Lecture 21 -- Docking
Review of Normal Mode Analysis

- What is the basic assumption of normal mode analysis?
- Why an energy minimization is needed?
- What are the three normal modes of water?
- What is wave number?
- What motions are responsible for the Amide I band in infrared spectroscopy? Does its frequency change with the secondary structure?
- For an harmonic oscillator, knowing the spring constant (k) and mass (m), what is the frequency of the oscillation?
- What are the definitions of eigenvalues and eigenvectors of a matrix?
- How do you calculate the eigenvalues and eigenvectors of a matrix?
- What are the general steps of the normal mode analysis of a protein?
- Does the NMA provide the amplitude of the motion?
- What is elastic network model?
Today’s agenda

- Principle component analysis (follow-up on Lecture 20)
- Docking--
  - Small molecule screening by docking
  - Manual docking in MOE
  - Docking and solvation (ContextShapes)
  - Anchor points for Docking (SmoothDock)
21.1 Principle Components Analysis
Normal Modes versus Principle Components

<table>
<thead>
<tr>
<th>Normal Modes</th>
<th>Principle Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Single input structure</td>
<td>• Many input structures.</td>
</tr>
<tr>
<td>• Assumed to be in energy minimum</td>
<td>• Assumed to be equilibrated.</td>
</tr>
<tr>
<td>• Harmonic motions</td>
<td>• Harmonic motions</td>
</tr>
<tr>
<td>• Solved as an eigenvector problem</td>
<td>• Solved as an eigenvector problem</td>
</tr>
<tr>
<td>• Uses Hessian matrix</td>
<td>• Uses covariance matrix</td>
</tr>
<tr>
<td>• Eigenvectors form a basis set for $R^N$</td>
<td>• Eigenvectors form a basis set for $R^N$</td>
</tr>
</tbody>
</table>

Compare, contrast
PCA as a least-squares approximation problem

\[ \vec{q}(t) \approx \sum_{\alpha=1}^{M} a_\alpha(t) \vec{\eta}_\alpha \]

where \( q(t) \), \( t=1,N \), are the coordinates of \( N \) atoms at time \( t \) in a MD simulation. Thus \( q_i \) is the coordinate \( i \) over the whole trajectory, \( q_i(t) \) is the coordinate \( i \) at time \( t \) (a scalar real value), and \( q \) is the entire MD trajectory.

\( M \) is the number of degrees of freedom (\textit{dof}).

\( \eta_\alpha \) is one of the \( M \) basis vectors (an eigenvector, a \textit{principle component}), and \( a_\alpha(t) \) is a time-dependent value for each \textit{dof}, each \( t \).

The vectors \( \eta_\alpha \) are eigenvectors of \( K \), the covariance matrix. They are determined by solving.

\[ \lambda_\alpha \vec{\eta}_\alpha = K_T \vec{\eta}_\alpha \]
Normal modes: Hessian matrix

Many atom case

\[
\begin{bmatrix}
\frac{\partial^2 E}{\partial x_1 \partial x_1} & \frac{\partial^2 E}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 E}{\partial x_n \partial x_1} \\
\frac{\partial^2 E}{\partial x_1 \partial x_2} & \frac{\partial^2 E}{\partial x_2 \partial x_2} & & \\
\vdots & \ddots & \ddots & \\
\frac{\partial^2 E}{\partial x_1 \partial x_n} & \frac{\partial^2 E}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 E}{\partial x_n \partial x_n}
\end{bmatrix}
\begin{bmatrix}
\Delta x_1 \\
\Delta x_2 \\
\vdots \\
\Delta x_n
\end{bmatrix}
=
\begin{bmatrix}
F_1 \\
F_2 \\
\vdots \\
F_n
\end{bmatrix}
\]

So the force vectors \( F \) as a function of the vector of displacement vectors \( \mathbf{x} \) is

\[
\mathbf{H} \cdot \mathbf{x} = \mathbf{F}
\]
Principle Components: Covariance matrix

\[
\begin{pmatrix}
\langle x_1 x_1 \rangle & \langle x_1 x_2 \rangle & \cdots & \langle x_1 x_n \rangle \\
\langle x_1 \rangle < x_1 \rangle & \langle x_1 \rangle < x_2 \rangle & \cdots & \langle x_1 \rangle < x_n \rangle \\
\langle x_2 x_1 \rangle & \langle x_2 x_2 \rangle & \cdots & \langle x_2 x_n \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle x_n x_1 \rangle & \langle x_n x_2 \rangle & \cdots & \langle x_n x_n \rangle
\end{pmatrix}
\begin{pmatrix}
\eta_{1,1} \\
\eta_{1,2} \\
\vdots \\
\eta_{1,n}
\end{pmatrix}
= \lambda_1
\begin{pmatrix}
\eta_{1,1} \\
\eta_{1,2} \\
\vdots \\
\eta_{1,n}
\end{pmatrix}
\]

\[\lambda_\alpha \vec{\eta}_\alpha = K_T \vec{\eta}_\alpha\]
Solving for eigenvectors

• Eigenvectors may be solved by setting the determinant of \( K = 0 \), then finding all of the solutions of that equation. Called eigendecomposition.

• More easily:
  – Multiply a random vector by \( K \), many times. The resulting vector is the first eigenvector \( \vec{\eta}_1 \) (first PC).
  – Project a random vector on the hyperplane normal to \( \vec{\eta}_1 \). Multiply by \( K \). Project the result on the hyperplane normal to \( \vec{\eta}_1 \). Repeat until converged. The result is \( \vec{\eta}_2 \). And so on.
Each element of the covariance matrix $K_{ij}$ is a correlation over time (t) between the coordinate $q_i$ and the coordinate $q_j$. The matrix is symmetrical because $K_{ij} = K_{ji}$
Principle components analysis, How-To

• Run a MD simulation. For maximum significance, this should be an *equilibrium* simulation. (As you know, we can only approximate equilibrium)

• Superpose samples on one of the samples (doesn’t matter which one) and center (by subtracting the center of mass)

• Calculate covariance matrix $\mathbf{K}$.

• Solve for eigenvectors.

• Project MD trajectory to PC-space.
All points can be constructed from PCs.

\[ u_1 \text{ is the first PC. } u_2 \text{ is normal to } u_1. \text{ Each point } x^i \text{ (blue) can be expressed as a point on } u_1 \tilde{x}^i, \text{ plus a fraction of } u_2. \]

Every atom configuration in a MD trajectory can be generated from PCs that were solved from that MD trajectory.
Projecting MD trajectory to PC-space.

• To get the PC coordinate, project the coordinate vector $q(t)$ onto the PC $\eta_\alpha$

$$q(t)^\alpha = q(t) \cdot \eta_\alpha$$

Each sample from a MD trajectory can be projected onto a few PCs. Then 2D contour plots can be drawn to show where the molecule was spending most of its time. Peaks in this plot can be projected back to real space.
PC faces represent the most common, most different faces. First PC is the average face. 2nd PC the most common differences from the average face. 3rd PC the most common differences from the first and second PCs, etc.
21.2 Docking
Docking is the process of predicting the stable three-dimensional structure of a bound pair of molecules -- a **pose**.

Alternatively, the prediction can be negative -- no pose.

There are two types of docking algorithm

- Descriptor-based
- Energy based
DOCKING- algorithms

• Descriptor-base algorithms
  – A type of QSAR (quantitative structure-activity relationships)
  – Fast.
  – Convert a binding site (receptor) into a set of descriptors.
    • Spheres
    • ellipsoids (alpha shapes)
    • internal distances
    • electrostatic potentials
  – Convert a small molecule (ligand) or a large set of small molecules, or a large number of conformations, into a set of descriptors
    • spheres, ellipsoids, etc.
  – Find correlations. Find complementary descriptors.
  – Generally must be followed up by an energy-based screen
DOCKING- algorithms

• Energy-base algorithms
  – Define the representation of the receptor and ligand.
    • all protein atoms.
    • all protein atoms and bound waters.
    • just backbone atoms.
    • just alpha-carbons.
    • all atoms with moveable sidechains.
    • all atoms with moveable backbone atoms and side chains
    • Spheres. Ellipsoids.
    • Surfaces. Masks. Fourier transforms.
  – Define an energy function
    • Physics-based
    • Knowledge-based
  – Search for the lowest energy pose.
Searching for the lowest energy pose

- Six-dimensional search -- $x,y,z,\phi,\psi,\kappa$
- Additional dimensions if any atoms are flexible.
- Slow. Requires tricks.
21.3 Drug discovery -- small molecule screening
In silico drug screening by docking

• Drug discovery costs are too high: ~$800 millions, 8~14 years, ~10,000 compounds (DiMasi et al. 2003; Dickson & Gagnon 2004)
• Drugs interact with their receptors in a highly specific and complementary manner.
• Core of the target-based structure-based drug design (SBDD) for lead generation and optimization.

A “Lead” is a compound that
  – shows biological activity,
  – is novel, and
  – has the potential of being structurally modified for improved bioactivity, selectivity, and drugeability.
What is Drugable?

• Lipinski’s rule of five: Poor oral absorption and/or distribution are more likely when
  – MW > 500
  – logP (log of the n-octanol/water partition coefficient) > 5 (too hydrophobic)
  – > 5 H-bond donors (number of OH and NH groups)
  – >10 H-bond acceptors (number of N and O atoms)
• If a compound meets 2 conditions above, it is not suitable for further development.
Sources of molecules for screening

- Drug-like: MDDR (MDL Drug Data Report) >147,000 entries, CMC (Comprehensive Medicinal Chemistry) >8,600 entries
- Non-drug-like: ACD (Available Chemicals Directory) ~3 million entries
- Literatures and databases, Beilstein (>8 million compounds), CAS & SciFinder
- Compounds with 3D structures: CSD (Cambridge Structural Database, www.ccdc.cam.ac.uk): ~3 million X-ray crystal structures for >264,000 different compounds and >128,000 organic structures
- Compounds you can buy
  - Available without exclusivity: various vendors (& ACD)
  - Available with limited exclusivity: Maybridge, Array, ChemDiv, WuXi Pharma, ChemExplorer, etc.
- Corporate databases: a few millions in large pharma companies
Success stories

- HIV protease inhibitor amprenavir (Agenerase) from Vertex & GSK (Kim et al. 1995)

- HIV: nelfinavir (Viracept) by Pfizer (& Agouron) (Greer et al. 1994)

- Influenza neuraminidase inhibitor zanamivir (Relenza) by GSK (Schindler 2000)
21.4 Descriptor-based methods for drug discovery
Docking Applications

- Determine the lowest free energy structures for the receptor-ligand complex
- Search database and rank hits for lead generation (virtual screening)
- Calculate the differential binding of a ligand to two different macromolecular receptors
- Propose modification of a lead molecules to optimize potency or other properties
- de novo drug design for lead generation
- protein interaction design (protein therapeutics)
Descriptor Matching Methods

- Distance-compatibility graph in DOCK (Ewing and Kuntz 1997): distances between sphere centers and distances between ligand heavy atoms
- Interaction site matching in LUDI (Boehm 1992): HBA<->HBD, HYP<->HYP
- Shape-matching in FRED (Openeye www.eyesopen.com)
- Shape chemical complementarity in SANDOCK (Burkhard et al. 1998)
- Surface complementarity in LIGIN: (Sobolev et al. 1996)
- H-bond matching in ADAM (Mizutani et al. 1994)
Limitations

• Limitation of experimental structures (Davis et al. 2003) in Protein Data Bank:
  – Hydrogens, water molecules, and metal ions may be absent.
  – Misplaced atoms (e.g., ~1/6 of N,O of Asn & Gln flipped, and His sidechains rotated in PDB).
  – Uncertainty from conformational flexibility and/or static disorder in crystals.
  – Overmodeling of waters, ligands.

• Limitations of structural descriptors:
  Coarse graining
  Polarizability of electrostatics ignored
  Flexibility not always modeled
  Other players usually ignored
  May distort space.
Challenges with descriptor matching methods

• Descriptor-matching: using pattern-recognizing geometric methods to match ligand and receptor site descriptors
  – geometric, chemical, pharmacophore properties, such as distance pairs, triplet, volume, vector, hydrogen-bond, hydrophobic, charged, etc.

• Challenges
  – Complete conformation and configuration space of ligand and receptor complex are too large.
  – Conformational flexibility of both ligand and receptor can’t be ignored.
  – Effect of water is hard to model.
  – Shape alone is not sufficient descriptor to identify low-energy conformations of a ligand-receptor complex (Jorgensen 1991).
Pharmacophore model

• Defined as a set of descriptors that correlate with function in a set of ligands for a particular receptor.
  – For example, all ligands that bind HIV protease have aromatics and have MW > 300.
  – All trypsin inhibitors have a positively charged side chain.

• Pharmacophores can be used to pre-screen potential leads.
Descriptor Matching Method: UCSF DOCK

- Place spheres on receptor to make “negative space”
- Find compatible sets of distances between sphere centers and ligand moieties.

Go to http://dock.compbio.ucsf.edu/Overview_of_DOCK/index.htm To see algorithm. Or google UCSF DOCK.
Docking site in HIV protease

Drug binding site in a cavity of protein
Ligand–protein docking:
Step 1: Creation of spheres to fit a cavity
Ligand-protein docking concept

Ligand–protein docking:
Step 2: Place a ligand to match the positions of spheres
Ligand-protein docking: Step 3: Check chemical complementarity.
Docking Software

DOCK: (Kuntz et al. 1982)
DOCK 4.0 (Ewing & Kuntz 1997)
DOCK 6.0
AutoDOCK (Goodsell & Olson 1990)
AutoDOCK 3.0 (Morris et al. 1998)
GOLD (Jones et al. 1997)
FlexX: (Rarey et al. 1996)
GLIDE: (Friesner et al. 2004)
ADAM (Mizutani et al. 1994)
CDOCKER (Wu et al. 2003)
CombiDOCK (Sun et al. 1998)
DIVALI (Clark & Ajay 1995)
DockVision (Hart & Read 1992)
FLOG (Miller et al. 1994)
GEMDOCK (Yang & Chen 2004)
Hammerhead (Welch et al. 1996)
LIBDOCK (Diller & Merz 2001)
MCDOCK (Liu & Wang 1999)
PRO_LEADS (Baxter et al. 1998)

SDOCKER (Wu et al. 2004)
QXP (McMartin & Bohacek 1997)
Validate (Head et al. 1996)
ContextShapes (Shentu et al, 2008)
SmoothDock (Camacho)
ZDOCK (Z. Weng)
PatchDock (Nussinov)
Accuracy of Docking

• **Reality Boundary**
  – Experimental errors: 0.1-0.25 kcal/mol (18-53% relative error in binding constant), as much as 0.65 kcal/mol (3 fold error in binding constant)
  – **Free energy calculation accuracy:** ~1 kcal/mol (5.4 fold error in binding constant) starting with an accurate geometric model & full sampling
  – **Entropy and solvation estimation need a sufficiently long simulation run with an accurate force field, an ensemble of explicit of water molecules, and full sampling**

• **Current**
  – Reproduce X-ray structure with rmsd<2Å: 50-90% achievable
  – Binding affinity: 1.5~2 log unit (32-100 fold, 2.05-2.73 kcal/mol)
  – Correlation between scores and affinities, r^2<0.3
  – Enthalpy ranking with minimization: ±5 kcal/mol
  (Wang et al. 2003; Erickson et al. 2004; and others.)
Challenges for Docking Method

• Large energies vs. small energy differences
• Find weakly potent compounds in pools of nonbinders
• High false positives and false negatives from *in silico* screen
• Explicit water are needed for: volume, change shape of the binding site, bridging interaction
• A scoring function that always has its global optimum in agreement with the experiment
• Good affinity prediction not necessarily leads to correct binding mode
• Speed and accuracy
Docking strategy using MOE

- **Prepared Receptor**
  - Protonate 3D
  - Rotamer Explorer
  - Protein Geometry (Site Finder)

- **1: Conformations**
  - Rotate Bonds

- **2: Placement**
  - Generate poses from conformations

- **3: Rescoring(1)**
  - Apply scoring function

- **4: Refinement**
  - MM or grid energetics

- **5: Pharmacophore Constraint**
  - Filter final poses

- **6: Rescoring(2)**
  - Apply scoring function

- **Ligand Sources**
  - MDB file
  - Selected Atoms
  - Selected Residues
  - Selected Chains
  - (With or without conformations)

- **MOE Database of Docked Ligands**
Work on Homework 12

• Download instructions from web site.
• Watch demo first.
• Follow along.
• Give yourself 6 hours.
• Due April 25.
21.5 Energy-based docking -- shape matching
Surfaces can be represented as a “ray ball” centered on a point on the surface. The rays are composed on 32 points. Each point is represented by a bit in a 32-bit word, 1 for inside the surface, 0 for outside.

1 ray ball (called a ContextShape) is composed of 1256 rays, or 5024 bytes of data.

Runs only on large-memory (40Gb) machines.
Two contextshapes (CS) are superimposed by their centers. The surface volume that overlap are proportional to the amount of water lost.

To rapidly rotate CS, a precalculated look-up table of rotated rays is stored. A rotated CS corresponds approximately to re-mapping of the ray indeces.

\[ \text{OV} = \text{Overlap volume} \]

\[ \text{OV}(\text{CS}_{\text{vol}}^L, \text{CS}_{\text{X}}^R, \pi) = \sum_{i=1}^{K} V(\text{CR}_i^L \land \text{CR}_i^R) \]

\( \land \) = logical AND operator.

\( \pi \) = pose

\( L \) = ligand, \( R \) = receptor

\( V \) = count bits function, \( i\pi \) = re-mapped indeces.
Molecular surfaces were calculated for each protein and were projected inward (Inner) and outward (Outer). Combinations of layered CS were used to score poses. The optimal combination of layer scores was determined by machine learning using known poses.
ContextShapes algorithm

1. $CS_R$ ← context shapes from receptor protein $P_R$;
2. $CS_L$ ← context shapes from ligand protein $P_L$;
3. Candidate-Pairs ← $\emptyset$;
4. foreach (Context Shape $CS_R$ in $CS_R$) do
5.   foreach (Context Shape $CS_L$ in $CS_L$) do
6.     Calculate the average Solid Angle sum $\Theta$ over $CS_R$ and $CS_L$;
7.     Calculate angle $\phi$ between Solid Vector of $CS_L$ and Reversed Solid Vector of $CS_R$;
8.     if ($\phi$ and $\Theta$ exceed corresponding pruning thresholds) then
9.       Skip to next context shape in $CS_L$;
10. foreach (Pose $\pi$ of context shapes $CS_R$ and $CS_L$) do
11.     Calculate the overlap volume $OV$, under the given pose $\pi$;
12.     if ($OV$ exceeds threshold value) then
13.       Reject pose $\pi$;
14.     Calculate the buried surface area $BSA$, under the given pose $\pi$;
15.     Only keep the best pose $\pi$ with the largest buried surface area;
16.     Insert the tuple ($CS_R$, $CS_L$, $\pi$, $BSA$) in Candidate-Pairs;
17. Sort Candidate-Pairs based on $BSA$ (decreasing order);
Comparing ContextShapes to other docking methods.

For the problem of rigid body docking, ContextShapes does better than PatchDock or ZDOCK. ZDOCK suffers from coarse-graining and a crude energy function. ContextShapes is limited to rigid docking -- no flexibility. Both ZDOCK and PatchDock are faster.
ZDOCK
Fast Fourier Transform algorithm to enable an efficient global docking search on a 3D grid, and utilizes a combination of shape complementarity, electrostatics and statistical potential terms for scoring.

http://zdock.umassmed.edu/

A designed ligand
ZDOCK model #5
PatchDock

http://bioinfo3d.cs.tau.ac.il/PatchDock/patchdock.html

- Molecular Shape Representation - a) compute the molecular surface, b) apply a segmentation algorithm c) keep only patches with 'hot spot' residues.
- Surface Patch Matching - Geometric Hashing and Pose-Clustering to match the patches. Concave patches are matched with convex, flat patches with any type.
- Filtering and Scoring - a) discard all complexes with unacceptable penetrations, b) rank according to a geometric shape complementarity score.

Docking programs give different results

designed in MOE  
ZDOCK  
PatchDock
Receptor in solid surface. Best pose (black). Highest scoring pose (blue). True pose (red)
Lessons from ContextShapes

• Binary matching is fast
• Rigid docking only predicts correctly if the true docked structure is used. (which is cheating)
• Flexible docking can be approximated by “soft” inflexible docking. (i.e. by allowing some interpenetration of surfaces).
• Energy-based docking is better than descriptors.
• True flexible docking is the way to go, but slow.