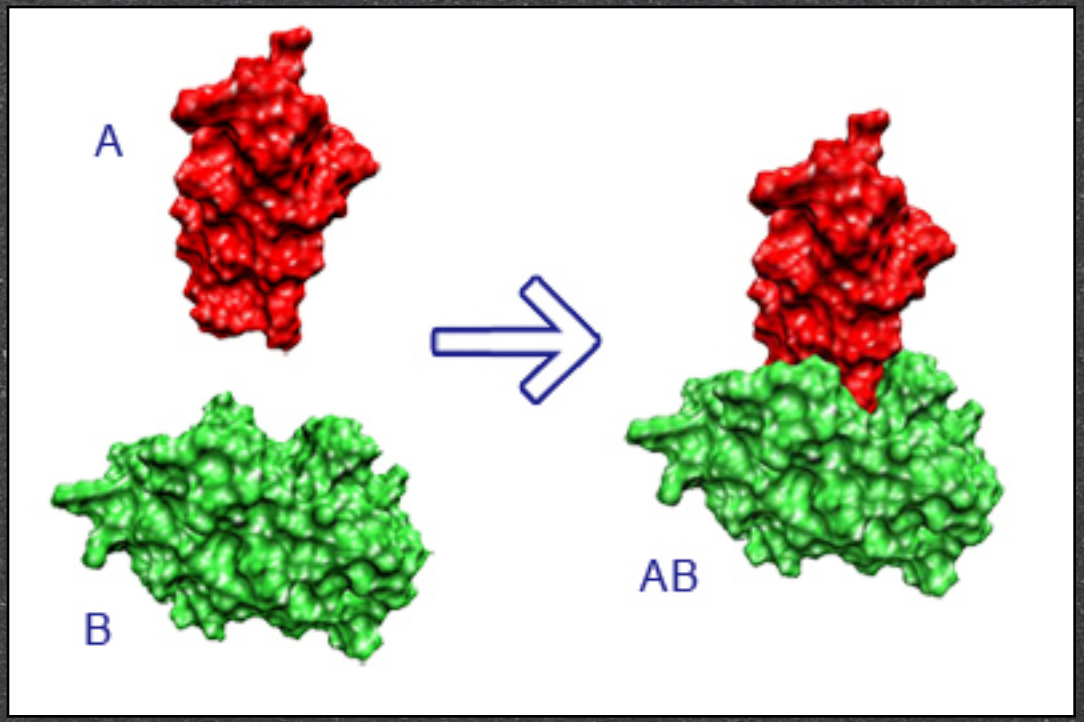
Protein Docking

Current trends in protein docking

Protein Docking Overview



Current Docking Methods

Current Docking Methods

- Molecular Dynamics
- Monte Carlo Methods
- Genetic Algorithm and Evolutionary Programming
- Fragment-based Methods
- Point Complementarity Methods
- Distance Geometry Methods
- Tabu Searches
- Systematic Searches
- Multiple Method Algorithms

Current Docking Methods

- Molecular Dynamics (MD)
  - Involves calculating the solutions to Newton's equations of motions.
  - Can be difficult. Since the hypersurface of a biological system is very rugged, the MD can become locked into a local minimum.
  - AMBER, CHARMM

Current Docking Methods

- Monte Carlo (MC) Methods
  - One of the most established and widely used stochastic optimization techniques.
  - One advantage over gradient based methods (such as MD) is that the algorithm can more easily jump over energy barriers in the biological systems hypersurface, making MC methods less susceptible to getting caught in local optima.
  - Involves applying random Cartesian moves to the system and accepting or rejecting the move based on a Boltzman probability.
  - Prodock, MCDOCK

Current Docking Methods

- Genetic Algorithm (GA) and Evolutionary Programming (EP)
  - Based on bio-operators found in Darwinian evolution (selection, recombination, and mutation)
  - Unlike MC and MD methods where a single starting structure is required, the GA method must generate a population of random solutions at its inception.
  - AutoDock, GOLD, DARWIN

Current Docking Methods

- Fragment-based Methods
  - The basic concept of Fragment-based docking methods can be described as dividing the ligand into separate portions or fragments, docking the fragments, and finally linking the fragments.
  - One major downfall of these methods is that the choice of the base fragment can be crucial to the quality of the results.
  - FlexX

Current Docking Methods

- Point Complementarity Methods
  - This method uses points of complimentarity between a ligand and a receptor’s binding site.
  - The major difference between this method and the fragment-based methods is its treatment of the ligand as a complete entity throughout the entire docking method.
  - Soft docking, FTDOCK, SANDOCK

Current Docking Methods

- Distance Geometry Methods
  - These methods determine the binding modes between protein and ligand through consideration of hydrogen bonding only.
  - This method samples the conformational space identifying plausible binding modes which are then used to direct an embedding algorithm.
  - DockIt

Current Docking Methods

- Tabu Searches
  - Tabu search is a local search method, that is used to search for solutions within a neighborhood, all the while modifying the neighborhood in which it is searching.
  - It gets its name from the data structure it uses to enhance its performance—a tabu list is made up of n solutions that have been visited in the past and are thus excluded from future neighborhoods. This feature helps keep the tabu search from getting stuck in a local optima.
  - PRO_LEADS

Current Docking Methods

- Systematic Searches
  - These methods systematically search the entire solution space.
  - Not the most efficient.
  - EUDOC, SYSDOC

Current Docking Methods

- Multiple Method Algorithms
  - Protein docking is still in its theoretical stages and thus far from solved.
  - The combination of several types of docking methods has been used to increase the effectiveness of a docking protocol.
  - In most combined methodologies, a computationally inexpensive method is first used, followed by a time consuming yet more accurate method to refine the population of solutions into a final docking solution.

PRO_LEADS (tabu search)

PRO_LEADS

- Begins with an initial random ligand conformation (referred to as the current solution).
- Random moves are then applied to the ligand (*current solution*) to create a population of solutions (typically 100 solutions).
- These solutions are then scored and ranked in ascending order where the highest ranking solution is then accepted as the new *current solution*.
- A new random population is then generated from the *current solution*, and the entire process is repeated for a user defined number of iterations.

PRO_LEADS

- To ensure diversity of solutions (i.e., attempt to avoid getting stuck on a local optima) a tabu list is used containing the last 25 *current solutions*.
- If a new best solution is the best so far, it is always accepted as the *current solution*. However, if it is not the best so far, the best non-tabu solution is chosen as the *current solution*.

PRO_LEADS

- With a validation test set of 50 ligand-protein complexes, PRO_LEADS achieved a success rate of 86%.

Protein Docking Using a Genetic Algorithm

Protein Docking Using a Genetic Algorithm

- Uses Connolly “critical” points.
- The idea of Connolly’s “critical” points is to preserve the important features (local maxima and minima) of the protein surface, while reducing the computational intractability associated with processing very large numbers of surface points.
- Connolly developed a method to extract these “critical” points (termed “knobs” and “holes”) from the Connolly surface of a protein.
- The “critical” points of the ligand and receptor are expected to match in a successful docking.

Protein Docking Using a Genetic Algorithm

- In this method, a GA was used to generate rotations of the smaller (query) protein relative to the larger (target) protein (which was held static).
- Both proteins were treated as rigid bodies in this method.

Protein Docking Using a Genetic Algorithm

- The first step in this method was to generate the proteins' Connolly dot surfaces.
- Then an initial population of chromosomes was generated at random (A chromosome consisted of 6 integer values representing the 6 degrees of freedom needed to define the movement of one rigid body relative to another).
- Each of the chromosomes were applied to the query protein's dots (translating and rotating the query proteins dot surface as defined by each chromosome) and the fitness of the resulting complex was calculated.

Protein Docking Using a Genetic Algorithm

- Fitness = number of matches – penalty
- Penalty = J * number of clashes, if any dots matched and 100, 000 otherwise.
- J = the penalty multiplier.
- Clashes = overlapping of a query protein's dot with an interior point of the target.

Protein Docking Using a Genetic Algorithm

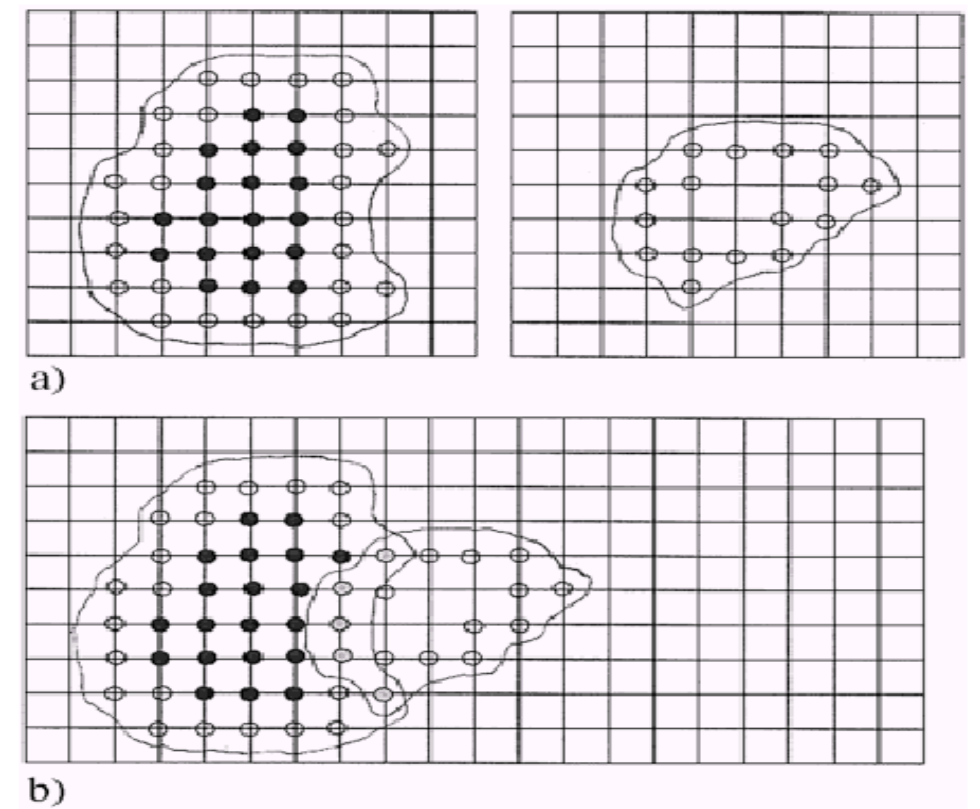


Fig. 1. **a:** Target and query proteins have their surfaces represented by grid point dots (○). The target also has interior points (●). **b:** Surface grid dots are matched (gray filled) to produce a putative complex. (N.B.: The proteins are positioned too closely together.)

References

- Taylor RD, Jewsbury PJ, Essex JW. *A review of protein-small molecule docking methods.* J Comput-Aided Mol Des 2002;16:151–166.
- Eleanor J.Gardiner and Peter Wilett, *Protein Docking Using a Genetic Algorithm.* Proteins: Structure, Function, and Genetics 44:44-56, 2001.
- Carol A. Baxter, Christopher W. Murray, David E. Clark, David R. Westhead, Matthew D. Eldridge *Flexible docking using tabu search and an empirical estimate of binding affinity.* Proteins: Structure, Function, and Genetics Volume 33, Issue 3, 1998. Pages 367-382

Molecular surface recognition: Determination of Geometric fit by correlation technique

Ephraim Katchalski-Katzir,
I. Sharvi,
M. Eisenstein

Introduction

- Shape complimentary matching is a necessary condition for protein-protein interaction and the other being the energy minimization.
- Geometric-based approach have advantage over other approaches for determine the shape complimentary.
- We present a geometric-based approach (FFT) which borrow the idea from pattern recognition matching.

Components

- Digital representation of protein
  - 3D atom coordinates (protein structure)
  - Protein a (probe) protein b (target)
  - Projected into 3D grid
- Calculation of correlation function
  - Using fast Fourier transformation
  - Discrete step in three dimension
- A scan of relative orientation
  - If surface-surface overlap, correlation is one
  - if there is no contact, correlation is zero

Method

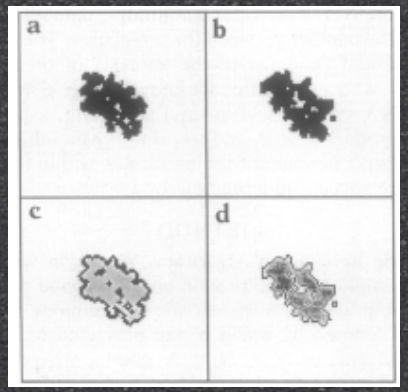
- a, b denotes the two molecule which are
- Projected into 3D grid of N*N*N points where l,m,n are indices.
- Any point in the grid is consider inside the molecule if there exist at least one atom with distance r (r =radii for van der waal surface)
- The resulting function are:

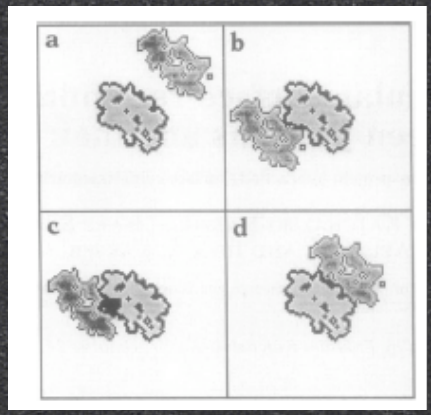
$$a_{l,m,n} = \begin{cases} 1 \text{ surface} \\ \rho \text{ inside} \\ 0 \text{ outside} \end{cases} \qquad b_{l,m,n} = \begin{cases} 1 \text{ surface} \\ \sigma \text{ inside} \\ 0 \text{ outside} \end{cases} \qquad [2]$$

ρ , σ points inside molecule

Method -cont

- The matching of surface is accomplished by correlation function defined as:
  - $-c_{\alpha,\beta,\lambda} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} \cdot b_{l+\alpha,m+\beta,n+\lambda}$ [3]
  - α, β,λ number of grid discrete step molecule b is moved with respect molecule a in each dimension.
  - Complexity $O(N^6)$





Method-cont

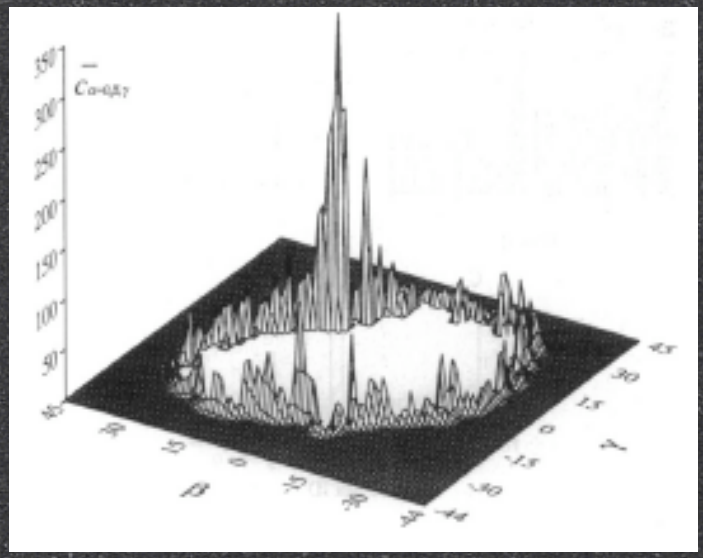
- If the shift vector {α,β,λ} is such that there is no contact the correlation is zero.
- If there is a contact between the surface the correlation contribution is positive
- A non-zero value is also obtained when one molecule penetrate the other. Thus a distinction is define:
  - σ is assigned a large negative value and ρ assigned a small positive value.
  - If one molecule penetrate the other than a it contribute a negative value to the correlation score

Fourier Transformation of correlation function

- A direct correlation between two function in the form described [3], expensive thus we take advantage of FFT.
- A discrete Fourier transformation of function $x_{l,m,n}$ is defined as:
  - $X_{o,p,q} = {}^{N}\sum_{l=1}{}^{N}\sum_{m=1}{}^{N}\sum_{n=1} \exp[-2\pi i(ol+pm+qn)/N] \cdot x_{l,m,n}$ [4]
  - Where $o,p,q = \{1,\ldots,N\}$ and $i=\sqrt{-1}$.
- we apply this transformation to equation [3], yield:
  - $C_{o,p,q} = A^*_{o,p,q} \cdot B_{o,p,q}$ [5]
  - Where C and B are DFT of c and b and A* is the complex conjugate of DFT of a.
- The inverse Fourier Transformation is defined as:
  - $\_c_{\alpha,\beta,\lambda} = (1/N^3)\ {}^{N}\sum_{o=1}{}^{N}\sum_{p=1}{}^{N}\sum_{q=1} \exp[-2\pi i(o\alpha + p\beta + q\lambda)/N] \cdot C_{o,p,q}$ [6]
  - One can use the Fast Fourier Transformation algorithm using equation [6] to calculate c for all orientation (360*360*180)
  - Complexity $O(N^3 \ln(N^3))$

Procedure - summary

1. Derive _a from C coordinates of a from [2]
2. Compute A*=DFT(_a) from [4]
3. Derive _b from C coordinates of b from [2]
4. Compute B =DFT(_b) from [4]
5. Compute C= A* . B fom [5]
6. Compute _c=IFT(C) from [6]
7. Look for sharp positive peakof _c
8. Rotate b to new orientation
9. Repeat step 3 to 8 until for all orientation
10. sort all peak by correlation score



Reference

E. Katchalski-Katzir, I. Sharvi, M. Eisenstein et al. “Molecular surface recognition: Determination of Geometric fit between protein and their ligands by correlation technique”. *Proc. Natl. Acad. Sci. USA.* pp. 2195-2199. March 1992.

Taking geometric to its edge: patch Docking (Unbounded Docking)

Dina schneidman-Duhovny,
Nussinov R,
Wolfson HJ

The idea

- The method is inspire by object recognition and image segmentation technique (computer vision)
- The idea could be compared to assembling a jigsaw puzzle. In solving a puzzle we try to match two pieces by picking one and searching for complimentary one.
- Likewise, given two molecule complex, we divide surface into patches and superimpose them using the algorithm.

Bounded/unbounded Docking

- Bounded -Co-crystallized molecule are separated artificially and the goal is to construct the resultant complex between them
- Unbounded - Protein complex are pulled apart in components and reassembled – good for confirmation change
- Method do the following:

1. Molecular surface fitting based on patches
2. Geometric hashing (for transformation detection)
3. Computation of shape complimentary using distance transformation.
4. Scoring
5. Filtering for “biological hot spot”

Method

- Divided into three major components

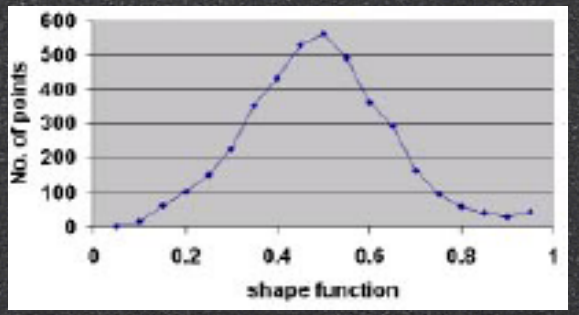
1. Shape representation - compute the surface for each molecule and detection geometric patches (concave, convex, flat surface)
2. Surface patch matching – match patches (concave to convex, flat to all type)
3. Filtering and scoring - scan and remove invalid configuration and rank according to complimentary score.

1.1. Shape representation

- Use the Connolly approach to compute the surface for each molecule.
- The calculated surface is preprocessed into a data structure: distance transformation grid and multi-resolution surface, used in the scoring routine
- Sparse surface representation is computed used in patches detection. It consist of points nicknamed “caps” (1), “pits” (3), “belts” (2)

1.2. detection of patches

- The goal is divide the shape into equal area with three type (concavities [“hole”], convexities [“knob”], flats [“flat”])
- construct a surface topology graph induced by the point of sparse surface.
  - G_{top} =(V_{top},E_{top}): V_{top}={critical point}, E_{top}={(u,v) if u and v belong to the same atom}
- Shape function calculation- group points into local curvature
  - A sphere of radii R is placed at the surface point.
  - The fraction of the sphere inside the solvent-exclude the protein volume is the shape function at this point.
  - The function point of all is calculated.
  - Score are sorted and the two cut point are identified separating knots and holes.



Example of patches

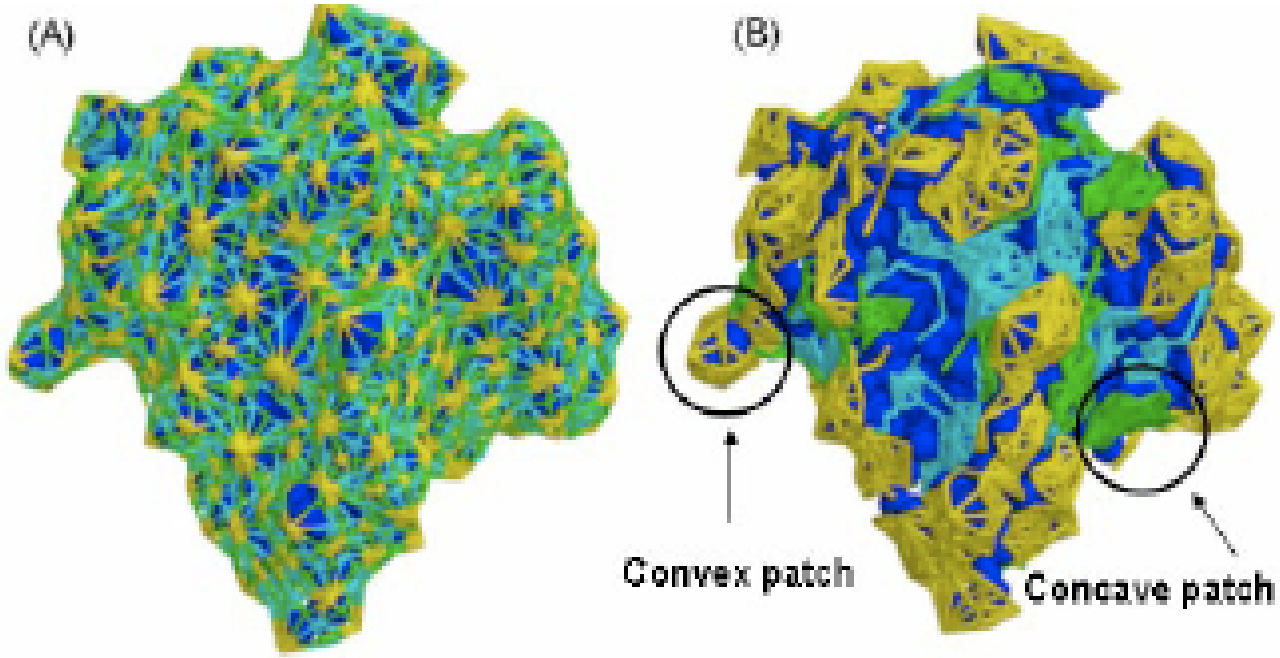


Fig. 2. (a) Surface topology graphs for trypsin inhibitor (PDB code 1ba7). The caps, belts and pits are connected with edges. (b) Geometric patches: the patches are in light colors and the protein is dark.

1.3. Patches detection

- A patch is connected set of critical points of the same type
- The idea is to divide the surface of molecule with non-intersection patches
- Construct G_{knob}, G_{hole}, G_{flat} as a subgraph of G_{top}
  - Run the graph algorithm for finding connected component
  - Run the merge and split routine to find equal surface area patches
- Detecting active sites
  - Hot-spot filtering (biological properties)

2. Surface patch matching

- The idea is to assembly a hypothetical docking configuration based on local geometric complimentary: knob<to>hole, flat<to>{flat,knob,hole}
- Two techniques are used which both utilizes geometric hash and pose clustering from computer vision
  - Single patch – one patch from receptor is matched with one patch from ligand.
  - Patch-pair – two patch from receptor are matched with two patches from ligand.

3. Filtering and scoring

- Since transformation is computed by matching local critical points within patches, it may results in “steric” clashes between receptor and ligands. We need to filter them out.
- Steric clash test – use the distance transform grid
  - For each candidate the transformation is applied to the surface of the ligand.
  - Compare the distance transform grid of the receptor with the coordinate of every point. If the distance is greater than a penetration threshold for each surface point then the transformation is filtered out.

3. Filtering and scoring

- Geometric scoring – divide the receptor into shells according to distance from the molecule surface
  - Each shell is represented by grid with distinction between interior, surface, and exterior atom.
  - Each shell is a range of distances from the transform grid e.g. [-5.0,-3.6),[-3.6,-2.2), [-2.2,-1.0),[-1.0,1.0),[1.0,-)
  - The score is the count of number of in each shell and geometric score is the weighted score of the shells.
- (this method is accurate but slow. There is another method faster but not accurate)
- The geometric score are ranked with biological information incorporated.

Reference

- Duhovny D, Nussinov R, Wolfson HJ. Efficient Unbound Docking of Rigid Molecules. In Gusfield et al., Ed. Proceedings of the 2'nd Workshop on Algorithms in Bioinformatics(WABI) Rome, Italy, Lecture Notes in Computer Science 2452, pp. 185-200, Springer Verlag, 2002

CTSS: A Robust and efficient Method for protein Structure Alignment Based on Local Geometric and Biological Features

Tolga Can
Yuan-Fang Wang

Method

- Based on theory of differential geometry on 3D space curve matching:
  - structural isomorphism of space curves is correspondence of their curvature and torsion values, expressed as a function of the intrinsic arc length.
  - Intrinsic arc length satisfies the property that C(s) |=| dC(s)/ds |=1, where C denote the space curve.
  - In reality such requirement cannot be satisfied but could be approximated using a smoothing function of Carbon alpha points.

Method -cont

- After smoothing C atom we compute the shape signature
  - A list of signature triplets, one for each of its residues
  - A signature triplet consist of secondary structure assignment, curvature and torsion value.
  - The signature are rotation and translation invariant
- Curvature provide local geometrical information
  - Build a hash table to index the space of invariant signature for each structure
  - Voting mechanism based on similarity of hash key is used to retrieve candidates

Method - cont

- To compre two protein we retrieve the signature from the hash tabe
- Use a dynamic programming algorithm to align pair-wise the signatures of the two protein structure
  - Result is a set of correspondences of structurally related residues
  - The C atom are superimposed and the RMSD value are computed for that subset

Procedure - summary

1. Calculate a spline fitting for C atom
2. Compute, for each residue, curvature and torsion value. Record the secondary assignment in signature too.
3. Compute a hash key of the signature
4. Construct a normalized scoring matrix based on distance between extracted feature
5. Run the local alignment algorithm
6. Superimpose the corresponding residue using RMSD

Reference

- Tolga Can,Yuan-Fang Wang, “CTSS: A Robust and efficient Method for protein Structure Alignment Based on Local Geometric and Biological Features” 2003