SIMILARITY – DISSIMILARITY COMPARISON STRUCTURAL FINGERPRINTS TUTORIAL

Authors: William J. Welsh, Ph.D.
Vladyslav Kholodovych, Ph.D.
University of Medicine & Dentistry of New Jersey
Robert Wood Johnson Medical School
675 Hoes Lane
Piscataway, NJ 08854 U.S.A.
(732) 235-3229 phone
(732) 235-3475 FAX
kholodvl@umdnj.edu
http://www2.umdnj.edu/~kholodvl

UMDNJ Fundamentals of Bioinformatics

SIMILARITY – DISSIMILARITY COMPARISON STRUCTURAL FINGERPRINTS TUTORIAL

Dr. William J. Welsh,
welshwj@umdnj.edu

and

Dr. Vladyslav Kholodovych
kholodvl@umdnj.edu

The following exercise will use MOE to calculate the structural fingerprints for a database of compounds using the MACCS Structural Keys, to calculate their similarity using the Tanimoto Coefficient, and then to sort them into separate clusters based on their similarity-dissimilarity.

Instructions

1. Create New Directory (e.g., Similarity)
2. Download the following four files from *www2.umdnj.edu/~kholodvl*:
   Database **ER.mdb**; and queries **q1**; **q2**; **q3.**
3. Start MOE
4. Open the database called **ER.mdb** in the MOE Database Viewer.

   This database contains 49 steroidal and non-steroidal compounds (ligands) that are known to exhibit binding affinity for the estrogen receptor. In MOE, it might be interesting to visualize and compare these compare on the screen.

5. In Database Viewer, select: Compute->Fingerprints (a window will open);
   Select "FP:MACCS Structural Keys" (not "Bit packed"), then hit "OK", and a new field called "FP:MACCS" will be created and displayed.

   Double-click on "Field MACCS". A new window will open, and inside it you will see the various structural keys that are present in any particular compound that you select.

   In the Database Viewer, MOE will show you which structural keys are found in each compound. For example, Compound #1 (located in first row) will show 32 keys present: 66 90 91 96 …. 162 163 164 165.

   Note: MACCS contains a total of 166 Structural Keys, corresponding to various groups (e.g., aromatic-C-aromatic) contained in organic compounds.

6. In Database Viewer, select: Compute-> Cluster Codes -> Fingerprint Based (window will open);

   Select "Set Fingerprint" (window will open): select "MACCS Structural Keys" for Fingerprint and "Tanimoto" for Similarity Metric (then hit "OK").

   Hit "OK" again to run the clustering analysis.

7. In Database Viewer, select: Compute -> Sort by Cluster (then hit "OK")

   Notice that compounds with similar structures are placed in the same cluster. For example, visually inspect and compare compounds in the following three clusters: #22, #40. All compounds in Cluster #22 will be similar to each other, and all compounds in Cluster #40 will be similar to each other. However, compounds in Cluster #22 will differ from compounds in Cluster #40.

8. In MOE Main Window, clear the screen by pressing the "Close" button, then hit "OK".

9. Open file **q1**, which contains the first Query compound. The compound will be displayed on the viewing screen.

10. Now calculate the Fingerprint for **q1**: In Database Viewer, select "Entry"->"Add Entry" (new window will open; leave all fields in their default values, and hit "OK"). A new entry will be added to the database. Select this entry, then select "Computer"->"Fingerprints" (FP:MACCS again). Check the "Selected Entries Only" field. Hit "OK".

11. In Database Viewer, select File->Similarity Search (window will open). In this window, select "Set Fingerprint" (window will open): select "MACCS Structural Keys" for Fingerprint and "Tanimoto" for Similarity Metric (then hit "OK").

    Select the button "Show Hits" for Visibility.
    Hit "Search" to run the similarity search.

12. You should find 5 similar compounds for **q1** in the Database Viewer. Manually calculate the Tanimoto Coefficient (TC) between **q1** and each of these 5 similar compounds. To view the fingerprints for any compound, double click on the FP:MACCS field for that compound. Create a Table in Word or Excel, and record the fingerprints for each of these compounds.

    The TC is calculated by comparing the fingerprints between any pair of compounds, and applying the Tanimoto Equation as given in class. For example, if Compounds A and B have the following structural keys

    A: **66** 67 **92 105** 111 145 **151 160 161 162** 163 164 165
    B: **66** 87 **92 105** 132 144 **151 160 161 162**

    Then TC = [7/(13 + 10 – 7) = 0.4375 (or, simply, 0.44).

Remember that the default or standard criterion for considering two compounds as "similar" is TC ≥ 0.85. Consequently, Compounds A and B in the example are not very similar.

13. Return to Step 8, and repeat the same process (Steps 8-12) for **q2** and **q3**.

14. Each Table (in Excel or Word) should show the Query (e.g., q1) and its Fingerprint, together with the Fingerprint for each similar compound. In the last column of the Table, enter the TC. See below for a typical representation of each Table.

Compound	**Fingerprint List**	**Number of Fingerpints**	**Number of Common Fingerprints with Query**	**Tanimoto Coefficient (TC)**
Query 1	52, 65, …162	22	-	-

Questions:

A. How many clusters were obtained in Step 7 above? Explain what you see by visually comparing any two compounds with the same cluster. Explain what you see by visually comparing any two compounds in different clusters.
B. Explain what is meant by the Tanimoto Coefficient (TC). What would be the TC for two compounds that share 9 out of 10 structural keys?
C. Although clearly indicated, MOE employs the Jarvis-Patrick method for clustering. Give a brief explanation of the Jarvis-Patrick clustering method. Can you name at least one other well-known clustering method?
D. For query **q1**, what cluster is most similar to it? (Cluster #1? Cluster #15?) How many compounds are contained in this cluster? How similar is **q1** to each compound in this cluster? Hint: Look at the TC values.
E. For query **q2**, what cluster is most similar to it? (Cluster #1? Cluster #15?) How many compounds are contained in this cluster? How similar is **q2** to each compound in this cluster? Hint: Look at the TC values.
F. For query **q3**, what cluster is most similar to it? (Cluster #1? Cluster #15?) How many compounds are contained in this cluster? How similar is **q3** to each compound in this cluster? Hint: Look at the TC values.