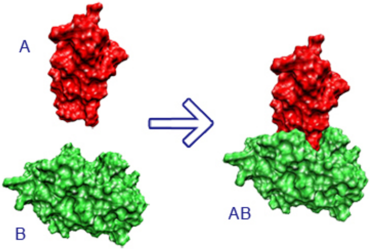
CS612 - Algorithms in Bioinformatics

Docking

May 7, 2025

Docking

Docking attempts to find the "best" matching between two molecules

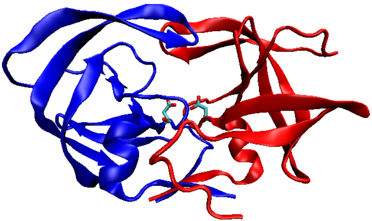


Docking – A More Serious Definition

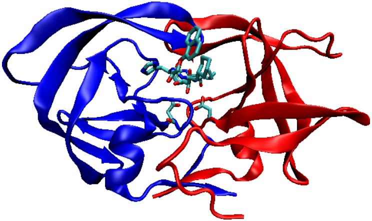
- Given two biological molecules determine:
- Whether the two molecules “interact”
- If so, what is the orientation that maximizes the “interaction” while minimizing the total energy of the complex
- **Goal:** To be able to search a database of molecular structures and retrieve all molecules that can interact with the query structure

Example: HIV-1 Protease, Finding Drug Candidates

Active Asp25

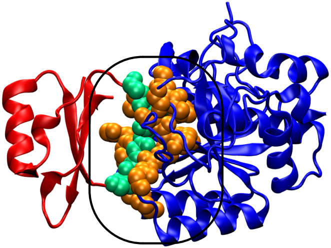


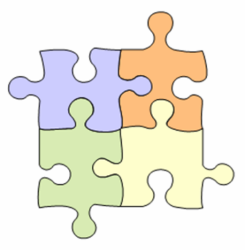
With inhibitor



Why is this difficult?

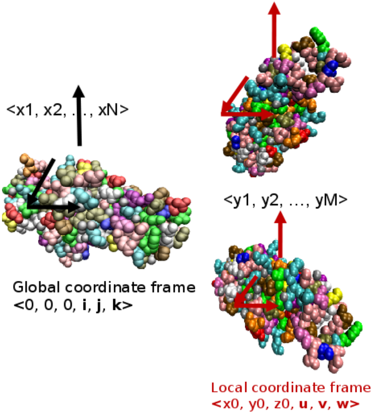
- Experimental detection of complexes is difficult
- Computational methods can come in handy...
- Computational docking methods try to find the "best" match between two or more molecules
- Goal: Find what is the orientation that maximizes the "interaction" while minimizing the total energy of the complex





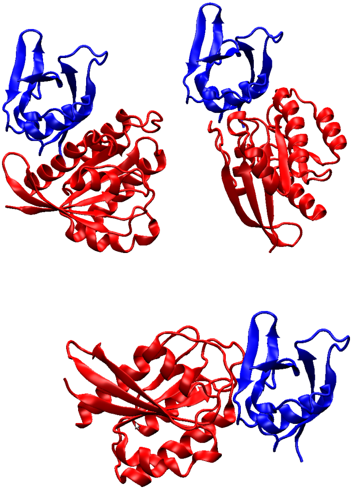
Why is this difficult?

- High dimensional search space: N X M + 6
- N, M = number of parameters to represent the unbound protein structures
- 6 = three rigid body translations and three rigid body rotations



Why is this difficult?

- Interaction site is not always known
- Geometry based methods often miss the correct binding site
- Energy differences between results are often small
- Structures may change upon binding



Why is this difficult?

- Key-lock model of docking
    - Assumes molecules are rigid
    - Docking primarily driven by shape complementarity
- Induced fit model of docking
    - Assumes molecules can induce changes to their structures
    - More difficult to account for induced fit

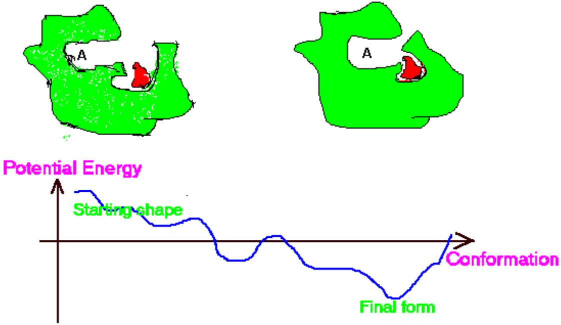
Bound vs. Unbound Docking

- In **bound docking** the goal is to reproduce a known complex where the starting coordinates of the individual molecules are taken from the PDB structure of the complex
- In other words, we take the two molecules from the complex, separate them and try to put them back together.
- In **unbound docking** the starting coordinates are taken from the unbound molecules.
- We take the PDB files of the separate molecules (not in a complex) and try to put them back together.
- It is a significantly more difficult problem, but more realistic.

Types of Molecular Docking

- Protein-Protein Docking
  - Both molecules usually considered rigid
  - 6 degrees of freedom
  - First apply steric constraints to limit search space and the examine energetics of possible binding conformations
- Protein-Ligand Docking
  - Flexible ligand, rigid receptor
  - Search space much larger
  - Either reduce flexible ligand to rigid fragments connected by one or several hinges, or search the conformational space using Monte Carlo methods or molecular dynamics

Forces that Drive Molecular Docking



Bound state of the system: lowest free energy of interaction between protein and ligand

Some Challenges in Pairwise Docking

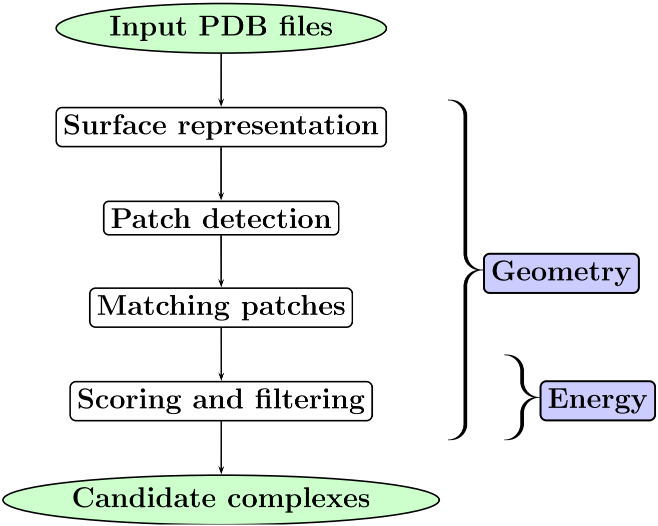
- **Where:** determination of active site (where to dock ligand)
    - Assumed known (Binding DB: Mostly for Protein-Ligand/Drug recognition sites)
    - Determined through geometry and/or homology methods
- **How:** orienting the ligand in the protein's active site
    - **Step 1:** representation/computation of the surfaces
    - **Step 2:** matching of the surfaces
- **Ranking:** an accurate yet efficient scoring function
    - Incorporates geometry (how well do the molecules fit)
    - Incorporates chemistry (which amino acids interact with which)

An Overall Framework for Pairwise Docking

- **Part 1:** Molecular shape representation
- **Part 2:** Matching of critical features
- **Part 3:** Filtering and scoring of candidate transformations

- The correct solution is found in 90% of the cases with RMSD under 5Å
- The rank of the correct solution can be in the range of 1 – 1000
- Needed: Fast yet accurate biochemical scoring function !!!

A More Detailed View of Docking Framework



Input PDB files

Surface representation

Patch detection

Matching patches

Scoring and filtering

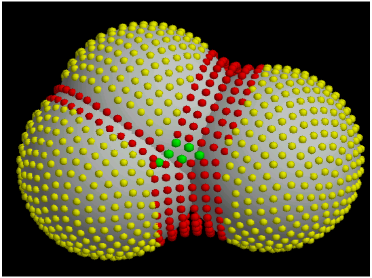
Candidate complexes

Geometry

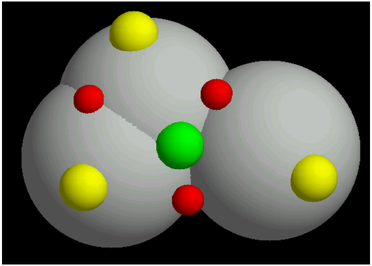
Energy

Surface Computation

Dense molecular surface (Connolly)

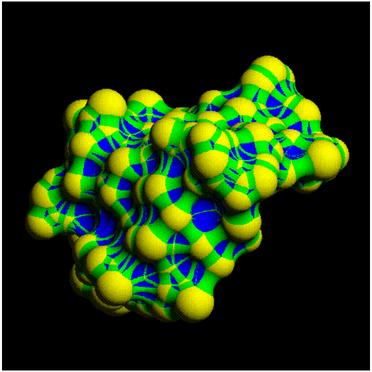


Sparse surface (Shuo Lin et al.)



- Most common: Connolly surface. Probe ball is rolled over the molecule, giving 3 types of points: caps (yellow points – belong to one atom), belts (red points – lie between two atoms), pits (green points – belong to the patches where the probe touches the 3 atoms).
- The representation is dense and can be reduced to local minima or maxima of the point patches. The shown sparse surface representation is by Shuo Lin

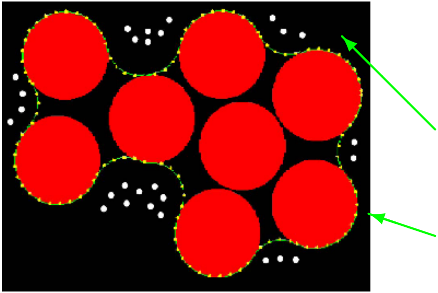
Connolly Surface



- Each atomic sphere is given the van der Waals (vdw) radius of the atom, resulting in the vdw surface representation
- Rolling a probe sphere over the vdw surface leads to the Solvent Reentrant Surface

Lenhoff Method for Complementary Surface

- Computes a "complementary" surface for the receptor, thus allowing the determination of possible positions for the atom centers of a prospective ligand

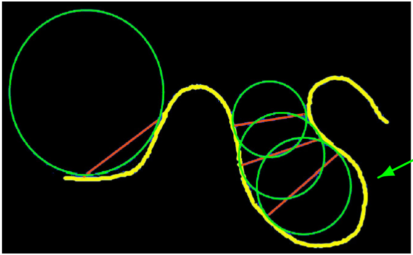


Potential atom centers of the ligand

Van der Waals surface

Kuntz Method for Clustered Spheres

- Uses clustered-spheres to identify cavities on the receptor and protrusions on the ligand
- Compute a sphere for every pair of surface points, i and j, with the sphere center on the normal from point i
- Regions where many spheres overlap are either cavities (on the receptor) or protrusions (on the ligand)

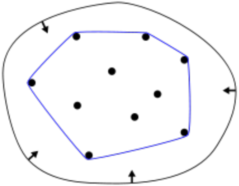


Cavity

Formalizing the Idea of Shape: Convex Hull

Definition (Convex Hull)

of a set X of points in the Euclidean plane or space, the *convex hull* is the smallest convex set that contains X. For instance, when X is a bounded subset of the plane, the convex hull may be visualized as the shape enclosed by a rubber band stretched around X.

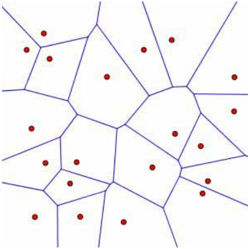


From wikipedia

Formalizing the Idea of Shape: Voronoi Diagrams

Definition (Voronoi Diagram)

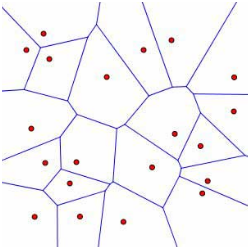
The Voronoi diagram of a point set P is a subdivision of the plane into cells with the property that each Voronoi cell of vertex p contains all locations that are closer to p than to every other vertex of P.



From http://www.ams.org/samplings/feature-column/fcarc-voronoi

Formalizing the Idea of Shape: Voronoi Diagrams

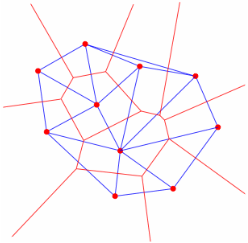
- Example: Dividing children into school districts based on their distance from a given school.
- For every pair of points, draw the line (plane in 3D) that passes in the middle (perpendicular to the line that passes between them).
- Cut off the line when it intersect with another line.



From http://www.ams.org/samplings/feature-column/fcarc-voronoi

Delauney Triangulation

- A triangulation of a three-dimensional point set S is any decomposition of S into non-intersecting tetrahedra (triangles for two-dimensional point sets).
- The **Delaunay triangulation** of S is the unique triangulation of S such that no circle (sphere) circumscribing a triangle (tetrahedron) in the triangulation contains any point in S.
- The Delaunay triangulation of a point set is a dual graph to the Voronoi diagram.
- Simply connect points from neighboring cells.

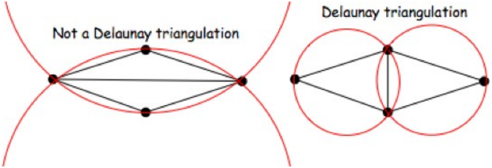


Delauney Triangulation

- The Delaunay triangulation of a point set is usually calculated by an incremental flip algorithm as follows:
  1. The points of S are sorted on one coordinate (x, y, or z).
  2. Each point is added in sorted order. Upon adding a point:
  3. The point is connected to previously added points that are "visible" to it, that is, to points to which it can be connected by a line segment without passing through a face of a tetrahedron.
  4. Any new tetrhedra formed are checked and flipped if necessary. Any tetrahedra adjacent to flipped tetrahedra are checked and flipped.
  5. This continues until further flipping is unnecessary, which is guaranteed to occur
- Naively, This algorithm runs in worst case $O(n^2)$ time, but expected $O(n^{3/2})$ time. With sorting – $O(n \log n)$

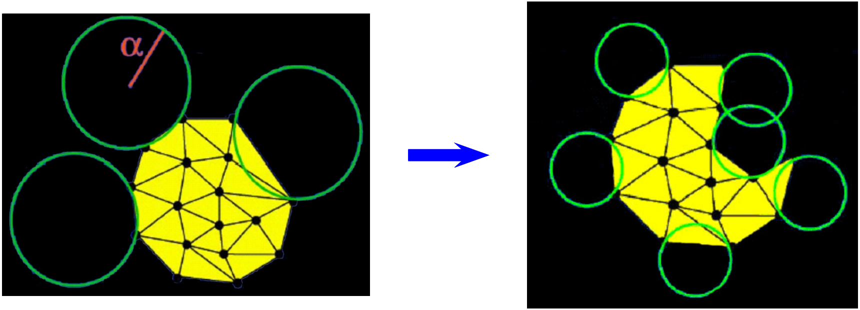
Delauney Triangulation – Flippling Points

- **Right:** The Delaunay triangulation of the four points.
- Note that the circumscribing circles on the left each contain one point of S, whereas the circles on the right do not.
- The transition from the triangulation on the left to that on the right is called an edge flip, and is the basic operation of constructing a two-dimensional Delaunay triangulation.
- Face flipping is the analogous procedure for five points in three dimensions.



Formalizing the Idea of Shape: Alpha Shapes

- In 2D, an "edge" between two points is "α-exposed" if there exists a circle of radius alpha such that the two points lie on the surface of the circle and the circle contains no other points from the point set

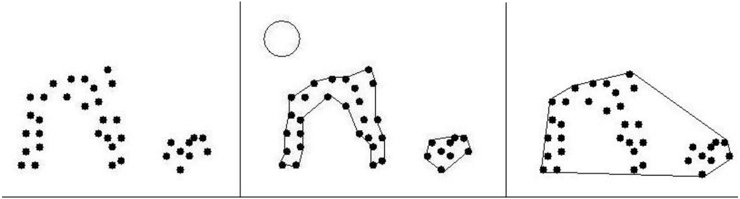


Formalizing the Idea of Shape: Alpha Shapes

- α-shapes are a generalization of the convex hull.
- Consider a point set S in 3D. Define an α-ball as a sphere of radius α. An α-ball is empty if it contains no points in S.
- For any α between zero and infinity, the α-hull of S is the complement of the union of all empty α-balls.
- For α of infinity, the α-shape is the convex hull of S. For α smaller than the 1/2 smallest distance between two points in S, the α-shape is S itself.
- For any α in between, one can think of the α-hull as the largest polygon (polyhedron) or set thereof whose vertices are in the point set and whose edges are of length less than 2α.
- The presence of an edge indicates that a probe of radius α cannot pass between the edge endpoints.

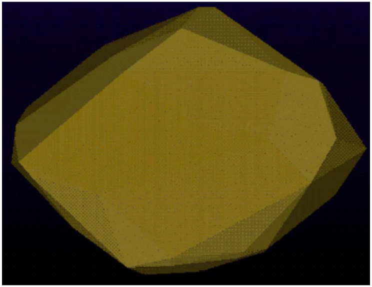
Alpha Shapes – Example

- **Left:** α is 0 or slightly more, such that an α-ball can fit between any two points in the set.
- The α-shape is therefore the original point set.
- **Middle:** the α-shape for α equal to the radius of the ball shown.
- This yields two disjoint boundaries, one of which has a significant indentation.
- **Right:** α is infinity, so an α-ball can be approximated locally by a line.
- α on this scale yields the convex hull of the point set.

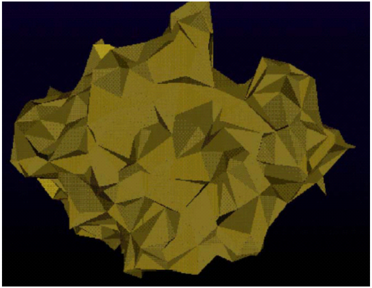


Examples of Alpha Shapes

$\alpha = \infty$, convex hull



α=3Å



Little insight into what value of alpha is needed – by trial and error

Computing Alpha Shapes from Delauney Triangulation

- Although Delaunay triangulation is incidental to α-shapes, note that the Delaunay triangulation maximizes the average of the smallest angle over all triangles.
- In other words, it favors relatively even-sided triangles over sharp and stretched ones.
- From the Delaunay triangulation the α-shape is computed by removing all edges, triangles, and tetrahedra that have circumscribing spheres with radius greater than α.
- Formally, the α-complex is the part of the Delaunay triangulation that remains after removing edges longer than α.
- The α-shape is the boundary of the α-complex.

Computing Alpha Shapes from Delauney Triangulation

- Pockets can be detected by comparing the α-shape to the whole Delauney triangulation.
- Missing tetrahedra represent indentations, concavity, and generally negative space in the overall volume occupied by the protein.
- Particularly large or deep pockets may indicate a substrate binding site.
- The volume of a molecule can be approximated using the space-filling model where each atom is modeled as a ball whose radius is α, where α is selected depending on the model being used: Van der Waals surface, molecular surface, SASA, etc.

Calculating Molecular Volume from Alpha Shapes

- Calculating the volume of a complex of overlapping balls is non-trivial because of the overlaps.
- If two spheres overlap, the volume is the sum of the volumes of the spheres minus the volume of the overlap, which was counted twice.
- If three overlap, the volume is the sum of the ball volumes, minus the volume of each pairwise overlap, plus the volume of the three-way overlap.
- In the general case, all pairwise, three-way, four-way and so on to n-way intersections (assuming there are n atoms) must be considered.
- Proteins generally have thousands or tens of thousands of atoms, so the general n-way case may be computationally expensive and may introduce numerical errors

Calculating Molecular Volume from Alpha Shapes

- calculate the volume of a protein, we take the sum of all ball volumes, then subtract only those pairwise intersections for which a corresponding edge exists in the α-complex.
- Only those three-way intersections for which the corresponding triangle is in the α-complex must then be added back.
- Finally, only four-way intersections corresponding to tetrahedra in the α-complex need to be subtracted.
- No higher-order intersections are necessary, and the number of volume calculations necessary corresponds directly to the complexity of the α number of atoms.

Recommended reading: Liang, H. Edelsbrunner et al. Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape. Proteins: Structure, Function, and Genetics, 33:1–17, 1998.
Class notes by Prof. Lydia Kavraki: `http://cnx.org/content/m11616/latest/`

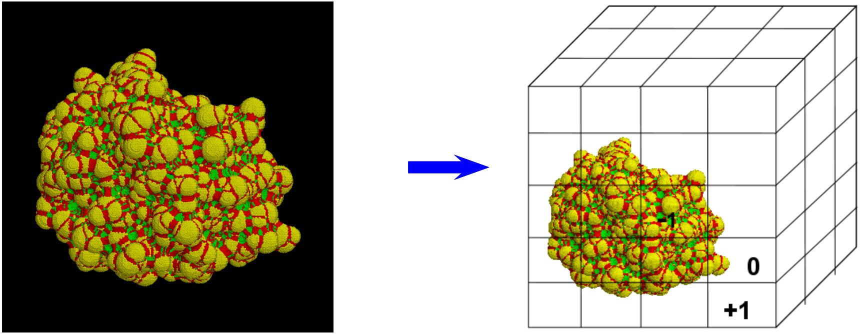
Surface Matching

- Find the transformation (rotation + translation) that will maximize the number of matching surface points from the receptor and the ligand
- First satisfy steric constraints:
  - Find the best fit of the receptor and ligand using only geometrical constraints
- Then use energy calculations to refine the docking
  - Select the fit that has the minimum energy

Docking Programs

- ZDOCK, RDOCK
- PyDock
- ClusPro
- AutoDock (Olson, Scripps)
- Bielefeld Protein Docking
- DOCK (Kuntz, UCSF)
- DOT
- FTDock, RPScore, MultiDock
- GRAMM
- Hex 3.0
- FlexX
- ICM Protein-Protein docking (Abagyan group, currently best)
- KORDO
- MolFit
- MPI Protein Docking
- RosettaDOCK (Baker, Washington Univ., Gray, Johns Hopkins Univ.)
- INVDOCK (Y. Z. Chen, NUS)

Grid-detection of Complementary Surface



Computing shape complementarity is based on determining regions in the grid that are not occupied by the protein's atoms but are filled by the ligand atoms

Docking Using Fast Fourier Transform (FFT)

- Grid detection is the basis of many docking algorithms.
- The algorithms project the two molecules A and B on a 3-D grid of $N \times N \times N$ points.
- Computing shape complementarity is based on determining regions in the grid that are not occupied by the protein's atoms but are filled by the ligand atoms.
- Each grid point is represented by two discrete functions:

$$\bar{a}_{l,m,n} = \begin{cases} 1 & \text{Grid point on surface of molecule A} \\ \rho & \text{Grid point in molecule A} \\ 0 & \text{Grid point outside molecule A} \end{cases}$$

and:

$$\bar{b}_{l,m,n} = \begin{cases} 1 & \text{Grid point on surface of molecule B} \\ \delta & \text{Grid point in molecule B} \\ 0 & \text{Grid point outside molecule B} \end{cases}$$

Docking Using Fast Fourier Transform (FFT)

- Matching of surfaces is accomplished by calculating the correlation between the discrete functions $\bar{a}$ and $\bar{b}$ is defined as

$$\bar{c}_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \bar{a}_{l,m,n} \cdot \bar{b}_{l+\alpha,m+\beta,n+\gamma}$$

- If the shift vector α, β, γ is such that there is no contact between the two molecules the correlation value is zero.
- If there is contact between the surfaces the correlation is positive.
- Large negative values are assigned to ρ in A and small nonnegative values to δ in B.
- Thus, when molecule B penetrates molecule A, it is a negative contribution to the overall correlation value.
- The correlation value for each displacement is the score for overlapping surfaces corrected by the penalty for penetration.

Docking Using Fast Fourier Transform (FFT)

- the Fourier transform allows for a fast calculation. The Discrete Fourier Transform (DFT) of a function xl, m, n is defined as:

$$X_{o,p,q} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} e^{-2\pi i(ol+pm+qn)/N} \cdot x_{l,m,n}$$

- Applying this to $\bar{c}$ above yields:

$$C_{o,p,q} = A^{*}_{o,p,q} \cdot B_{o,p,q}$$

C and B are the DFTs of $\bar{c}$ and $\bar{b}$, respectively and A^{*} is the complex conjugate of the DFT of $\bar{a}$.

Docking Using Fast Fourier Transform (FFT)

- Therefore we need to calculate A^* and B and simply multiply them. To revert to the original correlation we have to invert the Fourier Transform as:

$$\bar{c}_{\alpha,\beta,\gamma} = \frac{1}{N^3} \sum_{o=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} e^{2\pi i(o\alpha + p\beta + q\gamma)/N} \cdot C_{o,p,q}$$

- Using Fast Fourier Transform (FFT) it is possible to calculate the Fourier Transform in $O(N^3 \log N^3)$.
- In practice, molecule A is fixed and molecule B is rotated at fixed intervals of some Δ degrees, and the FFT is calculated for each orientation.
- This results in a complete scan of $360 \times 360 \times 180/\Delta^3$ orientations.

ClusPro

- The PIPER method is based on Fast Fourier Transform (FFT) correlation approach.
- It places protein A at the origin of the coordinate system on a fixed grid, and perturbs the second protein on a moveable grid.
- Then the docking energy is calculated based on FFT correlation function. The correlation function made up electrostatic interaction and desolvation contribution.
- The resulting conformations are clustered based on their IRMSD.
- For each cluster, the Van der Waals energy is minimized using the Charmm potential function for up to 300 steps with a fixed backbone to remove small steric clashes.

Flexible Docking?

- In reality, molecules may change their conformation upon binding.
- Introducing full flexibility is very very challenging.
- Introducing partial flexibility in the receptor and/or ligand flexibility is more feasible.

From Flexible Ligand to Flexible Receptor?

- Modeling full receptor flexibility is very difficult!
- In order for this process to become efficient, we must find a representation for protein flexibility that avoids the direct search of a solution space comprised of thousands of degrees of freedom.
- There are several methods available, and the accuracy of the results is usually directly proportional to the computational complexity of the representation.

Receptor Flexibility – Soft Receptor

- Soft receptors can be easily generated by relaxing the high VdW energy penalty
- The rationale is that the receptor structure has some inherent flexibility which allows it to adapt to slightly differently shaped ligands.
- If the change in the receptor conformation is small enough, it is assumed that the receptor is capable of such a conformational change.
- It is also assumed that the change in protein conformation does not incur a sufficiently high energetic penalty to offset the improved interaction energy between the ligand and the receptor.
- It is also quite easy to implement (relax the collision component).

Receptor Flexibility – Selecting Specific DOFs

- is it possible to select only a few degrees of freedom to model explicitly.
- They usually correspond to rotations around single bonds
- These degrees of freedom are usually considered the natural degrees of freedom in molecules.
- Rotations around bonds lead to deviations from ideal geometry that result in a small energy penalty when compared to deviations from ideality in bond lengths and bond angles.
- Selection of which torsional degrees of freedom to model is usually the most difficult part of this method because it requires a considerable amount of a priori knowledge.
- The torsions chosen are usually rotations of side chains in the binding site of the receptor protein.
- It is also common to further reduce the search space by using rotamer libraries.

Receptor Flexibility – Ensemble Docking

- One possible way to represent a flexible receptor for drug design applications is the use of multiple static receptor structures
- The best description for a protein structure is that of a conformational ensemble of slightly different protein structures coexisting in a low energy region of the potential energy surface.
- The structures can be determined experimentally either from X-ray crystallography or NMR, or generated via computational methods such as Monte Carlo or MD simulations.

Modeling Limited Receptor Flexibility

- Selection of specific degrees of freedom such as on designated amino acids on binding site
- Shown here: Acetylcholinesterase: Phe330 flexible – acts as swinging gate

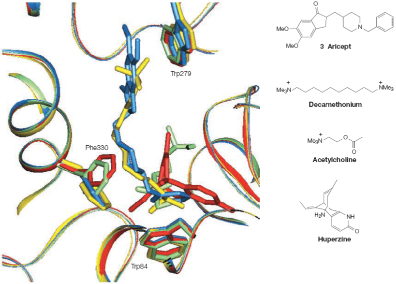


Figure 5 | **Multiple conformations of a single residue.** Overlay of native and three *Torpedo californica* acetylcholinesterase–ligand complexes using the protein C-α atoms (Protein Data Bank codes 2ACE, 1EVE, 1VOT and 1ACL). Key protein side chains are indicated by thick lines, as are the inhibitors. The colour codes are: donepezil (Aricept, blue), decamethonium (yellow), native (green), huperzine (red). The flexibility of Phe330, in comparison with the rigidity of the rest of the gorge, is highlighted.

Modeling Limited Receptor Flexibility

Moving larger number of amino acids (illustration on acetylcholinesterase)

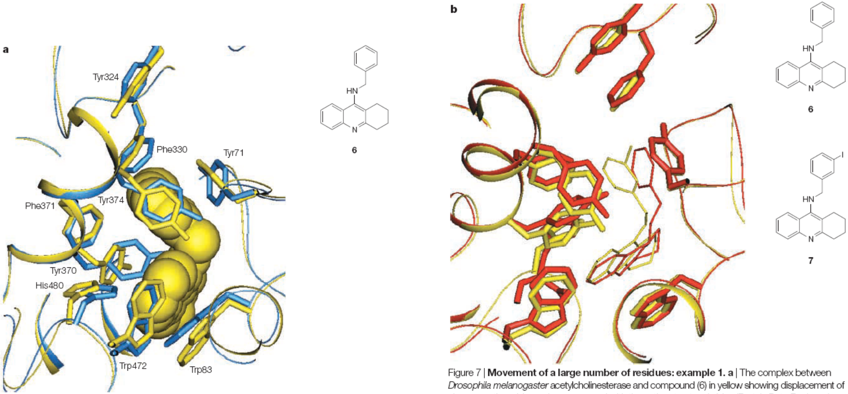
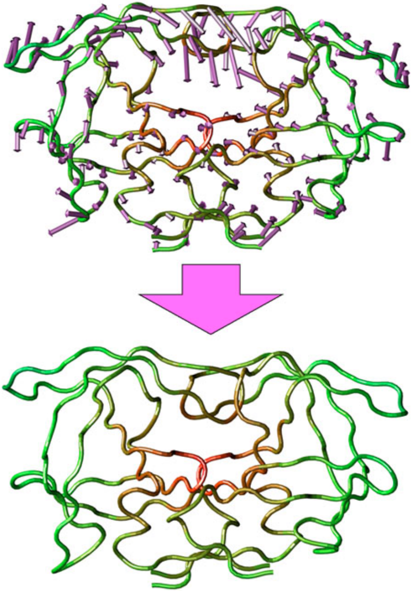


Figure 7 | **Movement of a large number of residues: example 1. a** | The complex between *Drosophila melanogaster* acetylcholinesterase and compound (6) in yellow showing displacement of nine aromatic residues when compared with the native structure in blue (Protein Data Bank codes 1QO9 and 1DX4). **b** | Overlay of the complexes between *Drosophila melanogaster* acetylcholinesterase complexed with compounds (6) in red and (7) in yellow. Both the side chains of the protein and the positions of the inhibitors are altered (Protein Data Bank codes 1QON and 1DX4).

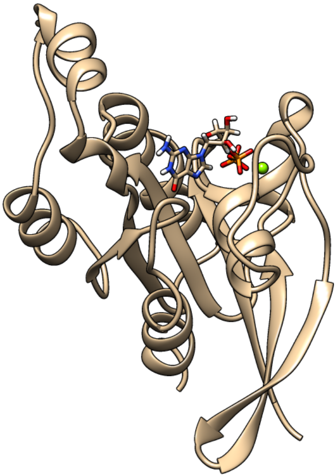
Receptor Flexibility – Collective DOF

- Collective DOF allows the representation of full protein flexibility without a dramatic increase in computational cost.
- One method is the calculation of normal modes for the receptor.
- Alternatively, we can use dimensionality reduction methods.
- The most commonly used method for the study of protein motions is principal component analysis (PCA).



Protein-Ligand Docking

- A ligand is a small molecule (drug, inhibitor, activator/deactivator etc.)
- Small molecules very often bind to proteins in specific sites
- Due to the small size of the ligand, flexibility is more feasible.



Protein-Ligand Docking - Autodock Vina

- Flexible ligand – define rotatable bonds.
- Limited flexibility in the receptor – user can define flexible side chains.
- Scoring function contains clashes, hydrophobic interactions and hydrogen bonds.
- Iterated Local Search global optimizer – a succession of steps consisting of a mutation and a local optimization are taken, with each step being accepted according to the Metropolis criterion.
- Local optimization is done using Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, that optimizes the energy function and its gradient with respect to the position and orientation of the ligand.

Critical Assessment of Protein Interactions (CAPRI)

- Protein-protein docking competition
- The equivalent of CASP for protein-protein docking
- Community-wide experiment that started in 2001
- Interesting review of docking methods: S. Vajda & C. J. Camacho. Protein-protein docking: is the glass half-full or half-empty? Trends in Biotechnology, 22(3):110-116, 2004.
- State-of-the-art methods – ClusPro (using FFT), ZDOCK (using FFT, desolvation energy, electrostatics and shape complementarity), RosettaDock (identifying near-native interactions near a starting point),...