to other applications
 - Cluster analysis
 - Molecular diversity analysis

Cluster analysis

Compute similarities and then cluster molecules so that molecules in the same (or different) clusters are similar (or dissimilar) to each other

Range of clustering methods available, e.g., Jarvis-Patrick (non-hierarchical) or Ward's (hierarchical) methods

Modern hardware/software enables clustering of files containing millions of molecules



Diversity analysis

- Similarity is a property of a pair of molecules; diversity is a property of a set of molecules
- Idea of choosing a representative subset from a large database, e.g., for biological testing
- Typical algorithm to select a set of dissimilar (e.g., 1-Tanimoto) molecules from a database
 1. Select a molecule and place in subset
 2. Calculate dissimilarity between each remaining molecule and the subset molecules
 3. Choose next molecule that is most dissimilar to the subset molecules
 4. If less than n subset molecules then return to 2

Comparison and evaluation of methods

- Use of datasets for which both structural and property/activity data are available, e.g., for comparing similarity searching methods
 - Given a known, bioactive reference structure, search it against a database that contains other molecules having the same activity
 - Note where the actives appear in the ranked list
 - A good similarity measure will cluster the known actives towards the top of the ranking
- Possible to identify good performers but no one measure is always the best, so idea of using multiple similarity searches

Data fusion

- Fusion of ranked list generated for same active compound (*similarity fusion*)
 - Do a similarity search for a reference structure and rank the database in order of decreasing similarity
 - Repeat with different descriptors, coefficients, etc.
 - Add the rank positions for a given structure to give an overall fused rank position
 - These fused rankings form the output from the search
- Consistency of search performance across a range of reference structures, types of fingerprint, biological activities etc.
- Increasing number of variations on this idea, e.g., use of multiple reference structures (*group fusion*)
- Analogous approaches (called *consensus scoring*) used in docking studies. Cf “wisdom of crowds”

Recent Sheffield research (all using 2D fingerprints)

- Interactions between the weighting scheme and the similarity coefficient
 - The Tanimoto's performance can be adversely affected by some types of weighting scheme
- Design of comparative studies
 - How many reference structures are required to differentiate between similarity measures?
- Scaffold-hopping
 - Can fingerprints provide at least some scope for scaffold-hops in similarity searching?
- Registration of orphan drugs
 - Collaboration with the European Medicines Agency (EMA)
 - Focus on individual similarity values

Orphan drugs

- Orphan drugs are medicines to treat people with rare diseases, where the numbers involved will not sustain the costs of conventional drug discovery
- The EU provides a range of incentives to encourage the development of such drugs, including market exclusivity
 - Once orphan drug status has been conferred, no similar molecule can come to market for ten years
 - How to define “similar molecule” for this purpose?

Registration process for orphan drugs

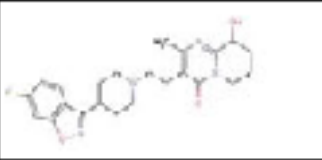
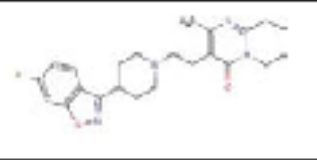
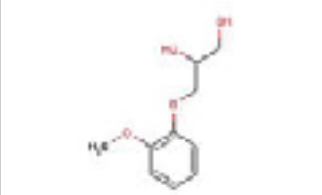
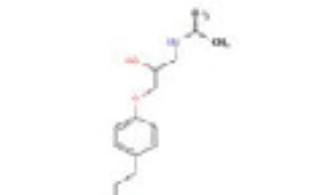
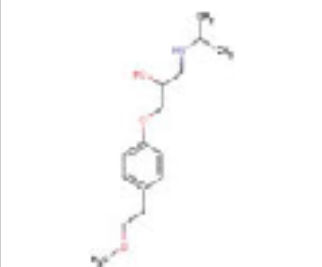
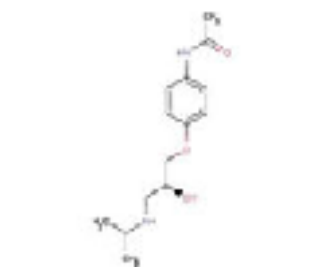
- This is done by the EMA Committee for Medicinal Products for Human Use (CHMP), which decides if a proposed molecule is similar to an existing orphan drug
- Orphan drug status conferred only if **not** similar on the basis of mode of action, physicochemical properties, and structural nature
- Structural similarity to date based on human judgement: can this be quantified?

Training-set based on expert judgements

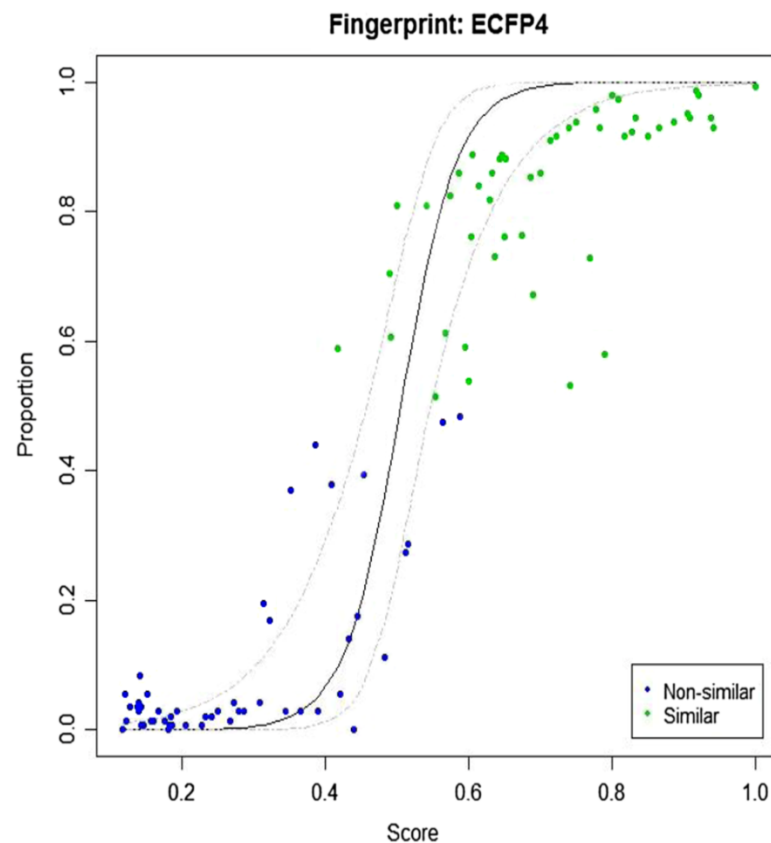
- 143 experts (from regulatory authorities in the EU, USA, Japan and Taiwan) assessed the similarity (Yes/No) of a training-set containing 100 pairs of molecules from DrugBank
- Similarities for each such pair computed using a range of 2D fingerprint
- Is there a fair degree of consistency between the expert judgements and do these correlate with the computed scores?

Answer: Yes

Typical expert judgements

Molecule A	Molecule B	Yes	No	Similarity
		0.93	0.07	0.865
		0.14	0.86	0.432
		0.59	0.41	0.595

Plot of proportion of experts saying similar against similarity score



Logistic regression

- Logistic regression yields an equation of the form: $\text{logit}(p) = \beta_0 + \beta_1 s$ (where p is the probability that a pair will be judged to be similar given a computed similarity of s)
- Training-set used to give values for β_0 and β_1 , and the equations were then applied to a test-set of 100 molecule-pairs previously considered by the EMA CHMP
- A value of $p > 0.5$ means that a pair is predicted to be similar and simple 2D fingerprints (BCI, Daylight, ECFP4 etc) had $> 95\%$ correct predictions across the test-set

Conclusions

- Measures of structural similarity underlie many processes in chemoinformatics
- Measures based on 2D fingerprints and the Tanimoto coefficient perform remarkably well given the simplicity of the procedures
- Fusion methods can be used to combine the results obtained from different measures
- The orphan drug application is a real-world application where the focus is on individual pairs of molecules, rather than large databases