Section III

Cheminformatics - Basics: *Molecular Similarity (or Diversity)*

Dan C. Fara and Tudor I. Oprea

1. What is molecular similarity (or diversity)?

Molecular similarity is a complex notion that can only be described with reference to the immediate use for which it is intended. The “Molecular Similarity Principle”[1] states that structurally similar molecules tend to have similar properties – physicochemical as well as biological ones – more often than structurally dissimilar molecules. This concept has found extensive use in lead discovery and compound optimisation. For example, in designing libraries of compounds for lead generation, one approach is to design sets of compounds ‘similar’ to known active compounds in the hope that alternative molecular structures are found that maintain the properties required while enhancing e.g. patentability, medicinal chemistry opportunities or even in achieving optimised pharmacokinetic profiles.

Molecular similarity encompasses (i) bonding patterns, (ii) atomic positions, (iii) conformation, (iv) shape and (v) spatial disposition of molecular properties. The reverse of similarity is complementarity. In between lays molecular dissimilarity (or diversity). All notions of similarity involve perception of patterns and subsequent attempts to classify and quantify the patterns. Complementarity descriptions are equivalent to trying to express the relationship of the fit between a cast and its mould. If similarity can be expressed by a correlation coefficient lying between 0 and 1, then complementarity can be related on the same scale but with coefficients lying between 0 and -1.

The similarity research methods can be divided into those which are based on (i) discrete properties, and those which are founded on (ii) continuous properties. For example, a pharmacophore may be defined by three or more atoms separated by a set of distances; there may be some tolerances on the distances but the general pattern is discrete. A searching algorithm then finds only those structures from a database that satisfy the pharmacophore query. In a continuous space problem, the user may wish to optimize the overlap of a continuous field such as the electrostatic potential projected onto defined molecular surfaces. The result can then be expressed by a similarity coefficient.

2. Fragment descriptors for molecular similarity calculations

Similarity searching in large chemical databases needs representations of the molecules that are both (i) **effective**, i.e., can differentiate between molecules that are different, and (ii) **efficient**, i.e., quick to calculate, in operation. There is a general conflict between these two requirements in that the most effective methods of representation tend to be the least efficient to calculate, and *vice versa*, so a suitable compromise needs to be made. For instance, quantum-mechanical descriptions, such as the electron probability density function described by Carbo and Calabuig[2], take too long to calculate whereas, at the other extreme, simple atom and bond counts are generally too trivial to discriminate between many molecules. In the middle lay descriptors based on 2D and 3D substructural fragments or properties – see Figures 1 and 2.

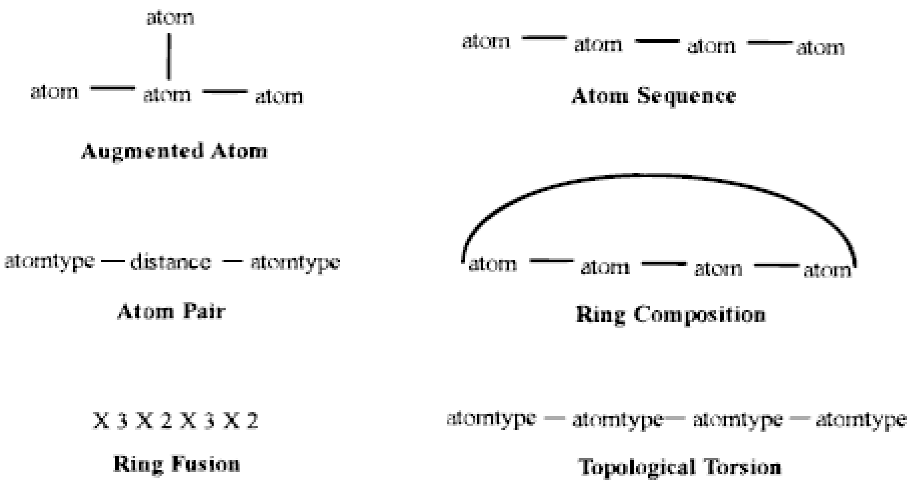

atom
atom — atom — atom
Augmented Atom
atom — atom — atom — atom
Atom Sequence
atomtype — distance — atomtype
Atom Pair
atom — atom — atom — atom
Ring Composition
X 3 X 2 X 3 X 2
Ring Fusion
atomtype — atomtype— atomtype — atomtype
Topological Torsion

Figure 1. Fragment 2D descriptors – examples (taken from reference[3])

These are the ones that are curently most commonly used for similarity searching. As can be seen, many of the 2D fragments that can be generated from a 2D connection table have equivalents in the 3D fragments that can be generated from a 3D connection table. However, there is much less consensus in the 3D area as to which are the best descriptors to use; moreover, the variety of available descriptors is greater, and new fragment types continue to be developed. Due to the fully connected nature of a 3D connection table, and the flexibility of 3D structures, the number of 3D fragments generated for a given class can be much larger than for the 2D equivalent, and the generation process can be more time-consuming.

However, given the wide variety of descriptor types available, it is necessary to select the most appropriate structural representation for a given application. Obviously, 2D fragment descriptors contain a lot of information about the physical properties and reactivity of a molecule and are quick to calculate. Augmented atoms are much localized, atom sequences are less so, and atom pairs span the whole of a 2D structure – see Figure 1. Ring descriptors are essential for cyclic structures, and topological torsions correlate well with 3D torsions except in highly folded molecules. Physical properties and topological indices can be useful representations of hydrophobic and electrostatic interactions. 3D shape descriptors can give useful information about dispersion and steric interactions. The larger 3D descriptors tend to be the most effective, but the flexibility of many molecules can increase the time required to generate 3D descriptors and/or can also decrease their effectiveness. Generalization of atom types from specific elements to groups or properties can help to reveal broader similarities. Overall, the trend is to use combined descriptors or descriptor sets which contain many different descriptor types, at many different level of generalization.

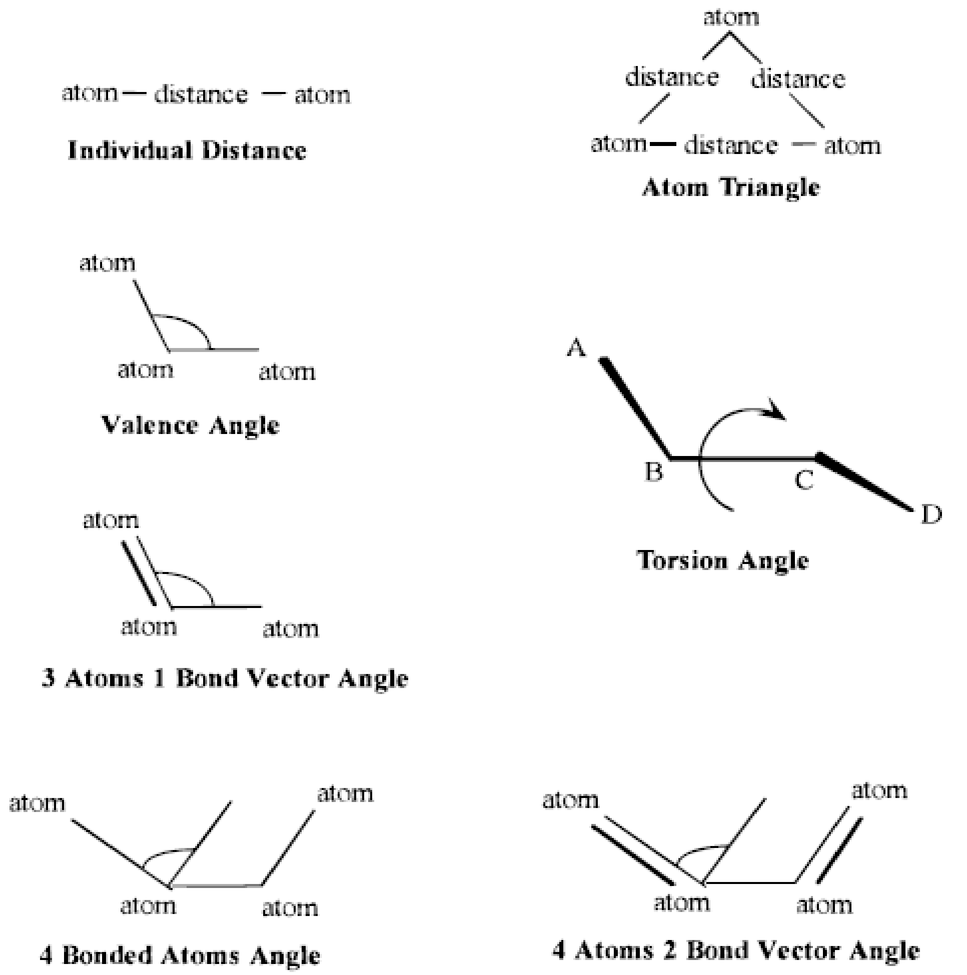


Figure 2. Fragment 3D descriptors – examples (taken from reference[3])

An alternative to these algorithmically generated descriptors is to define and to search for particular functional groups, which may be expressed in specific or general terms. One defined, the functional groups can be detected by scanning the connection tables for instances of them. A more efficient way is to use string-searching of a linear representation of the molecule, for instance by defining the functional groups in terms of SMARTS strings and using them to search SMILES representations[4].

3. Fragment descriptor encoding

The descriptors have to be encoded in order to enable similarity calculations to be carried out. The representation that is overwhelmingly used as a basis for similarity calculations in large databases is the fingerprint (the fixed-length bit-string) – see details on *Molecular Descriptors and Fingerprints*, section 2. Fingerprints.

4. Similarity and distance coefficients

At the center of any similarity search is the measure that is used to quantify the degree of resemblance between the reference structure and each of the structures in the database (real or virtual) that is being screened. A similarity measure comprises three components: the representation that is used to characterize the molecules that are being compared; the weighting scheme that is used to assign differing degrees of importance to the various components of these representations; and the coefficient that is used to determine the degree of relatedness between two structural representations.

Some coefficients are measures of the distance, or dissimilarity between structures (and have a value of 0 for identical structures), while others measure similarity directly (and have their maximum value for identical structures). In most cases the values that can be taken by a coefficient lie in the range from 0 to 1 or can be normalized to that range: this is typically effected by means of a function based on the values of the attributes for the two structures that are being compared, with the resulting coefficients being referred to as *association coefficients*. The zero-to-unity range provides a simple means for converting between a similarity coefficient and a complementary distance coefficient, namely, subtraction from unity. In some cases a similarity coefficient and its complement have been developed independently and are known by different names; i.e., the Soergel distance coefficient is the complement of the Tanimoto (or Jaccard) association coefficient.

Below are given the equations of the most common similarity (S) and distance (D) coefficients used for similarity searches[3]:

Euclidean: $EuD_{A,B} = (a + b - 2c)^{1/2}$ (1)

Hamming: $HaD_{A,B} = (a + b - 2c)$ (2)

Tanimoto: $TaS_{A,B} = c / (a + b - c)$ (3)

Cosine: $CoS_{A,B} = c / (ab)^{1/2}$ (4)

Tversky: $TvS_{A,B} = c / [\alpha(a - c) + \beta(b - c) + c]$ (5)

Where, a - unique bits turned on in molecule "A", b - unique bits turned on in molecule "B", c - common bits turned on in both molecule "A" and molecule "B", d - common bits turned off in both molecule "A" and molecule "B", n - the total number of bits in the fingerprint, and α, β - user-defined constants.

The Tanimoto, Euclidean, Hamming, and Cosine are all symmetric measures (equations 1 to 4). This means that the comparison of A to B yields the same number as the comparison of compound B to compound A. For the Tversky measure this is not the case (equation 5). The Tversky similarity coefficient $TvS_{A,B}$ does not necessarily equal the Tverksy $TvS_{B,A}$, which makes the measure asymmetric. The reason for this can be explained by the role of the two parameters, α and β. The usual user constraint is $\alpha + \beta = 1$. If the choice of α is set close to 1, e.g. 0.95 then $\beta = 0.05$. The $TvS_{A,B}$ for this choice would be close to 1.0 (the maximum) if the unique bits of compound A (when compared to compound B) are nearly zero. At high $TvS_{A,B}$ values, the interpretation is that compound A "fits into" compound B. If the choice of α is 0.5, then $\beta = 0.5$ and the $TvS_{A,B}$ is monotonic to the Tanimoto $TaS_{A,B}$. The Tversky measure therefore enables the grouping of chemical structures of different sizes that share features of the smaller of the two.

Figure 3 shows an example of a continuous similarity search for thrombin inhibitors in which the structural similarity of active compounds gradually ‘fades away’ when departing from known leads. Such similarity searches are consistent with the ‘Molecular Similarity Principle’ and the holistic view of molecular similarity.

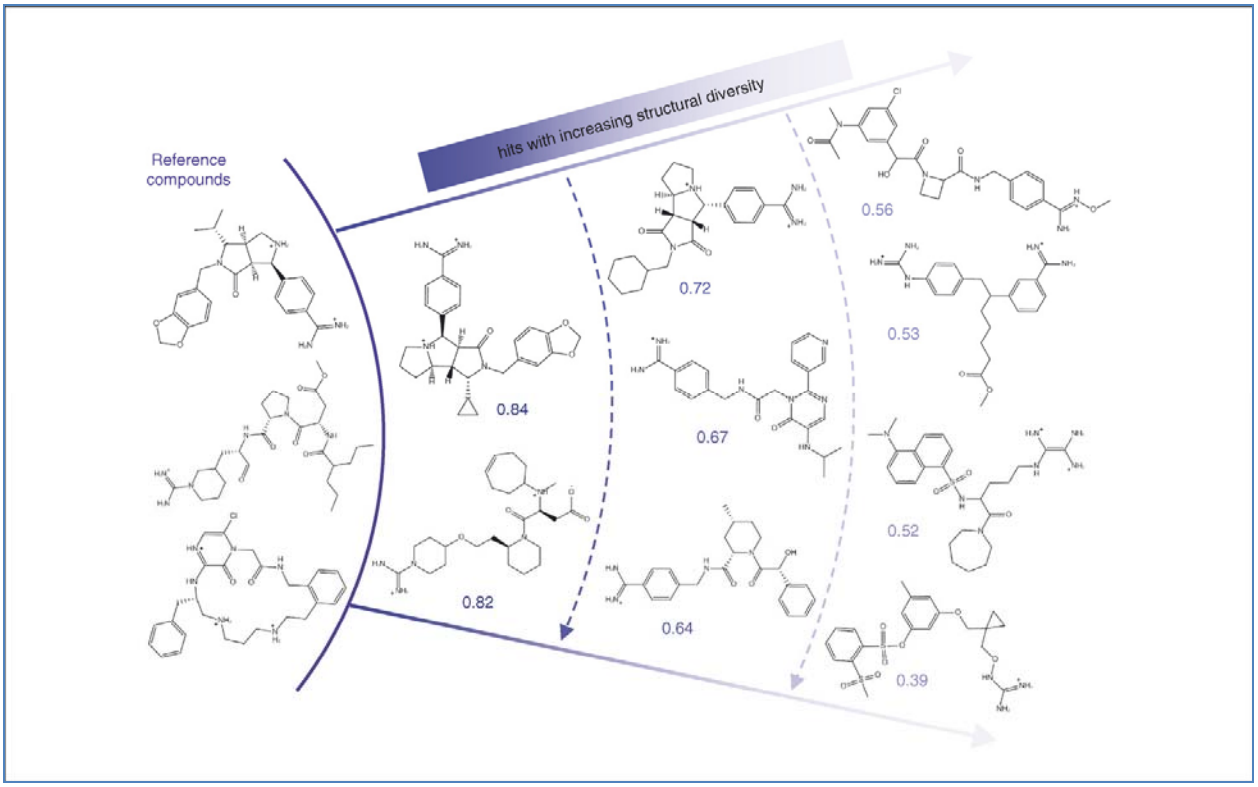


Figure 3. Structural spectrum of thrombin inhibitors. Starting with known inhibitors of thrombin, a sequence of hits ranging from close analogs to increasingly diverse structures was identified in a simulated LBVS campaign in a large screening database containing ~ 1.4 million compounds. Left: three of the five reference molecules. Right: examples of hits identified in a selection set of 250 database compounds. Hits are arranged in layers of increasing structural diversity (top down, from left to right). As a measure of structural similarity, the Tanimoto coefficient[5] is reported for each hit relative to the most similar of the five reference molecules. Reference molecules and hits were compared using a fingerprint consisting of the publicly available set of 166 MACCS structural keys (MDL Elsevier) – taken from reference[6].

5. Application – similarity search using RoadRunner

On the following exercise you will perform both similarity and batch similarity searching on the RoadRunner database using (i) dihydrotestosterone, (ii) estradiol, and (iii) 4-hydroxytamoxifen as query compounds. Roadrunner is a chemical database application developed and supported by the Informatics Core of the New Mexico Molecular Libraries Screening Center at the University of New Mexico Health Sciences Center.

- Go to Roadrunner
- Login into your account

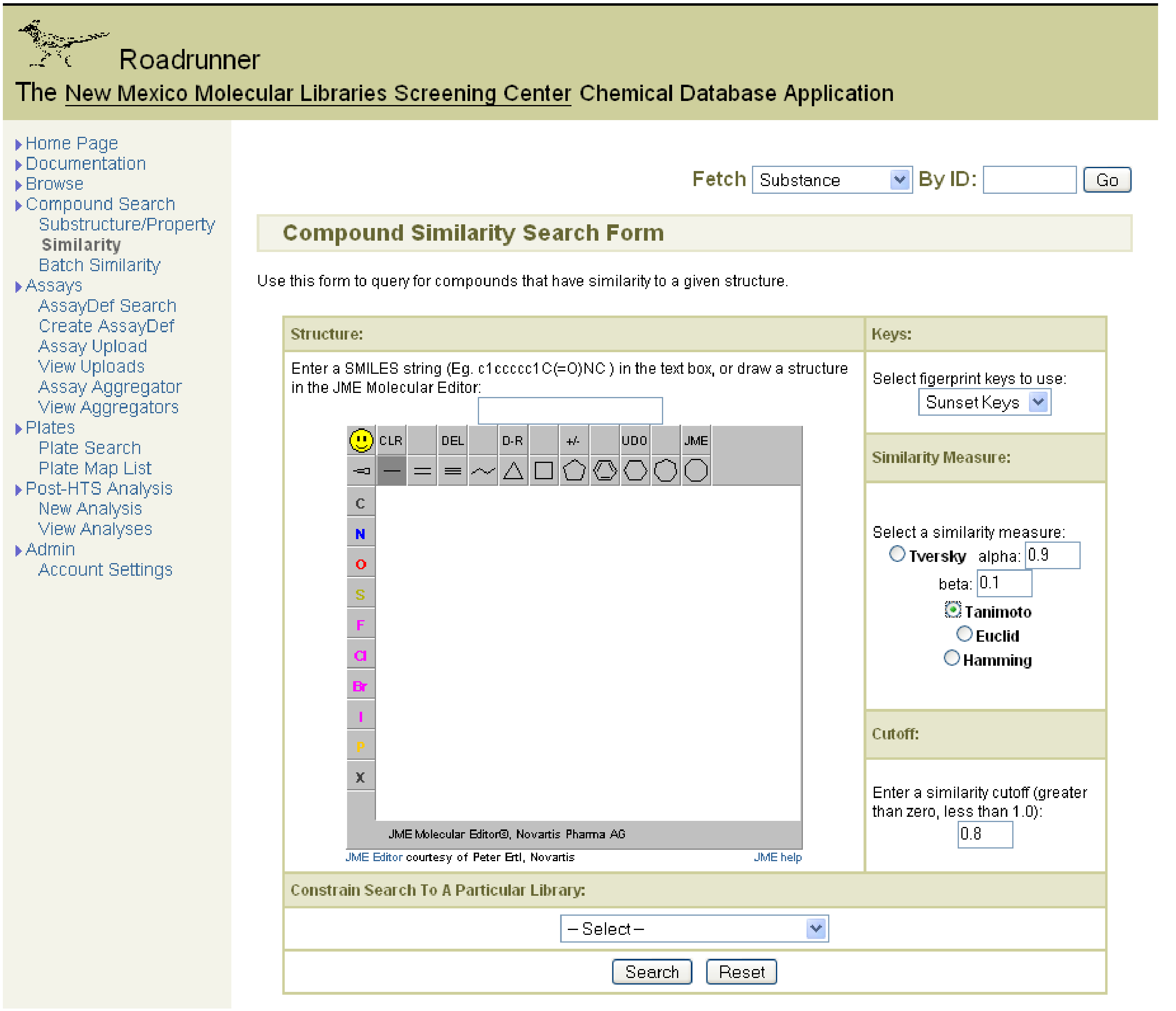
Similarity

- Go to *Compound Search* > *Similarity* and draw the structure in the Java Molecular Editor, or enter the SMILES string directly – see Figure 4.
- Select the fingerprint keys to be used: (i) Sunset or (ii) MDL
- Select the similarity measure: (i) Tanimoto, (ii) Tversky, (iii) Euclid, or (iv) Hamming
- Enter the similarity cutoff
- Select the library
- Click *Search*
- Analyze the results

For each of the query use the following:

1. First Sunset keys, and second MDL keys
2. Similarity measure: (i) Tanimoto, and (ii) Tversky
3. Cutoff: (i) 0.9, (ii) 0.8, (iii) 0.7 for each of the above combinations
4. Library: MLSMR (222107 compounds)

Quiz 1: Are the results (hits) influenced by the (i) type of the similarity measure, and (ii) the cutoff values? Please explain.



Roadrunner

The New Mexico Molecular Libraries Screening Center Chemical Database Application

Fetch Substance By ID:

Compound Similarity Search Form

Use this form to query for compounds that have similarity to a given structure.

Structure:

Enter a SMILES string (Eg. c1ccccc1C(=O)NC) in the text box, or draw a structure in the JME Molecular Editor:

JME Molecular Editor©, Novartis Pharma AG

JME Editor courtesy of Peter Ertl, Novartis

Keys:

Select figerprint keys to use: Sunset Keys

Similarity Measure:

Select a similarity measure:

○ Tversky alpha: 0.9 beta: 0.1

● Tanimoto

○ Euclid

○ Hamming

Cutoff:

Enter a similarity cutoff (greater than zero, less than 1.0): 0.8

Constrain Search To A Particular Library:

-- Select --

Figure 4

Batch similarity

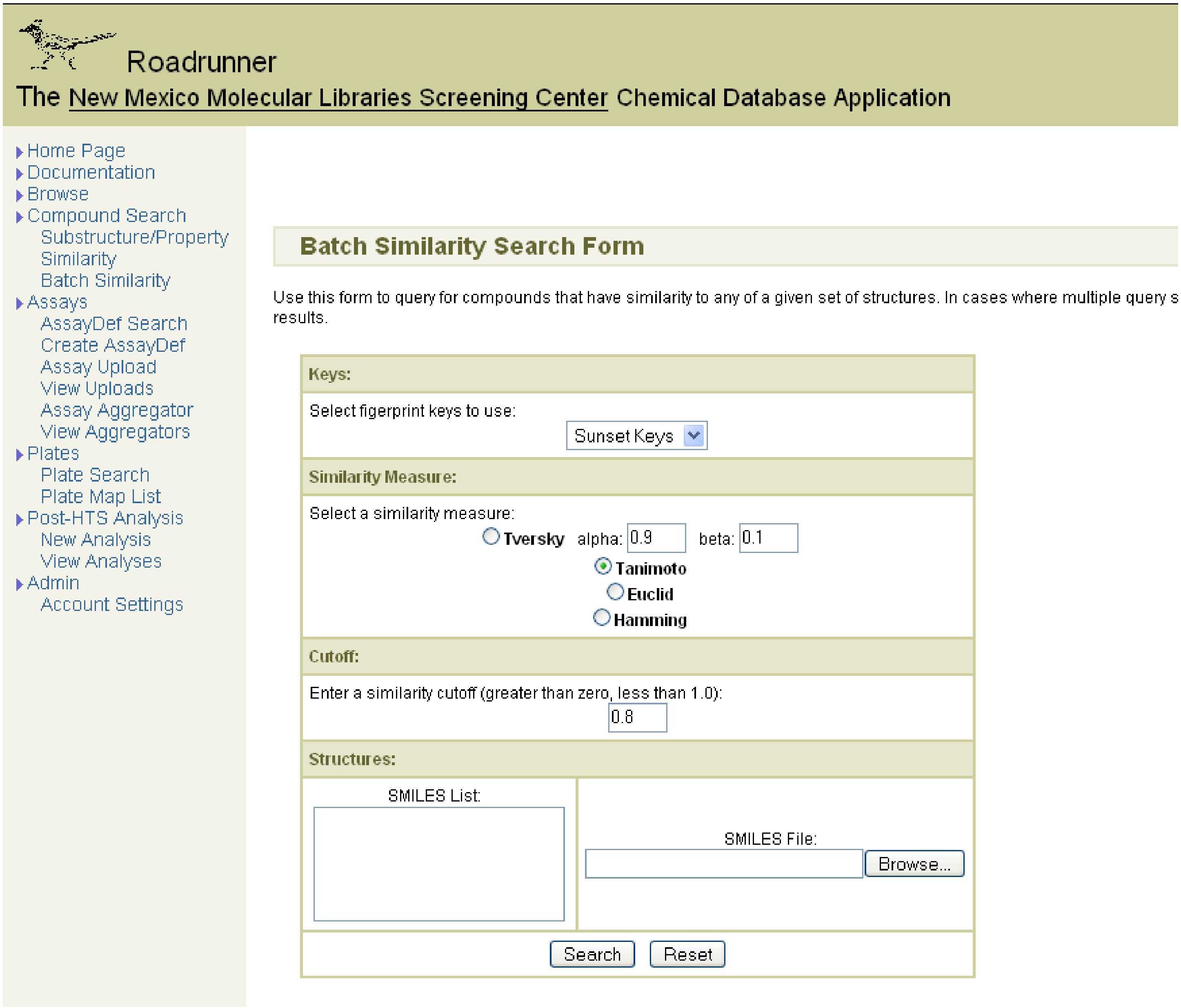
- Go to *Compound Search* > *Batch Similarity* – see Figure 5.
- Select the fingerprint keys to be used: (i) Sunset, or (ii) MDL
- Select the similarity measure: (i) Tanimoto, (ii) Tversky, (iii) Euclid, or (iv) Hamming
- Enter the similarity cutoff

- Enter the queries: (i) copy and paste all 3 SMILES, or (ii) upload a text file that contains the 3 SMILES
- Click *Search*
- Analyze the results

For each of the query use the following:

1. Sunset keys
2. Similarity measure: (i) Tanimoto
3. Cutoff: (i) 0.9, (ii) 0.8, (iii) 0.7

Quiz 2: Compare the results with those previously obtained (Similarity). Are there any differences?



Batch Similarity Search Form

Use this form to query for compounds that have similarity to any of a given set of structures. In cases where multiple query s
results.

Keys:

Select figerprint keys to use: Sunset Keys

Similarity Measure:

Select a similarity measure:

○ **Tversky** alpha: 0.9 beta: 0.1

● **Tanimoto**

○ **Euclid**

○ **Hamming**

Cutoff:

Enter a similarity cutoff (greater than zero, less than 1.0): 0.8

Structures:

SMILES List:

SMILES File:

Figure 5.

[1] Johnson, M.; Maggiora, G.M. eds (1990) Concepts and Applications of Molecular Similarity, John Wiley & Sons

[2] Carbo, R.; Calabuig, B. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 600-606.

[3] Willett, P.; Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.

[4] Brown, R.D.; Martin, Y.C. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 572-584.

[5] Willett, P. *J. Med. Chem.*, **2005**, *48*, 4183–4199.
[6] Eckert, H.; Bajorath, J. *Drug Discov. Today*, **2007**, *12*, 225-233.