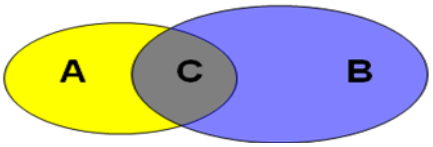
Structural similarity between molecules: Relative vs. Absolute character of similarity

The molecular similarity is context defined notion, i.e., it depends on the endpoint molecular are compared for. For example, two molecules could be similar with respect to AMES mutagenicity (having same functional groups which could damage DNA) but could be dissimilar with respect to acute aquatic toxicity (not having same functionalities damaging proteins/lipids or same mode of action). Such (dis)similarity is called mechanistic and it is different of the similarity based on abstract molecular features not directly related to functionalities conditioning the interactions with macromolecules (eventually associated with molecular endpoints). The latter is called structural similarity and it has more general character as compared to the mechanistic similarity. The structural similarity is often used in QSAR studies. In this appendix we would like to:

- Introduce the factors affecting the values of the structural similarity estimates (indices) and
- Illustrate that **the structural similarity estimates have relative but not absolute character**.

There are three main factors affecting the values of the calculated structural similarity indices:

a. Formulas for calculating similarity:



- Tanimoto
  $$\frac{c}{a+b+c}\cdot 100$$
- Dice
  $$\frac{c}{0.5\left[(a+c)+(b+c)\right]}\cdot 100$$
- Kulczynski-2
  $$\frac{1}{2}\left(\frac{c}{a+c}+\frac{c}{b+c}\right)\cdot 100$$
- Ochiai (Cosine)
  $$\frac{c}{\sqrt{(a+c)*(b+c)}}\cdot 100$$
- Yule
  $$\frac{c}{\min\left((a+c),(b+c)\right)}\cdot 100$$

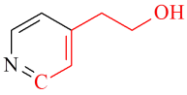
In the illustrative example two of the formulas have been used to illustrate the relative character of the structural similarity - Tanimoto and Dice. Based on the formulas they

are defined, Dice similarity index provides higher similarity estimates as compared to Tanimoto index (at constant of the other factors – molecular features and atom specification).

b. Molecular features. Multiplicity of molecular features could be used when molecules are compared:

- Atom pairs. The atom pairs is distance in bonds along the shortest path between an atom of type Atomi and an atom of type Atomj: Atomi – Atomj – Distance

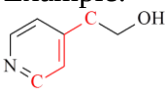
Example:



C{H1}{π1}_O{H1{π0}_O5 - atom pair of a carbon without hydrogen neighbours and one π-bond,5 bonds away from oxygen with one hydrogen neighbour and no π-bond

- Topological torsions. The topological torsions is sequence of four atoms consecutively bonded accounting the number of non-hydrogen atoms attached to them, and the number of incident π-bonds: Atomi – Atomj – Atomk – Atoml

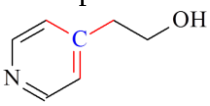
Example:



C{H1}{π1}_C{H1}{π1}_C{H0}{π1}_C{H2}{π0}

- Atom centered fragments (ACF). The atom centered fragment is a fragment containing the central atom and his first, second, etc. neighbours.

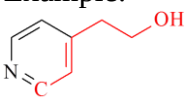
Example:



C{H1}{π1}_C{H0}{π1}_(C{H2}{π0})_C{H1}{π1}

- Path. The path is a sequence of 1 to 8 atoms

Example:



C{H1}{π1}_C{H1}{π1}_(C{H0}{π1})_C{H2}{π0}_C{H2}{π0}_O{H} {π0}

- Cycle. The Cycle extract all atoms cycles in the molecule

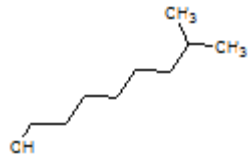
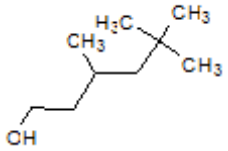
In the current example, two of the molecular features have been used, only - Atom pairs and Atom centered fragments (ACF). Both features provide different values for similarity estimates depending on the compared molecular structures (when keeping constant the other two factors).

c. Atom specification. Different atom specifications could be used when molecular features are defined:

- Atom type. Different types of atoms within molecule are taken into account, such as H, Cl, Br, Na, C, O, N, P, S, …
- Count H attached. Hydrogen atoms attached to the central atoms are taken into account
- Count heavy atom attached. Heavy atoms attached are taken into account
- Hybridization. Type of hybridization of atoms is taken into account: sp3, sp2, sp.
- Incident pi-bonds
- Valency. Valencies of atoms are taken into account.
- Charge. Atomic charges are taken into account, such as Na+, O-, N+, …
- Cyclic. Presence of cyclic structures are taken into account

In the current illustrative example two set of atom specifications have been used, only – atom type and atom type + number of attached H + hybridization. The more atom specifications are used when defining the molecular features the lower are the similarity estimates (when keeping constant the other two factors).

Similarity indices calculated for two chemicals with variation of the above three factors are listed in the table below:

Compared structures	**Similarity method**	**Molecular features**	**Atom characteristics**	**Measured value, %**
 Hydrolysis product of ester CAS 42131-25-9	Tanimoto	Atom pairs	Atom type	66.67
	Tanimoto	Atom pairs	Atom type; Count H attached; Hybridization	23.29
	Tanimoto	ACF*	Atom type	53.85
 Hydrolysis product of ester CAS 59219-71-5	Tanimoto	ACF	Atom type; Count H attached; Hybridization	25
	Dice	Atom pairs	Atom type	80
	Dice	Atom pairs	Atom type; Count H attached; Hybridization	37.78
	Dice	ACF	Atom type	70
	Dice	ACF	Atom type; Count H attached; Hybridization	40

*ACF – Atom centered fragments

A. General Conclusions:

- Calculated similarity indices between molecules could vary within a large range (from 24% - to 80%, in the above example), depending on the formula/features/specifications used
- There is no scientific or regulatory guidance defining the factors affecting calculations of similarity, in terms of used molecular formula, compared molecular features and atom specifications.
- **Similarity estimates between chemicals have relative but not absolute character** – all depend on the details taken into account when similarity is assessed – the more are the details used when two molecules are compared the lower are the similarity estimates between them.

B. Guidance when a set of analogues are subcategorized (clustered of similar analogs are formed). One could use:

- At constant values of the above three factors defining similarity the lower is the similarity threshold the lower will be the number of clusters similar analogues and vice versa.
- The same number of clusters will be resulted at lower similarity thresholds, however, when more stringent requirement of the above three factors are applied when similarity is defined.