### **Molecular Similarity and Chemical Similarity**











**Chemical and Molecular Similarity** 

Molecular similarity importance

**Representations of chemical structures** 

Similarity and Dissimilarity measures

**Similarity searching** 

**Cluster and diversity analysis** 













The terms chemical and molecular similarity are often used synonymously but this may not be entirely accurate.

**Chemical similarity** is based primarily on the physicochemical characteristics of compounds (e.g., solubility, boiling point, log P, molecular weight, electron densities, dipole moments, etc.)

**Molecular similarity** focuses primarily on the structural features (e.g., shared substructures, ring systems, topologies, etc.) of compounds and their representation.







Two Vascular Endothelial Growth Factor Receptor 2 ligands and different ways to assess their similarity.

by www.









The notion of chemical similarity (or molecular similarity) is one of the most important concepts in chemoinformatics.

It plays an important role in predicting the properties of chemical compounds, designing chemicals with a predefined set of properties and, especially, in conducting drug design studies by screening large databases containing structures of available (or potentially available) chemicals.

There are many ways to measure the similarity







### Substructure search



by www.







The rationale for similarity searching lies in the similar property principle [Johnson and Maggiora 1990] which states that structurally similar molecules tend to have similar properties.

Given a molecule of known biological activity, compounds that are structurally similar to it are likely to exhibit the same activity. This characteristic has been referred to as **neighbourhood behaviour** [Patterson et al. 1996].













Morphine

Codeine

Heroin

#### Compounds active at opioid receptors







#### SAPIT.NZA UNIVERSITÀ DI ROMA

### Neighbourhood principle (3)

Compounds such as the ibuprofen enantiomers, ibuprofen and paracetamol, or diclofenac and lumiracoxib, appear visibly similar.

From a medicinal chemistry point of view, however, this assessment may not be generally agreed upon since small chemical differences can lead to important changes in specificity profiles (e.g., diclofenac vs lumiracoxib) or compounds containing different functional groups can be synthesized or derivatized in different ways (e.g., ibuprofen vs paracetamol).

Moreover, these COX inhibitors are involved in highly complex similarity–activity relationships (similar mechanism of action) that also cannot easily be separated from a medicinal chemistry perspective.



Cyclooxygenase (COX) inhibitors and their activity profiles are compared. HSL = Hormone-Sensitive Lipase



Similarity versus activity. Three vascular endothelial growth factor receptor 2 ligands are shown that represent different (vertical vs horizontal) similarity–activity (potency) relationships.

There are many exceptions to the principle but it is an excellent ruleof-thumb in the absence of more detailed knowledge

People's judgements of similarity are inherently subjective, so need to provide a quantitative basis, a similarity measure, for assessing the degree of resemblance







The main difficulty with similarity searching is that assessing the degree of similarity between two objects is subjective.

In order to be able to quantify the similarity between two molecules, a similarity searching method requires two components:

- a set of numerical descriptors that can be used to compare molecules;
- a similarity coefficient which provides a way of quantifying the degree of similarity based on the 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







## Molecular descriptors

**OD**: bond counts, molecular weight, atom counts

1D: fragment counts, H-Bond acc/don, Crippen, PSA, SMARTS

2D: topological descriptors (Balaban, Randic, Wiener, BCUT, kappa, chi)

3D: geometrical descriptors, surface properties, COMFA

4D: 3D coordinates + conformations









### Molecular fingerprints



- Lots of types of fingerprints
- "Keyed" fingerprints indicate the presence or absence of a structural feature
- Length can vary from 166 to 4096 bits or more
- Fingerprints usually compared using the Tanimoto metric













### **Similarity measures**

Name	Formula for continuous variables	Formula for binary (dichotomous) variables
Tanimoto (Jaccard) coefficient)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2 - \sum_{i=1}^{N} x_{iA} x_{iB}}$ Range: -0.333 to +1	$S_{AB} = \frac{c}{a+b-c}$ Range: 0 to 1
Dice coefficient (Hodgkin index)	$S_{AB} = \frac{2\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2}$ Range: -1 to +1	$S_{AB} = \frac{2c}{a+b}$ Range: 0 to 1
Cosine similarity (Carbó index)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\left[\sum_{i=1}^{N} (x_{iA})^2 \sum_{i=1}^{N} (x_{iB})^2\right]^{1/2}}$ Range: -1 to +1	$S_{AB} = \frac{c}{\sqrt{ab}}$ Range: 0 to 1
Euclidean distance	$D_{AB} = \left[\sum_{i=1}^{N} (x_{iA} - x_{iB})^2\right]^{1/2}$ Range: 0 to $\infty$	$D_{AB} = \sqrt{a+b-2c}$ Range: 0 to N
Hamming (Manhattan or City-block) distance	$D_{AB} = \sum_{i=1}^{N}  x_{iA} - x_{iB} $ Range: 0 to $\infty$	$D_{AB} = a + b - 2c$ Range: 0 to N
Soergel distance	$D_{AB} = \frac{\sum_{i=1}^{N}  x_{iA} - x_{iB} }{\sum_{i=1}^{N} \max(x_{iA}, x_{iB})}$ Range: 0 to 1	$D_{AB} = \frac{a+b-2c}{a+b-c}$ Range: 0 to 1

• Similarity (S) or distance (D) coefficients in common use for similarity searching in chemical databases.

by www.

- For binary data:
  - **a** is defined as the number of bits set to "1" in molecule **A**
  - **b** as the number of bits set to "1" in molecule **B**
  - **c** as the number of bits that are "1" in both **A** and **B**.

[Willett et al. 1998.]







# Basic bit count terms of similarity

Symbol	Description	
onlyA	number of bits set "on" in fingerprint A but not in B	A 1 0 1 1 1 0 0 0 B 1 1 0 1 1 0 1 0
onlyB	number of bits set "on" in fingerprint B but not in A	A 1 0 1 1 1 0 0 0 B 1 1 0 1 1 0 1 0
bothAB	number of bits set "on" in both fingerprints	A 1 0 1 1 1 0 0 0 B 1 1 0 1 1 0 1 0
neither AB	number of bits set "off" in both fingerprints	A 1 0 1 1 1 0 0 0 B 1 1 0 1 1 0 1 0
A	number of bits set "on" in fingerprint A	
B	number of bits set "on" in fingerprint B	





### **Euclidean**



#### Formula: $Sim_{Euclid}(A,B) = \sqrt{\frac{bothAB + neitherAB}{onlyA + onlyB + bothAB + neitherAB}}$ $\mathbf{p} = \mathbf{A}$ Range: q = B $\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \dots + (q_n-p_n)^2}$ [0.0 - 1.0] $=\sqrt{\sum_{i=1}^n (q_i-p_i)^2}.$ Example: $\sqrt{\frac{bothAB+neitherAB}{onlyA+onlyB+bothAB+neitherAB}} = \sqrt{\frac{3+1}{1+2+3+1}} = \sqrt{\frac{4}{8}} = 0.707$ A 1 0 1 1 1 0 0 0 B1101101





### Manhattan



#### Formula:

$$Sim_{Manhattan}(A, B) = \frac{onlyA + onlyB}{onlyA + onlyB + bothAB + neitherAB}$$

#### Range:

[1.0-0.0]

#### Example:





### Cosine





Calculates the ratio of the bits in common to the geometric mean of the number of "on" bits in the two fingerprints.





### **Tanimoto (Jaccard)**

#### Formula:

$$Sim_{Tanimoto}(A, B) = \frac{bothAB}{|A|+|B|-bothAB} = \frac{bothAB}{onlyA+onlyB+bothAB}$$

Range:

$$[0.0 - 1.0]$$
  $J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$ 

by www.RC

Example:





### Tversky



 $S(X,Y)=rac{|X\cap Y|}{|X\cap Y|+lpha|X-Y|+eta|Y-X|}$ 

 $Sim_{Tversky}(A,B) = rac{bothAB}{lpha*onlyA+eta*onlyB+bothAB}$ 

The Tversky similarity measure is asymmetric. Setting the parameters lpha=eta=1.0 is identical to using the *Tanimoto* measure.

The factor  $\alpha_1$  weights the contribution of the first 'reference' molecule. The larger  $\alpha_1$  becomes, the more weight is put on the bit setting of the reference molecule.

Setting  $\alpha = \beta = 1$  produces the <u>Tanimoto coefficient</u>. Setting  $\alpha = \beta = 0.5$  produces <u>Dice's coefficient</u>.

Range:

variable

Example:

If we consider **A** to be the prototype and **B** to be the variant, then  $\alpha$  corresponds to the weight of the prototype and  $\beta$  corresponds to the weight of the variant.

$$\begin{array}{l} \textbf{A} \boxed{101111000} \\ \textbf{B} \boxed{1101110110} \\ \end{array} \qquad \begin{array}{l} \frac{bothAB}{\alpha \ast onlyA + \beta \ast onlyB + bothAB} \left( \alpha = 2.0, \beta = 1.0 \right) = \frac{3}{2.0 \ast 1 + 1.0 \ast 2 + 3} = \frac{3}{7} = 0.438 \end{array}$$







### 2D similarity (1)





Codeine 0.99 similar



Heroin 0.95 similar



by www.

Methadone 0.20 similar

Similarities to morphine calculated using Daylight fingerprints and the Tanimoto coefficient.







### 2D similarity (2)























#### **Binning Clustering**

# Cluster and diversity analysis

**Hierarchical Clustering** 

#### **Multidimensional Scaling**

### References

- Leach, Andrew R., and Valerie J. Gillet. An introduction to chemoinformatics. Springer Science & Business Media, 2007.
- Baldi, Pierre, and Ramzi Nasr. "When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values." Journal of chemical information and modeling 50.7 (2010): 1205-1222.

by www.

- Riniker, Sereina, and Gregory A. Landrum. "Open-source platform to benchmark fingerprints for ligand-based virtual screening." Journal of cheminformatics 5.1 (2013): 1.
- Maggiora, Gerald, et al. "Molecular Similarity in Medicinal Chemistry: Miniperspective." Journal of medicinal chemistry 57.8 (2013): 3186-3204.
- Bajusz, Dávid, Anita Rácz, and Károly Héberger. "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?." Journal of cheminformatics 7.1 (2015): 1.
- Duesbury, Edmund, John Holliday, and Peter Willett. "Maximum common substructure-based data fusion in similarity searching." Journal of chemical information and modeling 55.2 (2015): 222-230.
- Award Symposium 2015: Jürgen Bajorath. American Chemical Society, 2016.
- Schyman, Patric, Ruifeng Liu, and Anders Wallqvist. "General Purpose 2D and 3D Similarity Approach to Identify hERG Blockers." Journal of chemical information and modeling 56.1 (2016): 213-222.
- http://www.daylight.com/dayhtml/doc/theory/theory.finger.html
- https://docs.eyesopen.com/toolkits/python/graphsimtk/measure.html















ChemMine Tools	× +	
$\leftrightarrow$ $\rightarrow$ C $\triangle$ (i) Not sec	ure   chemmine.ucr.edu/myCompounds/	☆ 📇 🗣   🏶 🗄
Apps 📙 Università 📙 WWW	📙 Mail 📙 Ricerca 📙 Vari 📙 SoftwareVari 📙 NatComp 📙 METZ 📙 IM	P KI Kahoot!   Learning G »
ChemMine Tools		Î
Help		
Downloads		
Optional: Login or register to		
WORKBENCH My Compounds		
My Compounds	Show Structures Download SDF Download Smiles	
Add Compounds	No compounds uploaded.	
TOOLS		
Past Jobs		
Upload Numeric Data		







ChemMine Tools	× +	
← → C ① Not se	cure   chemmine.ucr.edu/accounts/signup/	☆ 📇 🗣   🏶 🗄
Apps 📙 Università 📙 WWV	V 🦲 Mail 🦲 Ricerca 🦲 Vari 🛄 SoftwareVari 📒 NatComp 🛄 METZ 📘 IMP	Ki Kahoot!   Learning G >>
ChernMine Tools		Ê
WORKBENCH My Compounds	Signup <sub>Email:</sub>	
Add Compounds	Create password:	
<b>TOOLS</b> Past Jobs Upload Numeric Data	Repeat password:	
Cluster Physicochemical Properties Similarity Workbench		







ChemMine Tools	× +		
← → C ① Not set	cure   chemmine.ucr.edu/myCompounds/		🖈 🔚 🖣 🛛 😵 🗄
Apps 📙 Università 📙 WWV	/ 🦲 Mail 📙 Ricerca 📙 Vari 📃 SoftwareVari 📃	, NatComp	📙 IMP   Ki Kahoot!   Learning G »
About			
Help			
Downloads			
Optional: Login or register to	o save data.		
	My Compounds		
WORKBENCH		ownload Smiles	
My Compounds	Show Structures Download 3D1 DC	ownload sinnes	
Add Compounds	No compounds uploaded.		
TOOLS	No compounds uploaded.		
Past Jobs			
Upload Numeric			







ChemMine Tools	× +	
← → C ☆ ③ Not sec	ure   chemmine.ucr.edu/accounts/signin/	☆ 📇 ●   🏶 🗄
Apps 📙 Università 📙 WWW	Mail 🦲 Ricerca 🦲 Vari 🛄 SoftwareVari 🧾 NatComp 🛄 METZ 🛄 IMP	K Kahoot!   Learning G »
ChemMine Tools		
WORKBENCH	Signin <sub>Email:</sub>	
My Compounds		
Add Compounds	Password:	
TOOLS	Remember me for a month	
Past Jobs Upload Numeric Data	Signin Forgot your password?	
Cluster		
Physicochemical Properties Similarity Workbench		







ChemMine Tools	× +	
← → C ① ▲ Not sec	ure   chemmine.ucr.edu/accounts/signin/	야 ☆ 📻 🌒 😵 :
Apps 🦲 Università 📙 WWW	7 📙 Mail 📙 Ricerca 🔜 Vari 🛄 SoftwareVari 📙 NatComp 🛄 METZ	, IMP 🕺 Kahoot!   Learning G »
ChemMine Tools		
<b>WORKBENCH</b> My Compounds	Signin Email: rino.ragno@uniroma1.i	
Add Compounds	Password: ••••••	
TOOLS	Remember me for a month	
Past Jobs Upload Numeric Data	Signin Forgot your password?	
Cluster		
Physicochemical Properties Similarity Workbench		











ChemMine Tools	X W Methadone - Wikipedia X +		
← → C ☆ ●	en.wikipedia.org/wiki/Methadone	\$	📕 🌒   🍪 🗄
Apps Università	📕 WWW 📕 Mail 📒 Ricerca 🛄 Vari 🔜 SoftwareVari 🛄 NatComp 🛄 METZ 🛄 IMP K	Kahoot!   Learning G 🔇	Save to Mendeley »
🗱 Apps 🧕 Università	treatment and in the reduction or cessation of heroin use as measured by self-report and urine/hair analysis but did not affect criminal activity or risk of death. <sup>[16]</sup> Treatment of opioid-dependent persons with methadone follows one of two routes: maintenance or detoxification. <sup>[17]</sup> Methadone maintenance therapy (MMT) usually takes place in outpatient settings. It is usually prescribed as a single daily dose medication for those who wish to abstain from illicit opioid use. Treatment models for MMT differ. It is not uncommon for treatment recipients to be administered methadone in a specialist clinic, where they are observed for around 15–20 minutes post dosing, to reduce risk of diversion of medication. <sup>[16]</sup> The duration of methadone treatment programs range from a few months to several years. Given opioid dependence is characteristically a chronic relapsing/remitting disorder, MMT may be lifelong. The length of time a person remains in treatment depends on a number of factors. While starting doses may be adjusted based on the amount of opioids reportedly used, most clinical guidelines suggest doses start low (e.g. at doses not exceeding 40 mg daily) and are incremented gradually. <sup>[19][20]</sup>	CAS Number         76-99-3@           PubChem <u>CID</u> 4095@           UPHAR/BPS         5458@           DrugBank         D800333@           DrugBank         D800333@           JNII         UC6VBE7           KEGG         D8195@           CheBI         CHEBI680           ChEMBL         ChEMBL60           CheMBL         DTXSID70           Dashboard (EPA)         100.000.90           Chemical and phys         Formula	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓
	Methadone maintenance has been shown to reduce the transmission of blood borne viruses associated with opioid injection, such as hepatitis B and C, and/or HIV. <sup>[19]</sup> The principal goals of methadone maintenance are to relieve opioid cravings, suppress the abstinence syndrome, and block the euphoric effects associated with opioids.	Chirality Racemic m IILES CCC(C(C)=CC=CC=C1)(C2=CC C)C)=C InChI XV (what is this?) an provide a slow recover	[hide] =CC=C2)CC(N(C) [show] (verify)

by www.





,	ChemMine Tools
ChemMine Tools X	W Methadone - Wikipedia X +
← → C ☆ A Not secure	e   chemmine.ucr.edu/myCompounds/addCompounds/
Apps 🔜 Università 📃 WWW	📕 Mail 📙 Ricerca 📕 Vari 📕 SoftwareVari 📕 NatComp 📕 METZ 📕 IMP K Kahoot!   Learning G 🧿 Save to Mendeley 🔹
ChemMine Tools About	Help Downloads Logged in as rino.ragno@uniroma1.it   Logout
WORKBENCH	Add Compounds           SMILES Input         Draw a Molecule         SDF Input         SDF Upload         PubChem Import
My Compounds	SMILES Input Draw a Molecule SDF Input SDF Upload PubChem Import Enter smiles entry:
Add Compounds	CN1CC[C@]23C4=C5C=CC(O)=C4O[C@H]2[C@@H](O)C=C[C@H]3[C@H]1C5 Morphine
TOOLS	CCC(C(C1=CC=CC=C1)(C2=CC=C2)CC(N(C)C)C)=O Methadone
Past Jobs	
Upload Numeric Data	
Cluster	
Physicochemical Properties	
Similarity Workbench	Tags
Target Search	Opioids ×
SEARCH	Test
Structural Similarity	Opioinds





ChemMine Tools X	W Methadone - Wikipedia × +	
← → C ☆ ③ Not secure	chemmine.ucr.edu/myCompounds/	☆ 🜉 🔍 🧶 :
🗰 Apps 📃 Università 📃 WWW	Mail 🔜 Ricerca 📃 Vari 🔜 SoftwareVari 📃 NatComp 📃 METZ	IMP KI Kahoot!   Learning G 📀 Save to Mendeley »
ChemMine Tools About	Help Downloads	Logged in as rino.ragno@uniroma1.it   Logout
WORKBENCH	Success: Added 2 compounds.	×
My Compounds	My Compounds	
Add Compounds	Show Structures Download SDF Download Smile	25
TOOLS		
Past Jobs	Select all Deselect all Add Tags Remove Tags	Delete Selected
Upload Numeric Data	Showing 1 to 2 of 2 entries	
Cluster		Search:
Physicochemical Properties	Show 10 🗢 entries	
Similarity Workbench	CID 11 Name 11 Formula	t↓ Tags t↓ Options t↓
Target Search	Methadone C21H27NO	Test, Opioinds, 🛛 🔍 🧪 Opioids,
SEARCH	Morphine C17H19NO3	Test, Opioinds, 🔍 🧪
Structural Similarity		Opioids,

by www.





	ChemMine Tools	by www.RCM	
ChemMine Tools X	W Methadone - Wikipedia X +		
$\leftrightarrow$ $\rightarrow$ C $\triangle$ (i) Not secure	e   chemmine.ucr.edu/similarity/	☆ 📻 🗣 🛛 💝 🗄	
Apps 📙 Università 📃 WWW	🔜 Mail 🧾 Ricerca 🛄 Vari 🛄 SoftwareVari 🛄 NatComp 🛄 M	IETZ 📙 IMP Ҟ Kahoot!   Learning G 🔇 Save to Mendeley »	
ChemMine Tools About	Help Downloads	Logged in as rino.ragno@uniroma1.it   Logout	
WORKBENCH	Compound Similarity Select two compounds to compare from the grid bel	ow.	
My Compounds	Selected Compounds		
Add Compounds			
TOOLS			
Past Jobs			
Upload Numeric Data			
Cluster			
Physicochemical Properties			
Similarity Workbench	All ×		
Target Search	Update		
SEARCH	Morphine Methadon	e	
Structural Similarity	· · · ·	•	







































