Nomenclature of ISIDA fragments (Fragmentor2015)

1. Nomenclature

The ISIDA fragments are a combination of a molecular graph colouration and a fragmentation of this graph calculated with the ISIDA Fragmentor2015 software. To characterize the different fragment, they are coded according to the following:

TopologicalFragmentationColourationType-BondInclusion-(LowerLength-UpperLength)-CountingType_Options

Where:

1. TopologicalFragmentation is a roman number and corresponds to the following fragmentation:
    - I – Sequences
    - II – Atom-centred fragments
    - III – Triplets

2. ColourationType is a chain of letters starting with a capital and followed by only lower case letters. The following codes have been used up to now:
    - A – Atom symbol
    - Ph – Pharmacophoric typing
    - Ep – Electrostatic potentials
    - Ba – Benson atoms

3. BondInclusion simply indicates the inclusion of bond orders in the string with a capital B. If only bonds are used then no ColourationType will appear.

4. LowerLength and UpperLength are the number of atoms to be included at minimum and maximum respectively. Note that a LowerLength=2 and UpperLength=5 will create fragments with at minimum a topological distance of 1 and maximum a topological distance of 4.

5. CountingType corresponds to the type of weight used to count the occurrences of fragments:
    - ms – micro-species (pH dependent counting)

    When none is indicated then the direct count is used (weight =1).

6. Options indicate special options used during the fragmentation and are listed below:
    - P – AtomPairs
    - R – Restricted (only for atom-centred fragments)
    - AP – AllPaths

- FC – FormalCharge representation
- MA – MarkedAtom
- MP – MarkedPair
- SF – StrictFragmentation
- AD – AllDynamic (Bonds)
- OD – OneDynamic(Bond)
- W – Wildcard

Options are separated by a hyphen (-).

Example: IIPhB(3-5)ms_P-FC

The ISIDA fragments were outlined in four publications:

- ISIDA SMF fragments are detailed in [Sub05].
- ISIDA Fuzzy Pharmacophoric Triplets (FPT) are detailed in [Fuz06] and [Fuz08].
- ISIDA Property-Labelled Fragments (IPLF) are detailed in [ISI10].

2. Definitions

1. Topological Fragmentation Types

Fragmental descriptors are obtained from fragmenting the molecular graph (2D) and counting the fragments occurrences. ISIDA descriptors include three basic patterns of fragmentation: a) Sequences, b) Atom-centred fragments and c) Triplets, which are explained in the following sub-sections.

a. Sequences (I)

Sequences are strings of successive connected atoms and/or bonds in the molecular graph. It corresponds to the shortest possible path between each pair of atoms.

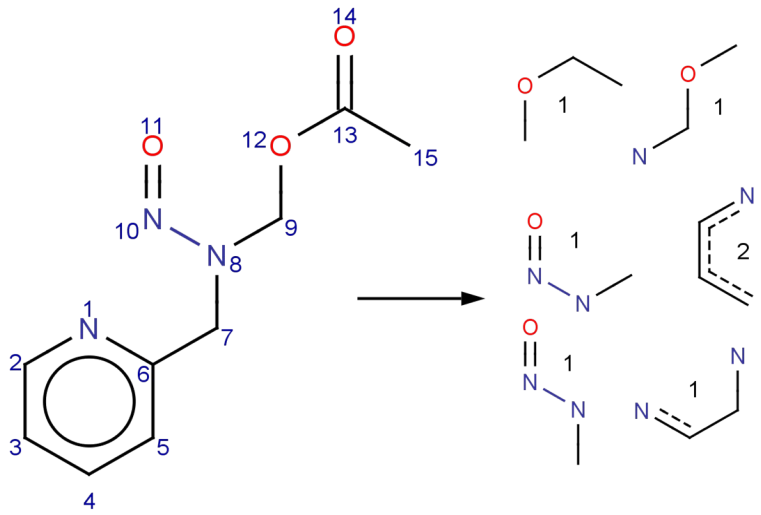


Figure 1 - Sequences of length 4 and their count from a molecular graph

b. Atom-centred fragments (II)

Atom-centred fragments start from an atom and encode the connected atoms to a certain topological distance (sphere). These include so-called neighbouring atoms (sphere =1) or augmented-atoms as well as extended augmented atoms (sphere > 1).

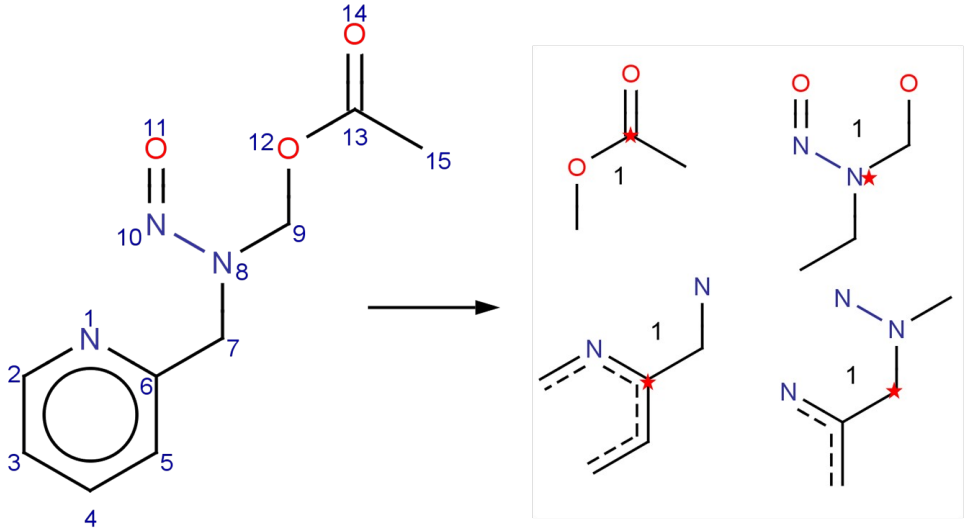


Figure 2 - Atom-centred fragments of sphere 2 and their count. The central atom is indicated by a star

c. Triplets (III)

Triplets are all the possible combinations of 3 atoms in a graph with the topological distance between each pair indicated. For example, the triplet formed by the atom number 1, 11 and 13 in Figure 2 will yield a triplet of the type: N5O5C6 where d(1,11)=5, d(11,13)=5 and d(1,13)=6.

2. Graph Colourations

Fragment descriptors are often calculated on the molecular graph with the nodes (aka vertices) indicated as the atom symbol and the edges as the bond orders. However, nodes (and even edges) may be associated to other properties, i.e. another colouration. Nodes without any colouration will be recognised as “flagless” by the fragmenting program (symbol used: §). Nodes may also have several colourations – all the combination will then be taken into account when creating fragments.

a Atom symbols (A)

The nodes are simply coloured by the atomic symbol of the atom.

d. Pharmacophoric typing (Ph)

The atomic pharmacophoric types are attributed with the ChemAxon's PMapper[Che11] according to the following rules:

- Aromatic atoms are flagged as “R”
- Carriers of positive charges are flagged as “P”
- Centres of negatively charged functional groups are flagged as “N”
- Any oxygen or nitrogen bound to a hydrogen is flagged as “D” (HB donor)
- Any oxygen or nitrogen or negative sulphide or thiourea (=S) is flagged as “A” (HB acceptor)

- Any carbon or halogen except if concerned by the rules above is flagged as “H” (hydrophobe)

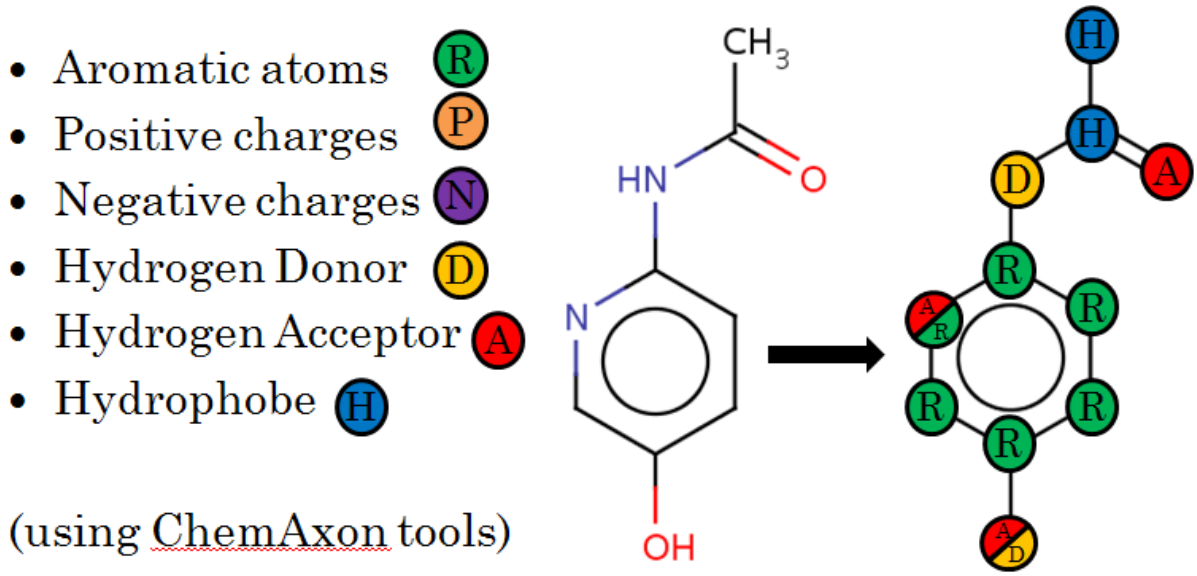


Figure 3 - Pharmacophoric rules and example of pharmacophoric graph colouration

e. Electrostatic potentials (Ep)

The electrostatic potential colouration is based on Gasteiger’s partial charges (using ChemAxon Calculation Plugin[Che111]). The electrostatic potential V_i of each atom i are calculated according to:

$$V_i = \sum_{j \neq i} \frac{q_j}{d_{ij}} + \frac{q_i}{d_0}$$

with q_j the partial charge on atom j, q_i the partial charge on atom i, d_{ij} the topological distance between atom i and j and d_0 a virtual distance to take into account the concerned atoms charges (d_0 = 0.4 after some empirical adjustments, aimed at ensuring that polar positive and polar negative heavy atoms in typical organic compounds were classified in agreement with chemical common sense).

The V values are then binned into 5 categories:

- N - negative ($V_i \leq -0.28$),
- n - slightly negative ($-0.32 < V_i \leq -0.08$),
- 0 - neutral ($-0.12 < V_i \leq +0.12$),
- p - slightly positive ($+0.08 < V_i \leq +0.32$) and
- P - positive ($V_i \geq +0.28$)

The overlapping bins permit ambiguous cases to be represented by both flags (for example V_i=-0.3 would return a “N/n” flag for atom i. Both possible patterns corresponding to flags N and n for that atom will be generated).

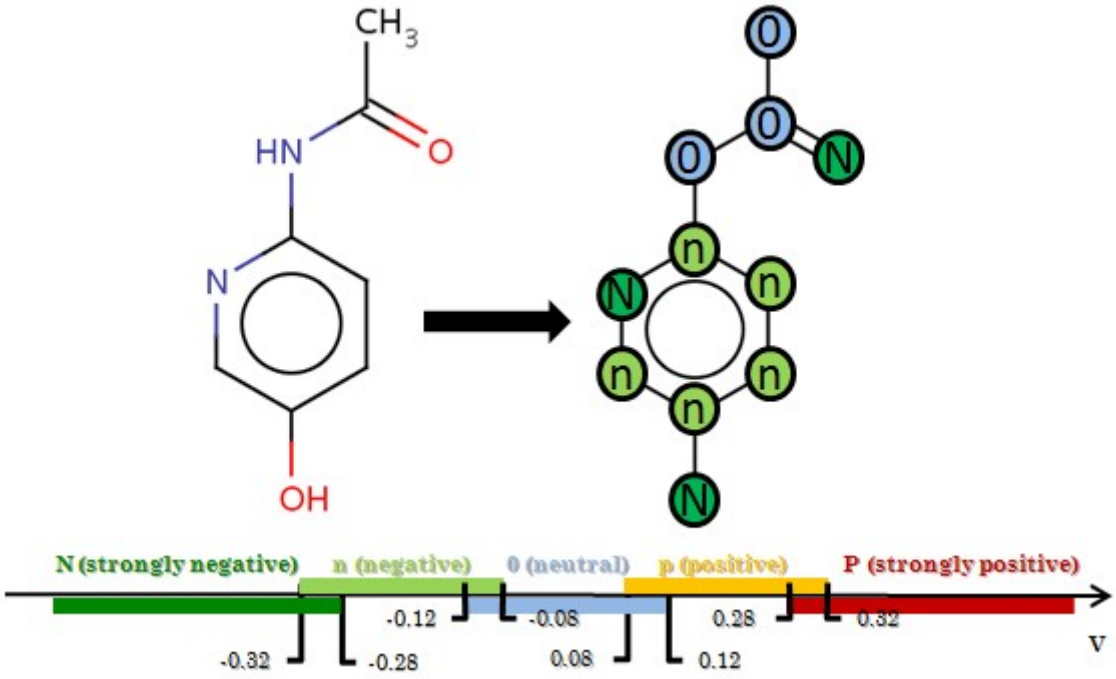


Figure 4 - Electrostatic potential colouring example and binning

f. Benson atoms

The Benson atoms colouration was used in older versions of the fragmentor and encodes the different carbon types. The following table gives the different possibilities in their priority order:

Table 1 - Benson Atom typing

Priority	Type of atom	Benson code
1.	Aromatic C	CB
2.	Triple-bonded CN (C#N)	CN
3.	Triple-bonded C (C#)	CT
4.	Twice double-bonded C (=C=)	CA
5.	Ketone (C=O)	CO
6.	Double-bonded C (C=)	CD

If a carbon can be classified in two of the types, the most important one (priority number =smallest) will be used.

3. Bonds

The inclusion of the bond order information in the fragments is indicated by a capital B following the graph colouration code. The following table sums up the bond order types, which also include the dynamic bonds of condensed graphs of reaction, found in the descriptors:

Bond Type	Symbol	Bond Type	Symbol
Simple	-	Single bond cut	18
Double	=	Double bond cut	28
Triple	+	Triple bond cut	38
Aromatic	*	Aromatic bond cut	48
Single or Double	5	Single bond to double bond	12
Single or Aromatic	6	Single bond to triple bond	13
Double or aromatic	7	Single bond to aromatic bond	14
Any bond type	?	Double bond to single bond	21
Special bond type II	_	Double bond to triple bond	23
Single bond in cycle	.	Double bond to aromatic bond	24
Double bond in cycle	:	Triple bond to single bond	31
Triple bond in cycle	#	Triple bond to double bond	32
Special bond type I – Hydrogen bonds	~	Triple bond to aromatic bond	34
Single bond creation	81	Aromatic bond to single bond	41
Double bond creation	82	Aromatic bond to double bond	42
Triple bond creation	83	Aromatic bond to triple bond	43
Aromatic bond creation	84	Unknown bond	YY

4. Fragment length

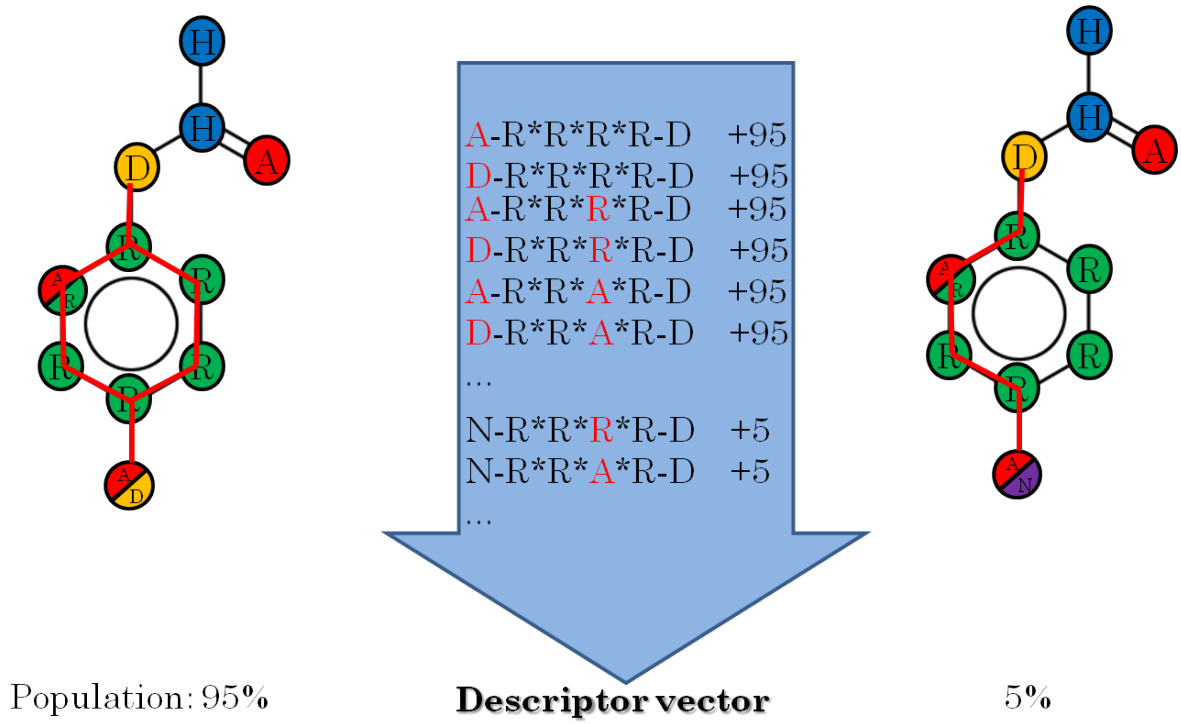
The lower boundary length and upper boundary length are used to define the different lengths of fragments.

5. Counting Type

At the moment, only two types of weights can be used:

- Standard count: the weight = 1 and the count corresponds to the number of occurrences of a fragment.
- ms = micro-species dependent count. Each micro-species at a defined pH (usually 7.4) is defined in the SDF with a specific weight which corresponds to the population level (in %) of the micro-species. The occurrence of fragments in each micro-species will be multiplied by the weight and added to the total count (Refer to Figure 5).

Figure 5 - Micro-species population level dependent counting



6. Other Options

a Atom Pairs (P)

Only the extremities of the fragment are kept and the topological distances between those are indicated. It can only be used with type I and II fragments:

I. In the case of sequences, the result is an atom pair and is represented by the colourations of the first and the last atom in the sequence and separated by the topological distance between them.

II. In the case of extended augmented atoms, the result will be a multiplet including the centre atom and the leafs separated by their topological distance. At radii 0 and 1, they correspond to extended augmented atoms without AP option.

This option can be used jointly with bond information inclusion.

g. Restricted Atom-Centred Fragments (R)

This option can only be used with extended augmented atoms (II) and the resulting fragments differ by a restriction on the sequences' path lengths from the centre: only paths of the defined length are represented.

h. All Paths exploration (AP)

By default, sequences correspond to the shortest paths possible between two atoms. When the AP option is used, all the possible paths between the two atoms are kept as fragments.

i. Formal Charge Representation (FC)

Formal charges of atoms are represented in the fragments with the addition of "_FC"+Formal Charge value. In a sense, it is an extension of the atom colouration to include more chemical relevant properties.

j. Marked Atom (MA)

Only fragments starting/ending from the marked atom will be generated.

k. Marked Atom Pair (MP)

Only sequences between two marked atoms will be generated. This option can only be used with sequences (I) and all the possible pairs of marked atoms will be considered.

l. Strict Fragmentation (SF)

The descriptors have been restricted to a set of descriptors defined by the user. In practice, usually a previous fragmentation was done and used to limit the generation of descriptors on other molecules.

m. All Dynamic Bonds (AD)

In the case of Condensed Graph of Reaction (CGR), special dynamic bonds are defined. With the AD option, only fragments containing **only** dynamic bonds are outputted.

n. One Dynamic Bond (OD)

Similar to the AD option, the OD option will output only fragments containing at least 1 dynamic bond.

o. Wildcard (W)

Intermediate atoms in fragments are systematically replaced by a "wildcard flag" (?) in all possible combination between the fragment with all flags and its paired counterpart. For example, the sequence H-H-H-H-A will display all possibilities of replacing the intermediate atoms: H-?-H-H-A, H-H-?-H-A, H-H-H-?-A, H-?-?-H-A, H-H-?-?-A, H-?-?-?-A. The last fragment corresponds to the paired sequence.

3. Examples:

1. IAB(2-5) corresponds to sequences of atom symbols and bonds ranging from 2 atoms to 5 atoms in the string (corresponds to 1 to 4 bonds between the two extremities).
2. IIPh(3-6)ms corresponds to pharmacophoric coloured extended augmented atoms with all possible paths between the central atom and the extremity taken into account, with length of 3 to 6 bonds between them.
3. IIIPh(2-8)ms correspond to pharmacophoric coloured triplets where each summit are separated at minimum by 1 bond and at maximum by 7 bonds.

4. References

1. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comp.-Aid. Mol. Design*. 2005, Vol. 19, 9-10, pp. 693-703.

2. Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* 2006, Vol. 46, 6, pp. 2457-2477.

3. Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* 2008, Vol. 48, 2, pp. 409-425.

4. ISIDA Property-Labelled Fragment. *J.Mol. Inf.* 2010, Vol. 29, 12, pp. 855-868.

5. ChemAxon PMapper 5.7.0. [En ligne] 2011. [Citation : 22 11 2011.] http://www.chemaxon.com/jchem/doc/user/PMapper.html.

6. *ChemAxon Calculation Plugin Cxcalc 5.7.0.* [En ligne] http://www.chemaxon.com/marvin/help/calculations/calculator-plugins.html.