Methods of Protein Structure Comparison

Introduction

Many of the algorithms discussed before require *comparing* two molecules (or molecular complexes). For example, when trying to fold a protein, how do we know whether our resulting protein is similar to the native structure. But how do we quantitatively measure the difference between two structures? We have to come up with a function that, when given two protein structures, tells us how similar or different they are from one another and this is not a trivial problem. Most intuitively, our function, $d(a, b)$ should have some or all of the properties of a distance function or a metric:

- The distance is non-negative: $d(a, b) \geq 0$ for all a, b.
- The distance of a molecule from itself, $d(a, a)$, is 0.
- The distance is symmetric. $d(a, b) = d(b, a)$.
- Triangular inequality: For every c - $d(a, c) + d(b, c) \leq d(a, c)$

To define such a function, we have to first decide on the right way to represent the molecule. As discussed earlier, there may be several ways to represent a molecule. First, we focus on representation as a set of atomic coordinates. In other words – a molecule with n atoms is represented by a set of $3 * n$ atomic coordinates, one for each of the location of the x, y, z coordinates of every atom. There are many ways to measure conformational similarity, and finding a good measure for conformational similarity is an active research area. What constitutes a good similarity measure also depends on the type of molecules we are interested in, as we will see later.

1 Root Mean Square Deviation (RMSD)

The most popular method for measuring the distance between two molecules is the RMSD. It is the average atomic distance between pairs of atoms from the two molecules. Formally, given two conformations A and B of a molecule of N atoms, represented as two $3 \times N$ matrices a and a. The i^{th} position of each matrix, a_i or a_i represents the (x, y, z) coordinates of atom i.

$RMSD(a, b)$ is defined as follows:

$$RMSD(X, Y) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |a_i - b_i|^2}$$

Where $|a_i - b_i|^2$ is the square Euclidean distance between points a_i and b_i, defined as:

$$|a_i - b_i|^2 = (a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2 + (a_{iz} - b_{iz})^2$$

RMSD is one of the simplest measures to quantify how different two protein conformations really are. It is intuitive and simple to compute by representing conformations as coordinate vectors. One problem is that the measure, as defined above, does not account for the absolute position and orientation of the two molecules. This way, even two identical conformations will turn up a very big RMSD if translated and/or rotated with respect to one another. However, this can be easily solved using least RMSD (lRMSD) as seen next.

Another extension of the RMSD measure is the weighted RMSD (wRMSD), which allows focusing on selected subsets of the atoms by assigning different weights to different atoms. It can be useful, for example, in downplaying the regions known to be inherently disordered:

$$wRMSD(X,Y) = \sqrt{\frac{\sum_{i=1}^{N} w_i |x_i - y_i|^2}{\sum_{i=1}^{N} w_i}}$$

1.1 Least RMSD – lRMSD

Least RMSD calculates RMSD after calculating the optimal alignment of two chains after removal of the changes due to rigid body transformations (translation and rotation) To remove translation, we simply align the centroids of the two conformations (centroid = average of all the coordinates). Let us define the centroids as follows: $c_a = centroid(a) = \frac{1}{N}\sum_{i=1}^{N} a_i$ and $c_b = centroid(b) = \frac{1}{N}\sum_{i=1}^{N} b_i$. Then we make both molecules centered at the same point. The easiest way is to "drag" both molecules so that their centers are at the origin. We do it by subtracting the centroid from every coordinate:

$$a'_i = a_i - c_a$$
$$b'_i = b_i - c_b$$

Notice that:

- The centroid of each molecule is now at $(0, 0, 0)$ (just calculate the geometric center and see!)
- While the position of the atoms changed, their relative positions with respect to one another remained the same, since we moved them all by the same magnitude. This is called a *rigid translation*, since we moved (translated) all the atoms as a rigid body.

Alternatively, "drag" molecule a to align with the centroid of b by setting each coordinate in a to:

$$a'_i = a_i - [c_a - c_b]$$

In any case, the centroids of the two molecules are now aligned.

Removing differences due to rotation are not as straight forward as removing translation, but still rather simple. Generally, we need to find the optimal transformation U that minimizes the distance E between y and the transformed x.

$$E = \frac{1}{N} \sum_{i=1}^{N} |Ux_i - y_i|^2$$

To solve this equation, we use some linear algebra for the eigenvector decomposition to find U:

$$NE = \sum_{i=1}^{N} (x_i^2 + y_i^2) - 2Tr(Y^T X')$$

So, after centering x and y to remove translations as described above, we do the following:

1. Store centered x and y as 3XN matrices (x,y,z) on rows for each of N points on columns)
2. Compute the transpose Y^T of matrix Y
3. Compute the covariance matrix $C = XY^T$
4. Apply SVD (Singular Value Decomposition) to the covariance matrix C
5. SVD yields matrices V, S, W^T such that $C = VSW^T$
6. Compute the determinant det(C) of the matrix C
7. compute the sign of this determinant: d=sign(det(C))
8. Finally, compute the optimal rotation U as

$$U = W \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} V^T$$

9. transform x by U and get a new vector U_x
10. lRMSD(x,y) = RMSD(Ux, y)

Despite being popular and simple to calculate, lRMSD has several major shortcomings:

1. Since we average the distance between pairs of atoms, we require a match list as input. In other words – every atom in x should have a matching atom in y. This limits us to conformations of the same protein chain. In other cases we need to define a match list and compute the RMSD only on this list.
2. Since it is an average of the euclidean distances, it tends to average out localized changes. Conversely, if a small perturbation occurs in a part of the structure, e.g. rotation of a hinge connecting two domains, lRMSD will report a large value (see Figure 1). The main reason is that lRMSD does not know how to attribute changes to specific atoms of the chain, since it distributes change equally (through the averaging) to all atoms in a protein chain

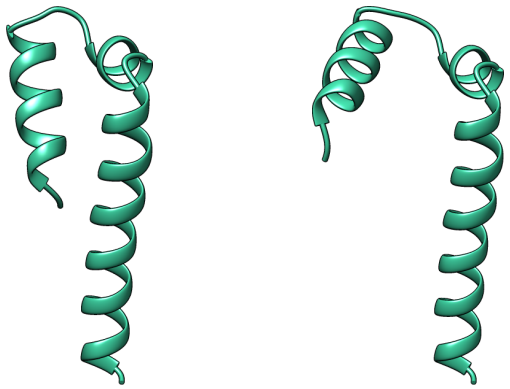


Figure 1: An example of two conformations of the same molecule. The only difference between them is one dihedral angle, which would result in shift of one helix with respect to the other, hence resulting in a large RMSD.

The root-mean-square distance measure accepts a list of pairs. It does not necessarily have to be atomic coordinates. We can also calculate the RMS of dihedral angles should we want to use an internal representation of the protein structure that includes bond lengths, planar angles, and dihedral angles. Usually, we describe the configuration of the backbone by the set of pairs of dihedral angle values, ϕ and ψ, which allows us to compare two structures without aligning them first. It should be noted that RMS of dihedral angles may give us vastly different results than atom-based RMSD. For example, modifying a small number of backbone dihedral angles can cause significant change to the structure, while having only marginal effect on the dihedral angle RMSD (as shown in the example above, Fig. 1). On the other hand, very similar structures are sometimes characterized by significant variations in their dihedral angles simply because these variations may partially cancel each other out. This can happen, for example, when there is a large-scale rotation of the peptide plane that takes the ϕ and ψ angles at residues i and i + 1 to different regions in the Ramachandran plot with a relatively small effect on the relative orientation of their side chains. This is called a peptide plane flipping. In particular, if the changes to $|\psi_i| + |\phi i + 1|$ are large but changes to $\psi_i + \phi_{i+1}$ are relatively small.

1.2 Local-Global Alignment

To overcome the main shortcomings of RMSD, CASP model evaluation uses two different tests: Global Distance Test (GDT) and Longest Continuous Segment (LCS). This measure conducts multiple superimpositions to find regions of similarity, switching between global alignment (GDT) and local alignment (LCS), striking the balance between the advantages of each superimposition. LCS finds the longest segments of residues that can fit under a selected RMSD cutoff. GDT is searching for the largest (not necessary continuous) set of "equivalent" residues that deviate by no more than a specified *distance* cutoff. The alignment between the first molecule and the second molecule is then evaluated using Local Global Alignment Scoring function (LGA_S). The steps above are repeated several times to find the complete set of local and global regions of 3D similarities between given two protein structures.

LCA: Both structures are scanned stepwise along their backbones and a moving search window is used to select segments for a comparison. The least RMSD method (see above) is then applied.

GDT: Given two molecules M1 and M2, for each selected pair of three, five and seven residue segments from both structures, calculate a superposition and an RMSD. Each calculated superposition is used as a starting point to an initial list of equivalent residues from both molecules. The list is examined and all the pairs of atoms whose distance is above a given cutoff are removed. The transformation is recalculated after removing these pairs. The above stage is repeated until there is no change in the number of atoms. The result is the largest set of (not necessarily contiguous) residues that can fit under a given distance cutoff.

The LGA Scoring Function: In the structure alignment search procedure, for each generated list of equivalent residues, the following values are calculated:

- LCS_vi is the percent of residues (continuous set) that can fit under an RMSD cutoff of v_i Å (for $v_i = 1.0, 2.0, \dots$)
- GDT_vi – an estimation of the percent of residues (largest set) that can fit under the distance cutoff of v_i Å (for $v_i = 0.5, 1.0, \dots$).

A scoring function (LGA_S) can be defined as a combination of these values and can be used to evaluate the level of structure similarity of selected regions. The combination can be done through a weight factor w which determines how much weight to give each of the two components.

GDT is dependent on the distance cutoffs which are chosen arbitrarily. Later methods replace it by a continuous distance dependent weight in the iterative weighted superimposition algorithm. The algorithm finds the better superimposable core between the two structures as follows:

1. The atomic equivalences are established between the two structures and a vector of per-atom weights $\{W_1, W_2, \dots, W_n\}$ is initialized to $\{1, 1 \dots, 1\}$.
2. A weighted superimposition is performed and the weighted RMSD is described above.
3. The deviations $\{d_1, d_2, \dots, d_n\}$ are calculated for all atom pairs, and their X-quantile, d_X is determined. X is an input parameter that defines the minimal size of the superimposable core to be found; by default it is set to 50%.
4. The new weights are calculated according to the formula. $W_i = \exp(-d_i^2/d_X^2)$. If two atoms are very closely superimposed, the weight is close to 1. Otherwise, it gets smaller as the deviation of pairs of atoms gets larger.
5. . Steps 2 – 4 are iterated until the weighted RMSD value stops improving or the specified maximum number of iterations is reached.

The similarity of the two structures can then be evaluated by the weighted RMSD or by taking the average of weights recalculated for the structure according to step 4 with d_X set to a fixed value, e.g., 2Å. The complement of this number, denoted superimposition error or E_{super}, ranges from 0 to 100% with lower values corresponding to more similar structure pairs:

$$E_{super} = 100\% \times (1 - \frac{1}{n}\sum_{i=1}^{n} \exp(-\frac{d_i^2}{d_X^2}))$$

This measurement is less sensitive to a small number of strongly deviating atoms due to the exponential scaling. Large discrepancies are captured and quantified.

1.3 TM-score (Template Modeling Score)

Another problem that one runs into when using RMSD to compare protein structures is that the RMSD distribution also depends on the size of the protein. This becomes important when the models of several different size proteins are evaluated in comparison with one another. TM-score aims to eliminate this dependence on the protein size. The score is a number in the range $(0, 1]$, where a higher score indicates a better match, and 1 is a perfect match. It is basically an extension of the approach using by the GDT score, as it uses a distance measure to assess similarity.

The score is defined as follows:

$$TM_{score} = \max\{\frac{1}{L_{target}} \sum_{i=1}^{L_{aligned}} \frac{1}{1 + (\frac{D_i}{D_0(L_{target})})}\}$$

Where L_{target} and $L_{aligned}$ are the lengths of the target protein and the aligned region respectively. D_i is the distance between the i^{th} pair of residues and $D_0(L_{target}) = 1.24\sqrt[3]{L_{target} - 15} - 1.8$ is a normalization factor derived from an analysis of a large scale of structures. The use of L_{target} eliminates the dependence of the score on the target size.

1.4 Comparing Protein Contacts

Contact-based measures rely on comparison of pairwise distances and/or interactions within one structure with the corresponding distances/interactions in the other structure rather than the distances between the corresponding points in the two structures. Therefore there is no need to superimpose the two structures. The Contact Area Difference (CAD) algorithm defines residue contact as the difference in accessible surface area when calculated for a pair of residues separately or together. While this contact area measure provides the most realistic assessment of fold similarity between the two structures, it is very sensitive to the way the side chains are packed. In other words, it requires specific residue pairs to be in contact with about the same area. If the side chains are not packed correctly even with roughly similar fold, the distance will be large.

A "contact" can be defined in several different ways. Given two residues whose C-α or C-β atoms are located at the distance of dÅ, the residue contact strength can be calculated as

$$f(d) = \begin{cases} 1 & \text{if } d < d_{min} \\ \frac{d_{max} - d}{d_{max} - d_{min}} & \text{if } d_{min} < d < d_{max} \\ 0 & \text{if } d > d_{max} \end{cases}$$

where d_{min} and d_{max} are predefined distance margin boundaries. The values of d_{min} and d_{max} can be chosen in such a way that the corresponding contact strengths are correlated with the pairwise residue contact areas which describe the real physical residue interactions. Cβ-Cβ contacts approximate contact areas more accurately than Cα -Cα , because on average, Cβ atoms are closer to their residues' center of mass. Later on, the Cβ atoms were replaced by virtual points, Cβ', located in the direction of the Cα-Cβ bonds at the distance of $1.5 \times dist(C\alpha - C\beta)$ from the C-α atom of each residue. The optimal margin boundaries were found to be $d_{min} = 4\mathring{A}$ and $d_{max} = 8\mathring{A}$.

A matrix of atomic contact strength is then defined for each one of the structures: C^R_{nn} for the first (reference) structure and C^M_{nn} for the second (model) structure. Each entry $[i, j]$ represents the contact strength between residues i and j. A contact similarity matrix $C^{R\cap M}$ is constructed using $C^{R\cap M}_{i,j} = Min(C^R[i, j], C^M[i, j])$ with a weight as $|C^{R\cap M}| = \sum_{i,j} C^{R\cap M}[i, j]$. This weight can be compared to either the weight of the contact matrices, $|C^R|$ or $|C^M|$, or the union of the two $|C^{R\cup M}|$, defined by $C^{R\cup M} = Max(C^R[i, j], C^M[i, j])$ (or their average). The three approaches result in quantities ranging from 0 to 100% and reflecting recall, precision, and accuracy with which the model reproduces the reference structure contacts.

For most pairs of experimentally determined structures of the same protein, protein flexibility and experimental errors lead to the contact strength differences of 5-20%. Small flexible fragments or even large domain movements have only minor effect on the contact strength matrices making the contact strength measures robust to elastic large scale deformations. At the same time, these measures are sensitive to major changes in packing occurring as a result of modeling errors.

An interesting insight is that for experimental structures, pairs may often differ in conformation (as reflected by superimposition error) or in contacts (as reflected by contact strength difference), but rarely in both. In contrast, computational models differ from their respective answers by both parameters simultaneously, especially in the more difficult modeling cases. This observation stressed the importance of applying complementary structure similarity measures that combine distance-based and contact-based approaches.

2 Other Quality Assessment: Shape Similarity

Sometimes assessment of cavities on the surface of a protein is more important than the description of the rest of the structure, especially when the goal is prediction of a binding site rather than of the entire structure (which can be thought of as a scaffold). Methods that assess surface area, solvent accessible surface area, that compute volumes, and detect cavities on proteins are very important in the context of protein binding and docking. To generate a model of the protein surface, we model each atom as a vdW sphere, the union of which gives the molecular surface

Not all molecular surface is accessible to solvent. Rolling a solvent ball over the VdW spheres traces out the solvent accessible surface area (SASA) . SASA is important to quantitatively determine interactions of the protein. Figure 3 shows two surfaces generated for the same molecule. One for a 1.4Å ball and one for a 2.4Å ball. Increasing the radius reduces the SASA due to more cavities that a bulkier ball cannot penetrate

Computational geometry methods that use Delaunay triangulations and alpha shapes assess SASA and other geometric descriptors of molecular surfaces, volumes, and cavities. We will come back to this topic in the context of molecular docking in Chapter **??**.

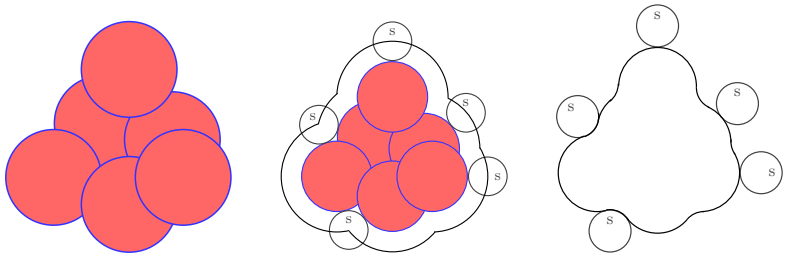


Figure 2: Stages in calculating molecular surface: (a) A molecule is represented by a set of vdW spheres. (b) A probe ball is rolled on the surface of the molecule, recording the accessible surface area. (c) The resulting outline indicates the solvent accessible surface.

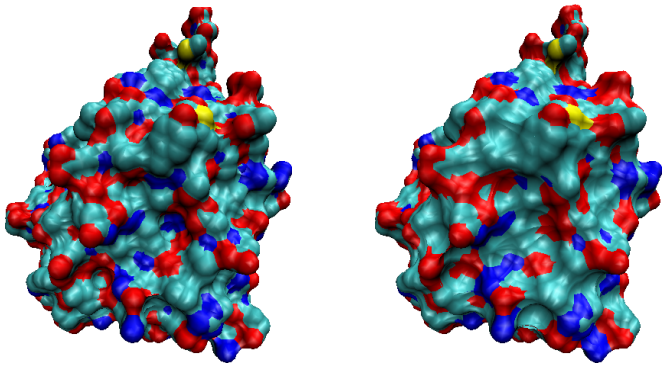


Figure 3: Two molecular surfaces made with a probe radius of 1.4Å (a) and 2.4Å (b)

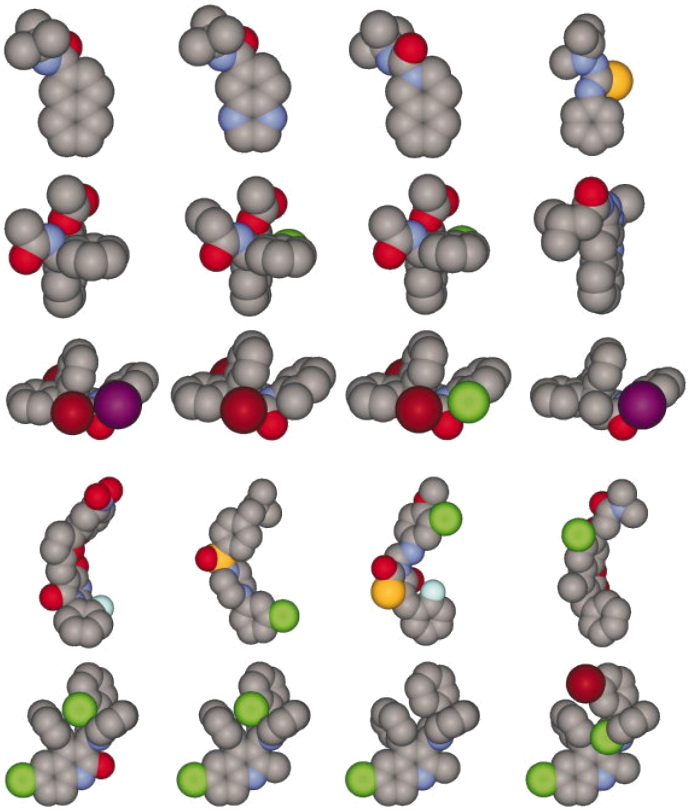


Figure 4: Examples of small compounds that can be used as drugs.

2.1 Ultrafast Shape Recognition (USR)

This method is an efficient global comparison of molecular shapes. It is suitable for comparing small molecules. Since it is very fast it can be used for large database search and does not require the molecules to be aligned. Drug design requires screening a number of potential compounds. The goal is to find a set of molecules which closely resemble a lead molecule from a HUGE database (millions of possible compounds).

Shape similarity may indicate similar binding properties and similar activity. The main idea is that the shape of a molecule is uniquely determined by the relative positions of the atoms, which are determined by the inter-atomic distances. The set of inter-atomic distances are constrained due to forces that hold the atoms together, so gathering statistics about them help us gain information about the shape of the molecule. Comparing statistics between different molecules help us assess how similar they are.

Given a molecule, the center of every atom is represented as a coordinate. The shape is described as 4 sets of atomic distance distributions from the following points:

- Center of mass – ctd
- Point closest to ctd – cst

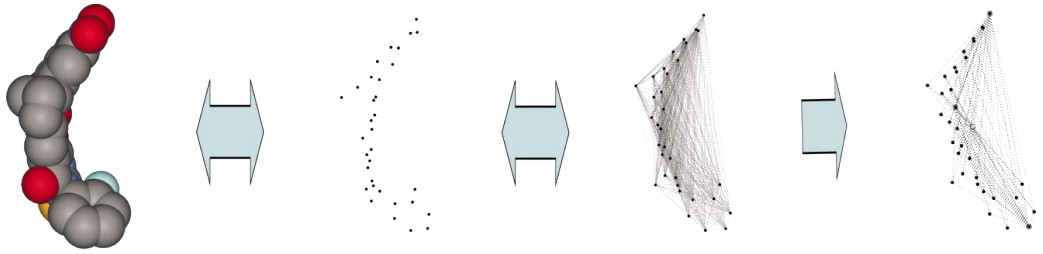


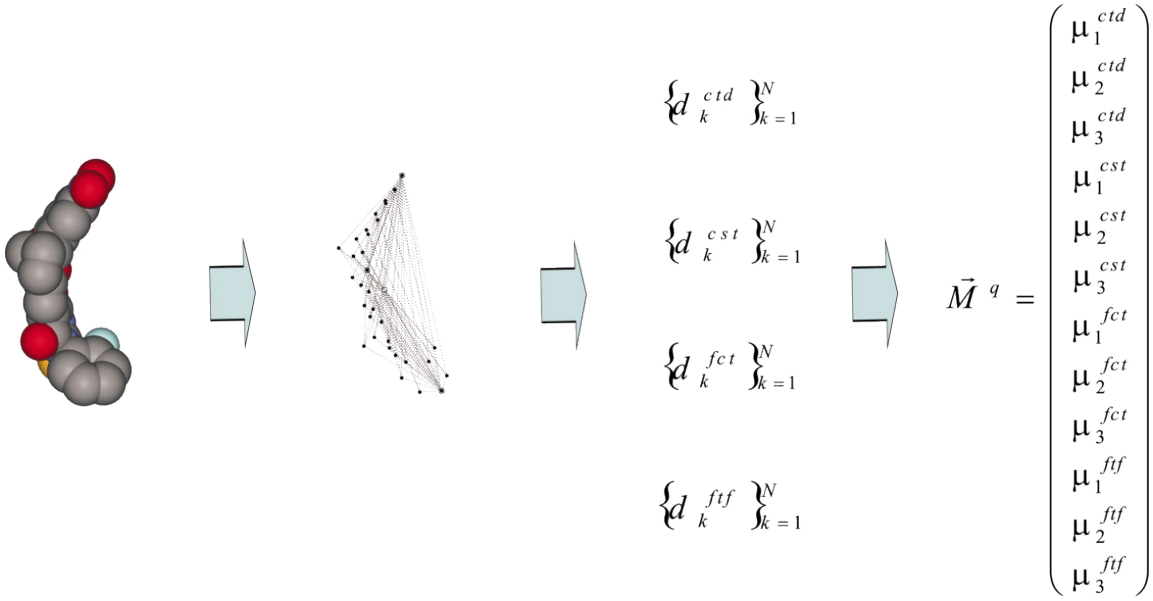
Figure 5: Using a set of points for calculating moments

- Point farthest from ctd – fct
- Point farthest from fct – ftf

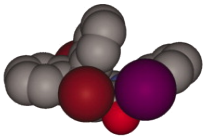
As seen on the right, the feature points are marked as circles. Notice that the center of mass is usually located outside the molecule.

The moments of the distributions for each one of the distances are calculated and stored as a feature vector. They estimate of the size, compactness and symmetry of the molecule. The distance between two molecules i and j is calculated as the Manhattan distance between their feature vectors M_i and M_j: $S_{lj} = \frac{1}{1+\frac{1}{12}\sum_{i=1}^{12}|\vec{M}_i^j - \vec{M}_i^l|} \in (0,1]$

The feature vector for the molecule is defined as follows:

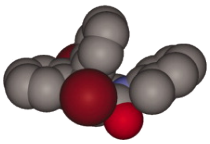


Where $\mu_1 =$ average, $\mu_2 =$ standard deviation $= \frac{1}{n}\sum_{k=1}^{n}(d_k^{cdt} - \mu_1^{ctd})^2$, $\mu_3 =$ skewness $= \frac{1}{n}\sum_{k=1}^{n}(d_k^{cdt} - \mu_1^{ctd})^3$ Here are two examples. In the first one, the similarity score is high, 0.966, and the molecules indeed look very similar.



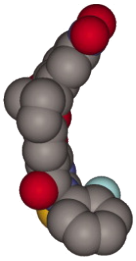
$$\vec{M}^q = (3.22,\ 1.38,\ 0.50,\ 3.46,\ 1.95,\ -0.47,\ 5.83,\ 6.80,\ -4.13,\ 5.71,\ 7.08,\ -1.70)$$

$$S_{qi} = \frac{1}{1 + \frac{1}{12}\sum_{l=1}^{12}\left|M_l^q - M_l^i\right|} \in (0,1] \qquad S_{qi} = 0.966$$



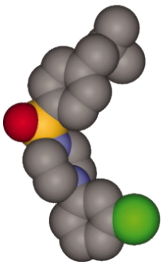
$$\vec{M}^i = (3.21,\ 1.36,\ 0.50,\ 3.45,\ 1.92,\ -0.51,\ 5.80,\ 6.83,\ -3.91,\ 5.71,\ 7.06,\ -1.73)$$

In the second example the score is 0.812 and even though the general shapes of the two molecules are similar, the differences are more noticeable.



$$\vec{M}^q = (4.44,\ 2.98,\ 1.04,\ 4.55,\ 4.70,\ 0.23,\ 8.30,\ 16.69,\ -22.97,\ 7.37,\ 15.64,\ 0.51)$$

$$S_{qi} = \frac{1}{1 + \frac{1}{12}\sum_{l=1}^{12}\left|M_l^q - M_l^i\right|} \in (0,1] \qquad S_{qi} = 0.812$$



$$\vec{M}^i = (4.39,\ 3.11,\ 1.36,\ 4.50,\ 4.44,\ 0.09,\ 8.34,\ 16.78,\ -23.20,\ 7.15,\ 16.52,\ 0.13)$$

Advantages and Disadvantages This is an extremely fast method due to calculation of only 4N distances and distributions. Also, it does not require the molecules to be aligned. On the other hand, it is very sensitive to small changes in the molecule shape, and does not directly account for chemical interactions and atom types. Because it is so sensitive to small changes in the shape, it works better for smaller molecules.

3 Further Reading

- For more information about shape computing see `http://cnx.org/content/m11616/latest/`