Molecular Similarity Methods

Courtesy of Prof. Jürgen Bajorath, University of Bonn
Each chosen descriptor adds a dimension to the reference space.

Calculation of $n$ descriptor values produces an $n$-dimensional coordinate vector in descriptor space that determines the position of a molecule:

$$\text{molecule } M_i = (\text{descriptor}_1(i), \text{descriptor}_2(i), \ldots, \text{descriptor}_n(i))$$
Descriptor spaces of variable design are utilized as chemical reference spaces for many chemoinformatics applications.

descriptor$_1$=0.1 for the two red molecules

descriptor$_n$=5.6 for the three green molecules

e.g. hypothetical $n$-D space
Projecting Molecules

- Molecules are projected into the descriptor space based on their descriptor values.

- Molecules having the same value for a descriptor should be projected to the same interception plane.

For example, in a hypothetical $n$-D space:

- $\text{descriptor}_1 = 0.1$ for the two red molecules.
- $\text{descriptor}_n = 5.6$ for the three green molecules.
**Chemical Reference Space**

- Distance in chemical space is used as a measure of molecular “similarity“ and “dissimilarity“

- “Molecular similarity“ covers only chemical similarity but also property similarity including biological activity
Distance Metrics in n-D Space

- If two molecules have comparable values in all the \( n \) descriptors in the space, they are located close to each other in the \( n \)-D space.
  - how to define “closeness” in space as a measure of molecular similarity?
  - distance metrics
Euclidean Distance in n-D Space

- “Ordinary distance”
- Each of $n$ variables (descriptors) is a dimension in $n$-dimensional space
Euclidean Distance in n-D Space

- Given two \( n \)-dimensional vectors, \( \mathbf{A} \) and \( \mathbf{B} \)
  - \( \mathbf{A} = (a_1, a_2, \ldots a_n) \)
  - \( \mathbf{B} = (b_1, b_2, \ldots b_n) \)

- Euclidean distance \( D_{AB} \) is defined as:
  \[
  D_{AB} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}
  \]

- Example:
  - \( \mathbf{A} = (3, 0, 1) \); \( \mathbf{B} = (5, 2, 0) \)
  - \( D_{AB} = \sqrt{(3 - 5)^2 + (0 - 2)^2 + (1 - 0)^2} = 3 \)
Manhattan Distance in n-D Space

- Given two $n$-dimensional vectors, $\mathbf{A}$ and $\mathbf{B}$
  - $\mathbf{A} = (a_1, a_2, \ldots, a_n)$
  - $\mathbf{B} = (b_1, b_2, \ldots, b_n)$

- Manhattan distance $D_{AB}$ is defined as:

$$D_{AB} = \sum_{i=1}^{n} |a_i - b_i|$$

- Example:
  - $\mathbf{A} = (3, 0, 1)$; $\mathbf{B} = (5, 2, 0)$
  - $D_{AB} = |3 - 5| + |0 - 2| + |1 - 0| = 5$
Descriptor-based Similarity

- When two molecules A and B are projected into an $n$-D space, two vectors, $\mathbf{A}$ and $\mathbf{B}$, represent their descriptor values, respectively.
  - $\mathbf{A} = (a_1, a_2, ..., a_n)$
  - $\mathbf{B} = (b_1, b_2, ..., b_n)$

- The similarity between A and B, $S_{AB}$, is negatively correlated with the distance $D_{AB}$
  - shorter distance $\sim$ more similar molecules
  - in the case of normalized distance (within value range $[0,1]$), similarity $= 1 – \text{distance}$

$\text{e.g. } D_{AB} > D_{BC} \Leftrightarrow S_{AB} < S_{BC}$
Descriptor-based Similarity: Example

- Four molecules in four-dimensional space
  - descriptor values listed in the table

<table>
<thead>
<tr>
<th>mol</th>
<th>logP(o/w)</th>
<th>b_rotN</th>
<th>a_acc</th>
<th>a_don</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.4250</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3.4700</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>0.7090</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2.4900</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

- logP(o/w): index of hydrophobicity
- b_rotN: number of rotatable bonds
- a_acc: number of hydrogen acceptors
- a_don: number of hydrogen donors
Descriptor-based Similarity: Example

- Euclidean distance between four molecules

- \( D_{12} = \sqrt{(2.4250 - 3.4700)^2 + (2 - 2)^2 + (3 - 3)^2 + (2 - 3)^2} = 1.45 \)

- \( D_{13} = \sqrt{(2.4250 - 0.7090)^2 + (2 - 0)^2 + (3 - 2)^2 + (2 - 1)^2} = 2.99 \)

- \( D_{14} = \sqrt{(2.4250 - 2.4900)^2 + (2 - 1)^2 + (3 - 2)^2 + (2 - 1)^2} = 1.73 \)

- \( D_{34} = \sqrt{(0.7090 - 2.4900)^2 + (0 - 1)^2 + (2 - 2)^2 + (1 - 1)^2} = 2.04 \)
Correct prediction
- molecule 2 is more similar to 1 than to molecule 3 or 4 based on 4-D Euclidean distance measurement
  - they are both angiotensin-converting enzyme inhibitors
- these four descriptors successfully detect this case of structure-activity similarity
Descriptor-based Similarity: Example

- Incorrect prediction
  - $D_{34} > D_{14}$ → molecule 1 is calculated to be more similar to 4 than molecule 3 to 4
    - molecule 3 and 4 belong to the same activity class, phospholipase A2 inhibitor
  - Here structure-activity similarity is not correctly accounted for in 4-D space
Descriptor-based Similarity: Example

- Euclidean distance using only three descriptors: b_rotN, a_acc, a_don (except logP)
  - \( D_{12} = \sqrt{(2 - 2)^2 + (3 - 3)^2 + (2 - 3)^2} = 1 \)
  - \( D_{13} = \sqrt{(2 - 0)^2 + (3 - 2)^2 + (2 - 1)^2} = 2.45 \)
  - \( D_{14} = \sqrt{(2 - 1)^2 + (3 - 2)^2 + (2 - 1)^2} = 1.73 \)
  - \( D_{34} = \sqrt{(0 - 1)^2 + (2 - 2)^2 + (1 - 1)^2} = 1 \)

- Selection of descriptors: critical
Molecular Fingerprints

- Bit string representations of chemical structure and properties
- 2D and/or 3D features of a molecule are typically encoded as a vector of binary values

```
Å

molecular fingerprint

= “1”

= “0”
```
Fingerprint Design Strategies

■ Keyed fingerprints:
  - one bit position is associated with exactly one predefined feature
  - “1”: presence of the feature in the molecule
  - “0”: absence of the feature in the molecule

■ Hashed fingerprints:
  - one molecular feature is mapped to several bit positions applying a hash function
  - it follows that a bit position is often set to “1” by more than one feature
Keyed Fingerprint Designs

- Structural keys (fragment-based fingerprints):

- Pharmacophore-based fingerprints:
MACCS Structural Keys

- “Molecular ACCess System” of MDL, 1979
- Account for the presence/absence of structural features:
  - Are there fewer than 3 oxygens?
  - Is there a S-S bond?
  - Is there a ring of size 4?
  - Is at least one F, Cl, Br, or I present?
- 166 keys
Daylight – Hashed Fingerprint

- Monitors molecular connectivity pathways of varying length (usually 0-7 bonds)
- Each pathway is mapped on a pre-set bit string
- All bit strings are joined by a logical OR operation

![Molecular structure diagram]

Pathway: O=CNC
... 
CCN=CC=CCl

individual bit strings for single pathways

bit string for the molecule
Fingerprint Distance Metrics

- Quantifying similarity or distance (dissimilarity) between two molecular fingerprints.

- Common relationship between distance (dissimilarity) and similarity metrics:
  \[ \text{distance} = 1 - \text{similarity} \]

- Can be applied to continuous or binary variables.
Popular Similarity/Distance Coefficients

- Similarity metrics:
  - Tanimoto coefficient
  - Dice coefficient
  - Cosine coefficient

- Distance metrics:
  - Euclidean distance
  - Hamming distance
  - Soergel distance
Tanimoto Coefficient (Tc)

- **Definition:**
  \[ s(A, B) = Tc(A, B) = \frac{c}{a + b - c} \]

- value range: [0,1]
- Tc is also known as Jaccard coefficient
- Tc is the most popular similarity coefficient
Example Tc Calculation

\[ T_c(A, B) = \frac{2}{4 + 4 - 2} = \frac{2}{6} = \frac{1}{3} \]

\( a = 4, \ b = 4, \ c = 2 \)
Dice Coefficient

- Definition:
  \[ s(A,B) = \frac{2c}{a + b} \]
  - value range: [0,1]
  - monotonic with the Tanimoto coefficient
Cosine Coefficient

- **Definition:**
  \[ s(A,B) = \frac{c}{\sqrt{ab}} \]

- **Properties:**
  - value range: [0,1]
  - correlated with the Tanimoto coefficient but not strictly monotonic with it
Hamming Distance

- Definition:

\[ d(A, B) = a + b - 2c \]

- value range: [0,N]  (N, length of the fingerprint)
- also called Manhattan/City Block distance
Soergel Distance

- **Definition:**
  \[ d(A, B) = \frac{a + b - 2c}{a + b - c} \]

- **Properties:**
  - value range: [0,1]
  - equivalent to (1 – Tc) for binary fingerprints
Clustering

- Process of dividing molecules into classes based on similarity in chemical reference space
  - molecules in the same cluster are similar to each other
  - molecules in different clusters are thought to be different from each other

- Typical property-prediction calculations
  - active vs. inactive molecules
  - molecules active against different targets
  - must be separated in different clusters
Clustering

- Unsupervised classification approach
  - no predefined cluster composition

- Types of clustering
  - Hierarchical
    - create a hierarchical decomposition of the set of objects
  - Non-hierarchical
    - find $k$ partitions, minimizing some objective function
  - And more...
Hierarchical Clustering

- Hierarchical decomposition of the data set with respect to a given similarity measure into a set of nested clusters
- Results represented by a dendrogram
  - nodes in the dendrogram represent clusters
  - can be constructed bottom-up (agglomerative approach) or top-down (divisive approach)
Hierarchical Agglomerative

- Basic Lance-Williams algorithm (common to all hierarchical-agglomerative methods)

  - procedure:
    - starts with table of similarities between all pairs of items (molecules)
    - at each step, the most similar pair of molecules (or previously-formed clusters) are merged together
    - until all items are in a single maximal cluster

  - slow: overall time requirements are $O(N^3)$
    - $O(N^2)$ to generate pairwise similarity table initially
    - table must be updated $N$ times, once for each agglomeration of clusters; $N$, number of items
Hierarchical Agglomerative

- Methods often differ in how they determine the similarity between clusters or between a molecule and a cluster

  - distance between an object and a cluster of objects
    - 1-NN: identify the most similar object from the cluster: its distance to the molecule is the distance between this cluster and that molecule
    - $k$-NN: take the top $k$ most similar objects from the cluster: their average distance to the molecule is the distance between this cluster and that molecule

  - distance between two clusters
    - three alternatives: single link, complete link, average link

„NN“: nearest neighbor
**Distance Function for Clusters**

- Given
  - a distance function $\text{dist}(p,q)$ (e.g. Euclidean distance)
  - two clusters, $X$ and $Y$

- The distance between $X$ and $Y$ can be calculated as:
  - **“single link”**
    - distance between most similar members of two clusters
    $$\text{dist}_{sl}(X,Y) = \min_{x \in X, y \in Y} \text{dist}(x, y)$$
  - **“complete link”**
    - distance between most dissimilar members
    $$\text{dist}_{cl}(X,Y) = \max_{x \in X, y \in Y} \text{dist}(x, y)$$
  - **“average link”**
    - average distance between cluster members
    $$\text{dist}_{al}(X,Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} \text{dist}(x, y)$$

$|X|, |Y|$: size of $X$ and $Y$
Hierarchical Divisive

- Starting from the maximal cluster, clusters are iteratively divided until only singletons remain.
Non-hierarchical Clustering

- Constructs a division of a set $S$ of $n$ objects into a set of $k$ clusters minimizing a distance function (e.g. total distance of clusters)
- Cluster number $k$ is pre-defined
K-Means Clustering

- **Objective:**
  - Given $k$, form $k$ clusters so that the sum of the distances between the mean of the clusters (cluster centers) and their members is minimal.
**K-Means Clustering**

- **Four steps:**
  1. partition the objects into \( k \) non-empty subsets
  2. compute the cluster centroids (means) to represent the clusters
  3. re-assign each object to the cluster with the nearest centroid
  4. go back to step 2 and reiterate until the cluster memberships do no longer change
K-Means Clustering: Example

- $k = 2$
K-Means Clustering: Pros and Cons

- $K$-means is fast ($O(Nk)$)

- However, it has several disadvantages
  - sensitive to the initial choice of seeds
  - can converge to a local (rather than global) optimum
  - tends to produce “spherical” clusters of similar size
  - difficult to decide which $k$ value to choose
K-Nearest Neighbor Clustering

- Best known example: Jarvis-Patrick method
  - identify top $k$ (e.g. 20) nearest neighbours for each molecule
  - two molecules join the same cluster if they share at least $k_{\text{min}}$ of their top $k$ nearest neighbours

- Tends to produce a few large and heterogeneous clusters and many singletons (single-member clusters)
Fuzzy Clustering

- Produces overlapping clusters, i.e. molecules may belong to more than one cluster

- Each molecule has partial membership of all clusters
  - degree of membership in each cluster is in range $[0.0, 1.0]$
  - for one molecule, its sum of membership over all clusters is 1.0

- A better representation of similarity relationships, but decision making is often more difficult
Number of Clusters

- **Hierarchical methods**
  - allow user to choose any level across the hierarchy
- **K-means**
  - arbitrarily defined by user
- **K-nearest neighbours**
  - generated by algorithm
  - depends on input parameters
Choice of Clustering Methods

- There is no general solution
  - as with similarity measures and structural descriptors
  
  - For many chemoinformatics applications preferred: Ward's hierarchical-agglomerative clustering
Cell-based Partitioning Methods

- Similarity between molecules
  - they map to the same partition (cell)

- Selecting a diverse subset from a data set
  - take a representative from each cell
Cell Definition

- Given a pre-defined (low-dimensional) chemistry space
  - each descriptor (property) is recorded along a separate orthogonal axis and divided into a series of value ranges ("bins")
  - the combinatorial product of these bins for all the properties defines a set of cells that covers the space

\[ \prod_{i=1}^{N} b_i \]

\(\text{Number of cells} = \prod_{i=1}^{N} b_i\)
Partitioning

- Binning of molecular weight (MW) and logP
- Set is diverse (15 / 16 cells occupied)
- Outlier is shown in red
  - only one compound in cell
Advantages of Cell-based Methods

- Empty cells with low occupancy can be readily identified
  - indicating regions of space that are under-represented
- Diversity of different subsets can be easily compared
  - by examining the overlap in the cells occupied by each subset
- Fast, with $O(N)$ time complexity
Drawbacks of Cell-based Methods

- Often restricted to low-dimensional space
  - number of cells increases exponentially with the number of dimensions
  - selection of a small set of relevant descriptors is critical
  - dimension reduction using methods like PCA is often required
Dissimilarity-based Compound Selection

- Identification of a **diverse subset** that is representative of a database
- Basic algorithm for DBCS:
  1. select a compound and place it in the subset
  2. calculate **dissimilarity** between each compound in the subset and remaining compounds in the data set
  3. choose the next compound as the one being **most dissimilar** to the compounds in the subset
  4. iterate until desired subset size is reached
Selecting the Initial Compound

- Several ways to select the first compound
  - random
  - most representative
  - most dissimilar to all other molecules

- The most representative molecule has the largest sum of similarities to all other molecules

- The most dissimilar molecule has the smallest sum of similarities to all other molecules
What Does Most Dissimilar Mean?

- Two most commonly used methods to assess dissimilarity: **MaxSum** and **MaxMin**

- MaxSum selects a compound with *maximal distance sum* to all other compounds

- MaxMin chooses a compound with maximal distance to its *$k$ nearest neighbor*
Sphere Exclusion

- Related to DBCS, but based on exclusion of compounds

- Removing compounds that fall below dissimilarity threshold to chosen compounds

- Corresponds to an exclusion hypersphere calculated around each compound
  - subsequent compounds can be selected using MaxSum / MaxMin
DBCS vs. Sphere Exclusion

MaxSum

MaxMin

Sphere exclusion

diverse subset → representative subset

most dissimilar initial compound