QSAR TOOLEOX

The OECD QSAR Toolbox for Grouping Chemicals into Categories

OECD (Q)SAR Toolbox v.4.4.1

Tutorial illustrating new options for structure similarity comparison

- Background
- Keywords
- Aims
- PubChem features
- The exercise
- Workflow

Background

- This presentation is designed to familiarize the Toolbox user with new structure similarity comparison features;
- Structure similarity options in the Toolbox have been expanded including PubChem substructure features.

- Background
- Keywords
- Aims
- PubChem features
- The exercise
- Workflow

Keywords

TARGET CHEMICAL - chemical of interest

MODULE – a Toolbox module is a section dedicated to specific actions and options (e.g. Profiling)

WORKFLOW – the use, in combination, of the different modules (e.g. prediction workflow: from input to report)

PROFILER - algorithm (rule set) for the identification of specific features of the chemicals. Several types of profilers are available, such as structural (e.g. Organic functional groups), mechanistic (e.g. Protein binding by OECD) and endpoint-specific (e.g. in vitro in vitro mutagenicity (Ames test) alerts by ISS) profilers.

ALERT - the profilers consist of sets of rules or alerts. Each of the rules consists of a set of queries. The queries could be related to the chemical structure, physicochemical properties, experimental data, comparison with the target or list with substances and external queries from other predefined profilers (reference queries).

CATEGORY – "group" of substances sharing same characteristics (e.g. the same functional groups or mode of action). In a typical Toolbox workflow, it consists of the target chemical and its analogues gathered according to the selected profilers

ENDPOINT TREE – Endpoints are structured in a branched scheme, from a broader level (Phys-Chem properties, Environmental Fate and transport, Ecotoxicology, Human health hazard) to a more detailed one (e.g. EC3 in LLNA test under Human health hazard-Skin sensitization)

DATA MATRIX – Table reporting the chemical(s) and data (experimental results, profilers outcomes, predictions). Each chemical is in a different column and each data in a different row

- Background
- Keywords
- Aims
- PubChem features
- The exercise
- Workflow

Aims

- To demonstrate to the Toolbox user how to compare two chemicals with respect to PubChem substructure similarity features;
- To demonstrate to the Toolbox user how to compare a chemical with a list of chemicals with respect to PubChem substructure similarity features.

- Background
- Keywords
- Aims
- PubChem features
- The exercise
- Workflow

PubChem features Overview

PubChem generates a binary substructure fingerprint for each chemical structure. These fingerprints are used by PubChem for similarity neighboring and similarity searching.

A substructure is a fragment of chemical structure. A fingerprint is an ordered list of binary (1/0) bits. Each bit represent a Boolean determination of specific atom or test features used further for similarity neighboring and similarity searching. Seven groups of PubChem features are defined and used:

- Hierarchical element counts;
- Rings;
- Simple atom pairs;
- Simple atom nearest neighbors;
- Detailed atom neighbors;
- Simple SMARTS patterns;
- Complex SMARTS patterns.

PubChem features Overview

Numbers in brackets show the number of common substructure features between the two compared structures out of all features found in a single chemical.

- Green colored features are common for both structures;
- Red colored features are unique .

Each of the *PubChem* features has **bit position** (1) which correspond to a **bit substructure** (2).



April 2020

PubChem features Hierarchical element counts

• **Hierarchical element counts** - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.

They include bit positions from 001 to 115.

Similarity form	– 🗆 ×						
Structure similarity: 96.65%							
Target chemical	Query chemical						
PubChem features (101/104)	PubChem features (101/105) ^						
#001 - ">= 4 H" (1/1) #010 - ">= 2 C" (1/1) #011 - ">= 4 C" (1/1) #015 - ">= 1 N" (1/1) #019 - ">= 1 O" (1/1) #020 - ">= 2 O" (1/1) #038 - ">= 1 Cl" (1/1)	#001 - ">= 4 H" (1/1) #010 - ">= 2 C" (1/1) #011 - ">= 4 C" (1/1) #015 - ">= 1 N" (1/1) #019 - ">= 1 O" (1/1) #020 - ">= 2 O" (1/1) #038 - ">= 1 Cl" (1/1)						
#1/9 - ">= 1 any ring size 6" (1/1) #180 - ">= 1 saturated or aromatic #256 - ">= 1 aromatic ring" (1/1) #284 - "C-H" (1/1) #285 - "C-C" (1/1) #286 - "C-N" (1/1)	<pre>#1/9 - ">= 1 any ring size 6" (1/1) #180 - ">= 1 saturated or aromatic #256 - ">= 1 saturated or aromatic #284 - "C-H" (1/1) #285 - "C-C" (1/1) #286 - "C-N" (1/1)</pre>						

PubChem features Rings in a canonic ESSSR ring set

Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set -٠ These bits test for the presence or count of the described chemical ring system. An ESSSR ring is any ring which does not share three consecutive atoms with any other ring in the chemical structure. For example, naphthalene has three ESSSR rings (two phenyl fragments and the 10-membered envelope), while biphenyl will yield a count of only two ESSSR rings.





PubChem features Simple atom pairs

• **Simple atom pairs** – These bits test for the presence of patterns of bonded atom pairs, regardless of bond order or count.

They include bit positions from 264 to 327.



PubChem features Simple atom nearest neighbors

 Simple atom nearest neighbors – These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") (1) or count, but where bond aromaticity (denoted by ":") (2) is significant.



PubChem features Detailed atom neighborhoods

 Detailed atom neighborhoods – These bits test for the presence of detailed atom neighborhood patterns, regardless of count, but where bond orders are specific, bond aromaticity matches both single and double bonds, and where "-", "=", and "#" matches a single bond, double bond, and triple bond order, respectively.



The OECD (Q)SAR Toolbox for Grouping Chemicals into Categories

PubChem features Simple SMARTS patterns

• **Simple SMARTS patterns** – These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.

They include bit positions from 461 to 713.



PubChem features Complex SMARTS patterns

Complex SMARTS patterns – These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.

They include bit positions from 714 to 881.



- Background
- Aims
- PubChem features
- The exercise
- Workflow

The Exercise

- In this exercise we will compare:
 - 1. Two chemicals with respect to *PubChem* substructure similarity features (we will use *m*-*Chloroaniline* and *benzoic acid*);
 - 2. One chemical with a list of chemicals with respect to *PubChem* substructure similarity features (we will use *m-Chloroaniline* and Skin sensitization ECETOC database).

- Background
- Keywords
- Aims
- PubChem features
- The exercise
- Workflow
 - **o** Substructure similarity between two chemicals

Substructure similarity between two chemicals



- 1. Go to *Profiling* module;
- 2. Right click over the Structure similarity profiler and select Options;
- 3. Uncheck all molecular features and select only *PubChem* features. The additional similarity options
- (e.g. Calculation and Atom characteristics) do not have influence to the PubChem features.;
- 4. Double click on the structure in left which will be our target. 2D editor window appears.

Substructure similarity between two chemicals



Substructure similarity between two chemicals



The structure which you have drawn appears (1). Double click on the structure in right (2) to draw the second chemical.

Substructure similarity between two chemicals



Click on the **Eraser** button (1) and remove the nitro group (2). Select the oxygen symbol (3) and click over the chorine atom (4). Now you are ready with drawing of benzoic acid and have to click on **OK** (5)

Substructure similarity between two chemicals

Main Model of fabries Biomedic (laced) Am park Biomedic (laced) <th>Similarity options</th> <th></th> <th>- 0</th> <th>- x</th> <th>Similarity form</th> <th>– 🗆 ×</th>	Similarity options		- 0	- x	Similarity form	– 🗆 ×
 Integration (Integration) Integration (Integratin) Integration (Integratin) Integr	Measure	Molecular features	Calculation		Structure sir	nilarity 53.75%
$ \begin{array}{ c } \hline c \ c \ c \ c \ c \ c \ c \ c \ c \ c$	Innimoto (Jaccard) Dice Lice Kulczynski-2 Ochiai(Cosine) Yule	Atom pairs Topologic torsions Atom centered fragments Path	 ○ Fingerprint ● Hologram 		Target Chemical	Query Chemical
$ \begin{array}{c} \text{Formula} \\ c/0.5[(a+c)+(b+c)] \\ \hline \text{Perclotent or derivations of a chemical is structure.} \\ \hline \text{A independent of a chemical is structure.} \\ \hline A independent of a chemical $		Cycles PubChem features Options	 Average by features Combine all features 		$\widehat{\bigcirc}$	
A C B count a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Structure Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system environment (nearest neighbors), etc., in a chemical structure. Image: a type of ring system environment (nearest neighbors), etc., in comment (nearest neigh	Formula c/0.5[(a+c)+(b+c)]	Description The PubChem System generates a binary substructure fingerprint for chemical structures. A substructure is a fragment of a chemical structure. A fingerprint is an ordered list of binary (1/0) bits. Each bit represence of, for example, an element	Atom characteristics Atom type Count H attached Count Heavy atoms attached Hybridization Incident pi-bonds Valeory		H ₂ N	ОН
$\frac{1}{2} \frac{1}{2} \frac{1}$	(A (C) B)	count, a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical		-	✓ PubChem features (43/76) #001 ">= 4 H" (1/1)	A PubChem features (43/84) ^/// #001 ">= 4 H" (1/1)
$ \begin{array}{c} for the second secon$		structure.			#010 - ">= 2 C" (1/1)	#010 - ">= 2 C" (1/1)
$\begin{array}{c} \# 015 - \$ = 1 \text{ N}^{\circ}(0/1) \\ \# 019 - \$ = 1 \text{ or } (0/1) \\ \# 038 - \$ = 1 \text{ C}^{\circ}(0/1) \\ \# 038 - \$ = 1 \text{ C}^{\circ}(0/1) \\ \# 109 - \$ = 1 \text{ or } or $	Structure			Define	#011 - ">= 4 C" (1/1)	#011 - ">= 4 C" (1/1)
$\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 41 + 43}$ $\frac{A + B + C}{33 + 10}$ $\frac{A + B + C}{33 $	Examin			benne	#015 - ">= 1 N" (0/1) #029 ">= 1 C!" (0/1)	#019 - ">= 1 O" (0/1) #020 ">= 2 O" (0/1)
$\frac{33}{41} \frac{43}{43}$ $\frac{2}{1} \frac{43}{1} \frac{43}{1$		B C			#038 - 2 = 1 cr(0/1) #179 - ">= 1 any ring size 6" (1/1)	#020 - >= 2 0 (0/1) #179 - ">= 1 any ring size 6" (1/1)
#256 - ">= 1 aromatic ring" (1/1) #256 - ">= 1 aromatic ring" (1/1) #284 - "C-H" (1/1) #284 - "C-H" (1/1) #285 - "C-C" (1/1) #285 - "C-C" (1/1) #285 - "C-C" (0/1) #285 - "C-C" (0/1) #300 - "N-H" (0/1) #333 - "C(-C)(-C)" (0/1) #333 - "C(-C)(-C)(-C)" (1/1) #334 - "C(-C)(-C)(-C)" (1/1) #333 - "C(-C)(-C)(-C)(-C)" (0/1) #334 - "C(-C)(-C)(-C)" (0/1) #345 - "C(-C)(-C)(-1) #355 - "C(-C)(-C)" (0/1) #355 - "C(-C)(-C)" (0/1) #356 - "C(-C)(-C)(" (1/1)		41 43			#180 - ">= 1 saturated or aromatic	#180 - ">= 1 saturated or aromatic
H2N i Similarity = 53.750% Details #284 - "C-H" (1/1) #284 - "C-H" (1/1) H2N Help #285 - "C-C" (1/1) #285 - "C-C" (1/1) #285 - "C-C" (1/1) W Default Help #286 - "C-N" (0/1) #287 - "C-O" (0/1) #286 - "C-L" (0/1) #333 - "C(-C)(-C)(-C)(-C)" (0/1) #333 - "C(-C)(-C)(-C)(-C)(-C)(-C)(-C)(-C)(-C)(-C					#256 - ">= 1 aromatic ring" (1/1)	#256 - ">= 1 aromatic ring" (1/1)
H₂M Default Help OH Default Help OK Cancel #285 - °C-C° (0/1) #285 - °C-O° (0/1) #285 - °C-C° (0/1) #285 - °C-O° (0/1) #285 - °C-O° (0/1) #285 - °C-C° (0/1) #333 - °C(-C)(-C)° (0/1) #333 - °C(-C)(-C)° (0/1) #333 - °C(-C)(-C)° (0/1) #334 - °C(-C)(-C)(-C)° (0/1) #345 - °C(-C)(-C)(-C)(-N)° (0/1) #343 - °C(-C)(-C)(-C)(-N)° (0/1) #355 - °C(-C)(-C)(°C)(-N)° (0/1) #355 - °C(-C)(-C)(°C)(-N)° (0/1) #355 - °C(-C)(-C)(-N)° (0/1) #355 - °C(-C)(-C)(°C)(-N)° (0/1) #357 - °C(-C)(-C)(°C)(-N)° (0/1)	Similarity = 53.	750% Details			#284 - "C-H" (1/1)	#284 - "C-H" (1/1)
bH Default Help OK Cancel #200 - C-N (0/1) #300 - "0-H" (0/1) #300 - "0-H" (0/1) #300 - "N-H" (0/1) #333 - "C(~C)(~C)" (1/1) #333 - "C(~C)(~C)" (1/1) #334 - "C(~C)(~C)(" (1/1) #343 - "C(~C)(~C)(" (0/1) #345 - "C(~C)(~C)" (0/1)	H ₂ N CI				#285 - "C-C" (1/1)	#285 - "C-C" (1/1) #287 - "C-C" (0(1)
OK Cancel #255 **C(-*C)(**C)(*) #333 - "C(-*C)(**C)(**C)(**C)(**C)(**C)(**C)(**C		он	Default	Help	#280 - C-N (0/1) #295 - "C-C!" (0/1)	#287 - C-O (0/1) #309 - "O-H" (0/1)
#333 - 'C(~C)(~C)'' (1/1) #334 - "C(~C)(~C)(~C)'' (0/1) #333 - 'C(~C)(~C)(~C)'' (0/1) #345 - "C(~C)(~C)(~C)'' (0/1) #343 - "C(~C)(~C)(-N)" (0/1) #345 - "C(~C)(~C)(-N)" (0/1) #345 - "C(~C)(~C))" (0/1) #355 - "C(~C)(~C)" (0/1) #352 - "C(~C)(~N)" (0/1) #356 - "C(~C)(-C)" (0/1) #352 - "C(~C)(~N)" (0/1) #357 - "C(~C)(-C)(-C)" (0/1)		<->	ОК	Cancel	#300 - "N-H" (0/1)	#333 - "C(~C)(~C)" (1/1)
#341 - "C(~C)(~C)(-N)" (0/1) #345 - "C(~C)(~H)" (1/1) #343 - "C(~C)(~C))" (0/1) #353 - "C(~C)(~O)" (0/1) #345 - "C(~C)(~H)" (1/1) #356 - "C(~C)(~C)" (0/1) #352 - "C(~C)(~N)" (0/1) #357 - "C(~C)(:C)" (1/1)					#333 - "C(~C)(~C)" (1/1)	#334 - "C(~C)(~C)(~C)" (0/1)
#343 - "C(~C)(~Cl)" (0/1) #353 - "C(~C)(~O)" (0/1) #345 - "C(~C)(~H)" (1/1) #356 - "C(~C)(:C)" (1/1) #352 - "C(~C)(~N)" (0/1) #357 - "C(~C)(:C)" (0/1)					#341 - "C(~C)(~C)(~N)" (0/1)	#345 - "C(~C)(~H)" (1/1)
#345 - "C(~C)(~H)" (1/1) #356 - "C(~C)(:C)" (1/1) #352 - "C(~C)(~N)" (0/1) #357 - "C(~C)(:C)(:C)" (0/1)					#343 - "C(~C)(~Cl)" (0/1)	#353 - "C(~C)(~O)" (0/1)
#352 - "C(~C)(~N)" (0/1)					#345 - "C(~C)(~H)" (1/1)	#356 - "C(~C)(:C)" (1/1)
					#352 - "C(~C)(~N)" (0/1)	#357 - "C(~C)(:C)" (0/1)
					< >	< >

Similarity between the two structures is calculated automatically with respect to the *PubChem* features (1). Click on **Details** button (2) to get more information about the common and unique structural features.

- Background
- Keywords
- Aims
- PubChem features
- The exercise
- Workflow
 - Substructure similarity between of two chemicals
 - Substructure similarity between single chemical and chemical list

Substructure similarity between single chemical and chemical list



Substructure similarity between single chemical and chemical list



Substructure similarity between single chemical and chemical list

QSAR TOOLEOX	Profiling > Data > Category definition > Data Gap Filling > Report				
Dealling Custom profile		The OECD QSAR Toolbo for Grouping Chemicals into Categories			
Appi, Visu Way Delete	Filter endpoint tree	Developed by LMC, Bul 12			
2 Doc ■ 1 Skin sensitization ECETOC	$\begin{array}{ c c c c c c c c c c c c c c c c c c c$	П нус			
	Structure info Structure info Parameters Physical Chemical Properties Support of the suppo				
	Ecotoxicological Information Comparison				
Profiling: Options _ f Select All Unselect Organic functional group Organic functional group Organic functional group Organic functional group Organic functional groups, Norbert Haider (checkmol Selectable Tautomers unstable Toxicological Repeated dose (HESS) Custom Outomatic function Scheme (PBT) p-benzoquinone precursors Selectable (TOP) Selectable (TOP) Selectable (TOP) Selectable (TOP) Selectable	Check the box in front of the <i>Structure similarity</i> profile (1) and click on Apply (2) .				
Metabolism/Transformations Options Oselected f Select All Unselect All Invert Observed Mammalan metabolism Observed Mammalan metabolism Observed Rat Invvo metabolism Observed Rat Invvo metabolism Observed Rat Liver S9 metabolism Simulated	Keep in mind that the chemical list which you are insert have no a target structure. Otherwise the structure similarity will be calculated based on the target structure in the data matrix.				

Substructure similarity between single chemical and chemical list

QSAR TOOLEOX Input Profiling Custom profile Apply Yrew New Delete	Profiling Data	Category definition Category	► Report	Similarity explain Structure simi Target Chemical	2 ×
Occuments Document 1	Filter endpoint tree Structure Structure info Structure info Parameters Physical Chemical Properties Environmental Fate and Transport Cotoxicological Information Human Health Hazards Proviling			H2N CI #001 - ">= 4 H" (1/1) #010 - ">= 2 C" (1/1)	4 PubChem features (65/88) #001 - ">= 4 H" (1/1) #002 - ">= 8 H" (0/1)
Profiling methods Options J 1 Selected f Select All Invert About Options Organic functional groups Organic functional groups (nested) 0 Organic functional groups (nested) 0 Organic functional groups (Norbert Haider (checkmol) Visurdure similarity Tautomers unstable 0 Toutomers unstable Example Prioritization Scheme (PBT) P=benzoquinone precursors V Options J Observed Mammalian metabolism O Selected 1 Invert d Oocumented Observed Mammalian metabolism 0 Select All Invert	C Empiric Structure similarity		Find Find Structure similarity • Explain • Delete prediction • Explain prediction • Transfer to target • Set AOP target • Use for AOP • Copy •	#011 - ">= 4 C" (1/1) #015 - ">= 1 N" (1/1) #038 - ">= 1 CI" (0/1) #179 - ">= 1 any ring size 6" (1/1) #180 - ">= 1 any ring size 6" (1/1) #285 - ">= 1 aromatic ring" (1/1) #285 - ">= 1 aromatic ring" (1/1) #285 - "C-C" (1/1) #286 - "C-N" (1/1) #295 - "C-C" (0/1) #300 - "N-H" (1/1) #333 - "C(-C)(-C)" (1/1) #343 - "C(-C)(-C)(" (1/1) #345 - "C(-C)(-C)" (1/1) #352 - "C(-C)(-N)" (1/1)	#010 - ">= 2 C" (1/1) #011 - ">= 4 C" (1/1) #012 - ">= 8 C" (0/1) #015 - ">= 1 N" (1/1) #016 - ">= 2 N" (0/1) #034 - ">= 1 S" (0/1) #179 - ">= 1 any ring size 6" (1/1) #180 - ">= 1 saturated or aromatic #186 - ">= 2 any ring size 6" (0/1) #187 - ">= 2 saturated or aromatic #256 - ">= 1 aromatic ring" (1/1) #258 - ">= 2 aromatic ring" (0/1) #284 - "C-H" (1/1) #286 - "C-N" (1/1) #294 - "C-S" (0/1)
Observed Rat Observed Rat Selection of E	more details a xplain (1). In ilarity (2) as	about the structu the <i>Similarity ex</i> well as all simil	ure similarity <i>plain</i> form yo ar and differe	by right click over u can see the exac ent features (3) be	the result and t percentage of tween the two

molecules.

Congratulations!

- You have now been familiarized with *PubChem* substructure similarity features;
- You have compared: 1) two chemicals and 2) single chemical and chemical list with respect to PubChem substructure similarity;
- Note, proficiency comes with practice!