CS612 - Algorithms in Bioinformatics

Distance Measurement

March 18, 2023

Quality Assessment – Definition

- How do we measure the quality of a structural model?
- We have to define a similarity (or distance) measure(s) to assess how different two conformations are.
- Usages:
  - Assessing the success of a folding algorithm.
  - Measure structural similarity between two different proteins which may be related
  - Measure the similarity (or complementarity) of the surfaces two potentially interacting molecules.
  - ...
- No one-size-fits-all quick fix ...

Quality Assessment – RMSD

- RMSD: Root Mean Squared Deviation
- The most popular distance measure between two conformations
- Average pairwise atomic distance
- given two conformations of a chain of N atoms, represent the conformations as two $3 \times N$ vectors a and b
- RMSD(a,b) is defined as:

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

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

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

Least RMSD – IRMSD

- Optimal alignment of two chains after removal changes due to rigid body transformations (rotations and translations)
- Removing translation:
  - Calculate the centroids (geometric centers) of both molecules:
    $c_a = centroid(a) = \frac{1}{N}\sum_{i=1}^{N} a_i$ and $c_b = centroid(b) = \frac{1}{N}\sum_{i=1}^{N} b_i$.
  - Then "drag" a to centroid of b through $a - [c_a - c_b]$
  - Check that the new c_a is now at c_b
  - Alternatively, and probably more easily, you can drag both a and b to have $(0, 0, 0)$ as centroid by subtracting the value of each centroid from every atom:

$$a_i' = a_i - c_a$$
$$b_i' = b_i - c_b$$

IRMSD – Removing Rotations

- Not as easy as removing translation
- Generally, we need to find optimal transformation U that minimizes the distance E between b and the transformed a

$$E = \frac{1}{N} \sum_{i=1}^{N} |Ua_i - b_i|^2$$

- Finding the optimal transformation U:
- After some linear algebra:
- Some more linear algebra uses eigenvector decomposition to find U:

$$NE = \sum_{i=1}^{N} (a_i^2 + b_i^2) - 2Tr(B^T A')$$

IRMSD – Removing Rotations

- Center the atoms of a and b (separately) by subtracting the centroid from each point
- Store centered a and b as $3 \times N$ matrices (x, y, z) on rows for each of N points on columns)
- Compute the transpose B^T of matrix B
- Compute the covariance matrix $C = AB^T$
- Apply SVD (Singular Value Decomposition) to the covariance matrix C
- SVD yields matrices V, S, W^T such that $C = VSW^T$
- Compute the determinant det(C) of the matrix C
- compute the sign of this determinant: d=sign(det(C))

IRMSD – Removing Rotations

- Final step: compute the optimal rotation U as

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

- Finding IRMSD after Computing U:
- transform a by U and get a new vector U_a
- IRMSD(a,b) = RMSD(Ua, b)

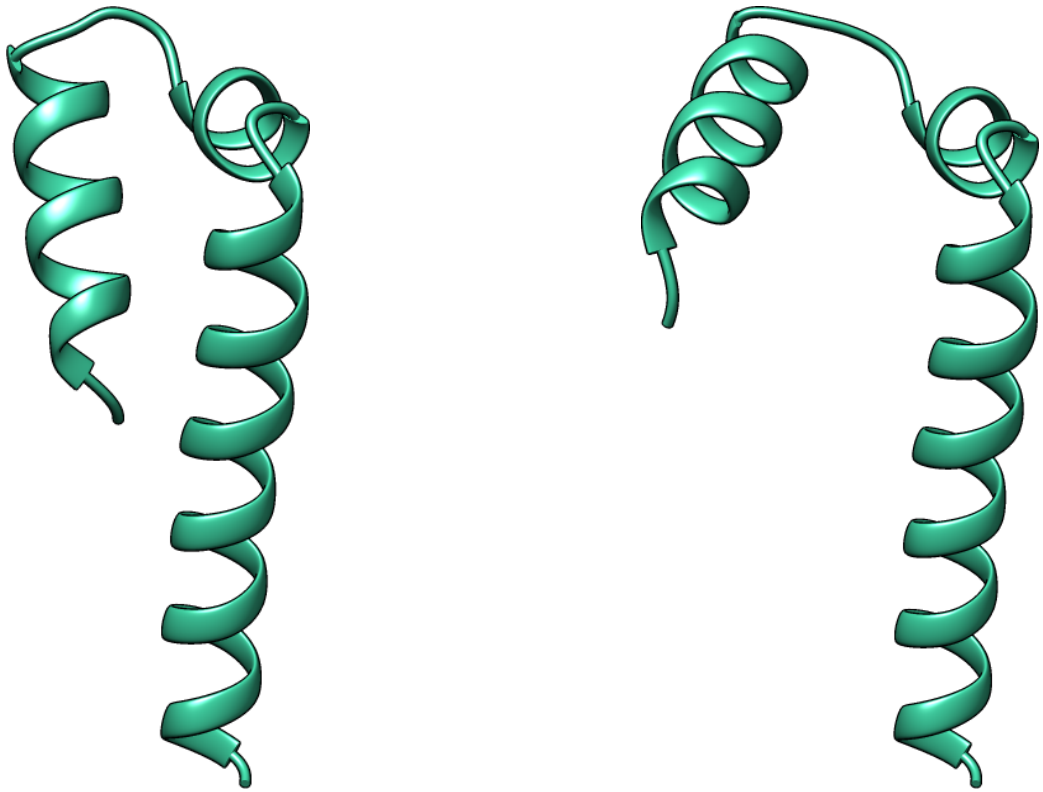
Quality Assessment – RMSD

- **Advantages:**
  1. Simple to compute by representing conformations as 3N vectors (N atoms)
  2. One of the simplest, most intuitive measures to quantify how different two protein conformations really are
- **Limitations:**
  1. Limited to conformations of the same protein chain
  2. Otherwise – atom-atom correspondence needed on different-length chains
  3. As an averaged measure, it is not very descriptive if changes are localized

IRMSD – Shortcomings

- IRMSD cannot capture localized changes: if a small perturbation occurs in a part of the structure, e.g. rotation of a hinge connecting two domains, IRMSD will report a large value
- Main reason: IRMSD does not know how to attribute changes to specific atoms of the chain
- IRMSD distributes change equally (through the averaging) to all atoms in a protein chain
- wRMSD can assign different weights to different atoms.
- Finding a good measure for conformational similarity is an active research area

IRMSD – Shortcomings



Dihedral RMSD

- We can also calculate the RMS of dihedral angles should we want to use an internal representation of the protein structure.
- It 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.
- On the other hand, very similar structures are sometimes characterized by significant variations in their dihedral angles because these variations may partially cancel out.

Quality Assessment through LGA

- Local-Global-Alignment (LGA) introduced by Adam Zemla in 2003 is being used as a more accurate similarity assessment than IRMSD in CASP
- LGA generates many different local superpositions to find regions where two conformations are similar: combines longest continuous segment (LCS) and global distance test (GDT) to find local and global similarities
- LCS superimposes the longest segments that fit under a selected RMSD cutoff
- GDT searches for the largest (not necessary continuous) set of 'equivalent' residues that deviate by no more than a specified distance cutoff

Zemla A., "LGA – a Method for Finding 3D Similarities in Protein Structures", Nucleic Acids Research, 2003, Vol. 31, No. 13, pp. 3370-3374
http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12824330

TM-Score (Template Modeling)

- RMSD distribution depends on the size of the protein.
- Sometimes we want to compare proteins of different sizes.
- The score is a number in the range $(0, 1]$, 1 being a perfect match.
- The score is defined as follows:

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

- 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 $D_0(L_{target})$ is a normalization factor.

Comparing Protein Contacts

- Contact-based measures rely on comparison of pairwise distances within one structure with the corresponding distances/interactions in the other structure.
- Therefore there is no need to superimpose the two structures.
- 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

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

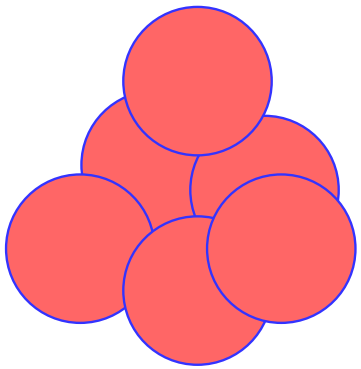
- The optimal margin boundaries were found to be $d_{min} = 4\mathring{A}$ and $d_{max} = 8\mathring{A}$

Comparing Protein Contacts

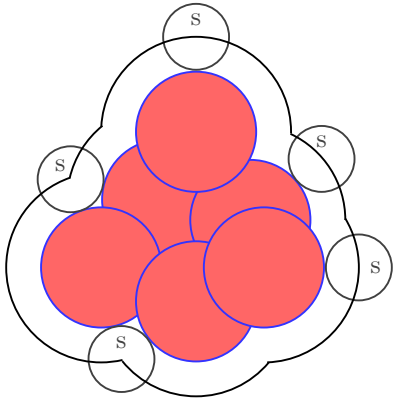
- For each protein define a matrix: C^R_{nn} for the first (reference) structure and C^M_{nn} for the second (model) structure.
- Each entry $[i, j]$ represents the contact strength between residues i and j.
- A contact similarity matrix $C^{R \cap M}$ is defined as $C^{R \cap M}_{i,j} = Min(C^R[i, j], C^M[i, j])$ 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, their union $|C^{R \cup M}|$, 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%

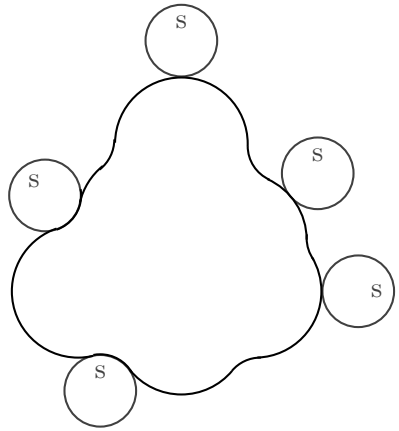
Other Quality Assessment: Shape Similarity

- Sometimes assessment of cavities on the surface of a protein is more important than 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 binding and docking



Model each atom as a vdw sphere, the union of which gives the molecular surface

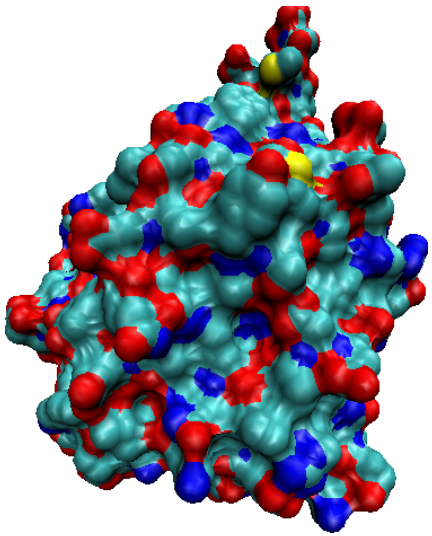




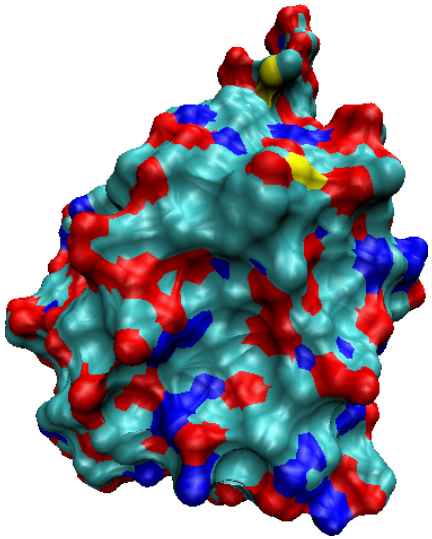
Not all molecular surface is accessible to solvent. Rolling a solvent ball over the vdw spheres traces out the solvent accessible surface area (SASA) . SASA is important to quantitatively determine interactions of the protein

Solvent Accessible Surface Area – SASA

- 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 – further reading about shape computing at `http://cnx.org/content/m11616/latest/`



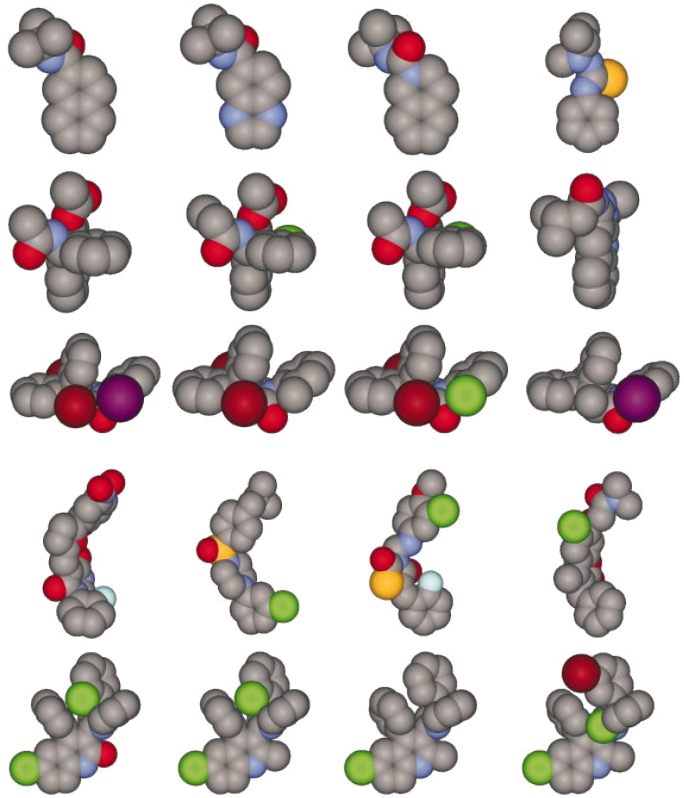
Ball radius 1.4Å



SASA for a 2.4Å ball. Increasing the radius reduces the SASA due to more cavities that a bulkier ball cannot penetrate

Ultrafast Shape Recognition (USR)

- Drug design – Screening a number of potential compounds.
- Find a set of molecules which closely resemble a lead molecule from a HUGE database.
- Shape similarity may indicate similar binding properties and similar activity.

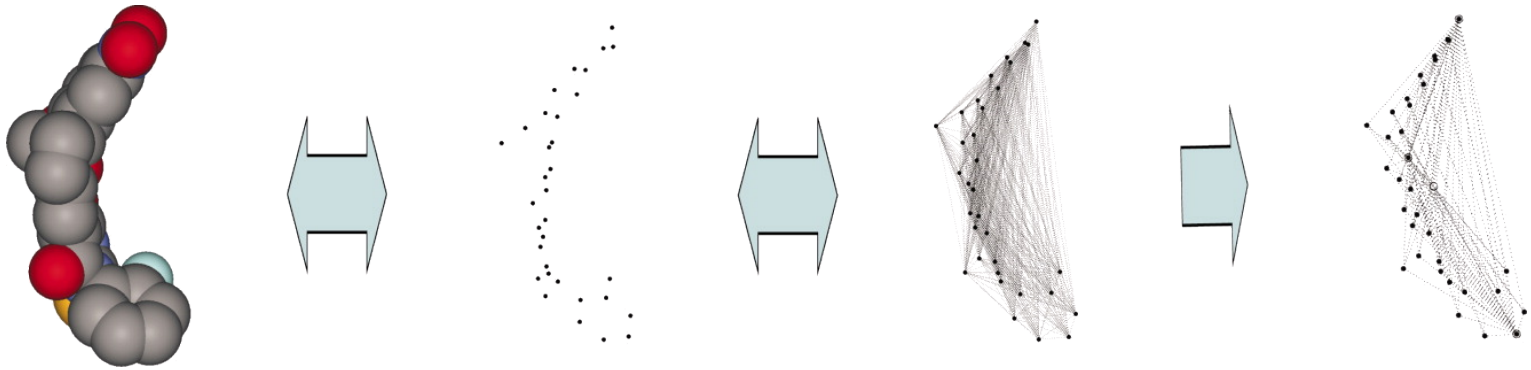


Overview

- Efficient global comparison of molecular shapes.
- The molecules are represented as feature vectors, representing the relative positions of the atoms.
- Does not require alignment of the molecules.
- Suitable for large database search.

Feature Vector Representation

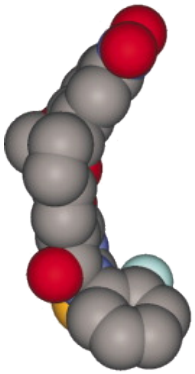
- The shape of a molecule is uniquely determined by the relative positions of the atoms.
- ... Which are determined by the inter-atomic distances.
- The set of distances can be constrained due to forces that hold the atoms together.



Feature Vector Representation

- The molecule is described as 4 sets of atomic distance distributions from feature points:
- Center of mass – ctd
- Point closest to ctd – cst
- Point farthest from ctd – fct
- Point farthest from fct – ftf
- The moments of the distributions are calculated and stored as a feature vector.
- Estimate of the size, compactness and symmetry of the molecule.
- Manhattan distance – $S_{lj} = \frac{1}{1+\frac{1}{12}\sum_{i=1}^{12}|\vec{M}_i^j - \vec{M}_i^l|} \in (0,1]$

Feature Vector for a Molecule



$$\{d_k^{ctd}\}_{k=1}^{N}$$

$$\{d_k^{cst}\}_{k=1}^{N}$$

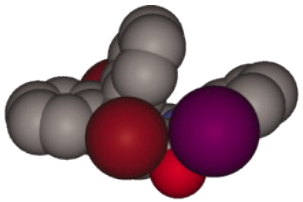
$$\{d_k^{fct}\}_{k=1}^{N}$$

$$\{d_k^{ftf}\}_{k=1}^{N}$$

$$\vec{M}^{q} = \begin{pmatrix} \mu_1^{ctd} \\ \mu_2^{ctd} \\ \mu_3^{ctd} \\ \mu_1^{cst} \\ \mu_2^{cst} \\ \mu_3^{cst} \\ \mu_1^{fct} \\ \mu_2^{fct} \\ \mu_3^{fct} \\ \mu_1^{ftf} \\ \mu_2^{ftf} \\ \mu_3^{ftf} \end{pmatrix}$$

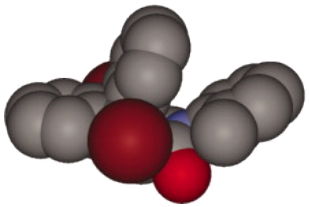
Where μ_1 = average, μ_2 = standard deviation = $\frac{1}{n}\sum_{k=1}^{n}(d_k^{cdt} - \mu_1^{ctd})^2$, μ_3 = skewness = $\frac{1}{n}\sum_{k=1}^{n}(d_k^{cdt} - \mu_1^{ctd})^3$

Similarity Score



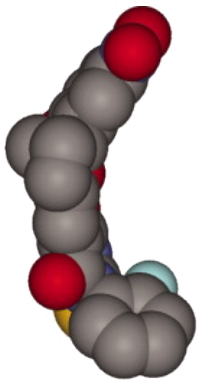
$$\vec{M}^q = (3.22,\ 1.38,\ 0.50,\ 3.46,\ 1.95,\ -0.47,\ 5.83,\ 6.80,\ -4.13,\ 5.71,\ 7.08,\ -1.70)$$

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



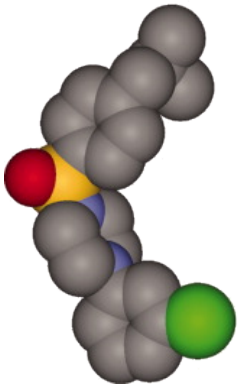
$$\vec{M}^i = (3.21,\ 1.36,\ 0.50,\ 3.45,\ 1.92,\ -0.51,\ 5.80,\ 6.83,\ -3.91,\ 5.71,\ 7.06,\ -1.73)$$

Similarity Score



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

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



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

Advantages and Disadvantages

- Extremely fast due to calculation of only 4N distances and distributions.
- Very sensitive to small changes in the molecule shape.
- Does not directly account for chemical interactions and atom types.