18.417 Introduction to Computational Molecular Biology	
Lecture 21: November 23, 1999	Scribe: Joe Cohen
Lecturer: Amy Keating	Editor: Camilla Fox

Computational Tools for Drug Discovery

(Adapted from notes handed out in class.)

Introduction

The sale of pharmaceuticals resulting from research in drug discovery is a multi-billion dollar per year industry. A single big-selling drug (such as Lipitor, which is at the center of the recent controversial Pfizer/Warner-Lambert deal) can bring revenues of \$4-5 billion per year. Bringing such a drug to market, however, is a slow and expensive process that involves several major steps:

process	**description**	**years**	**candidates**
lead discovery	Identifying a molecule (here called a ligand) that can inhibit some process important for disease, even if only weakly.	1-2	
lead optimization	Finding versions of this molecule that bind more tightly to the target of the drug (here called the receptor).	1-2	10^4
assays	*In vitro* and *in vivo* testing of efficacy, including ability of the drug to reach its target	1-2	
toxicology	Testing if the drug harms living tissue.	1-2	10^2
human safety	Testing if the drug harms humans.	1	10
human efficacy	Testing if the drug works in humans and determining dose.	1-2	
TOTAL		6-12	1

Computational tools are mostly used for the first two stages (lead discovery and optimization) and fall into two classes:

1. methods that use information about known inhibitors (no structural data available)
2. methods that use information about the structure of the biological target (structure-based drug design)

A useful overview of both kinds of methods can be found in [1].

Tools that don't use structural information

Methods of type 1 try to identify features common to known inhibitors to use in screening for new or improved drugs. Two ways of doing this are briefly outlined here.

Generation of a pharmacophore A pharmacophore is a 3D arrangement of functionally-important parts of a molecule. Pharmacophores can be identified by searching the possible low-energy conformations of known inhibitors to find recurring spatial relationships between important groups.

The conformational search can be quite large (since lots of drug molecules have many degrees of freedom) and can be approached using

1. systematic search
2. distance geometry methods
3. clique-detection

Quantitative Structure-Activity Relationship (QSAR) A QSAR study tries to establish a link between the ability of a certain molecule to perform its desired function and properties of that molecule.

A typical example of such a relationship is given by the Hansch equation:

$$\log\left(\frac{1}{C}\right) = k_1 \log P + k_2(\log P)2 + k_3\sigma + k_4$$

C = concentration required to give a certain response (lower C = more effective compound)

P = measure of hydrophobicity of molecule (partition coefficient between octanol/water)

σ = Hammett parameter; used in organic chemistry to describe the electronic properties of molecules

This can also include terms to describe steric properties, molecular volume, etc.

Generally a least squares fit of the parameters to the inhibition data is performed to give the constants k_i. Neural networks or genetic algorithms can also be used, for non-linear dependence on k_i.

Structure-Based Drug Design

Based on the structure of the biological target (usually a protein) either bound or unbound to an inhibitor or substrate, an inhibitor can be designed or improved upon. This is much harder if the structure is of the unbound complex. There are two main approaches:

Docking Scan a database of known molecules for those likely to bind well to the receptor. Generate possible binding geometries and evaluate these using some scoring function.

De Novo Design Build a "custom" ligand for a given receptor.

Two types of de novo design are "inside-out" and "outside-in". In each of these, the idea is to:

1. identify points on the receptor where a ligand can form favorable interactions.
2. choose ligand groups of atoms which are complementary to these sites.
3. connect them into a molecular framework in an exhaustive, unbiased way.

Favorable interaction sites are frequently found by placing multi-atom fragments on a grid inside the binding site and evaluating their energy. Another method is to use molecular dynamics or Monte Carlo search methods to place fragments in a site.

For inside-out approaches an initial good interaction is chosen as a starting point, and the rest of the molecule is "grown" from this site, seqentially choosing fragments from a library and connecting them using rules. GrowMol is one example of a successful inside-out method.

For outside-in approaches groups of atoms from the ligand are placed in the binding site, and then a search is made for possible fragments which can join them in an appropriate geometry. LUDI and CAVEAT are examples of outside-in programs.

Docking vs. Design

The key advantage of docking is that the molecules being considered are taken from a database of molecules that already exist and can therefore be immediately used in an assay. Most companies have proprietary databases of molecules that they have on-hand. The advantage in De Novo design is that new classes of molecules can be discovered (in case there are not any good candidates in the known databases). However, these designed molecules might be very difficult to make.

The main challenge for both methods is the size of the search space that must be considered. There are many possible different inhibitors to choose from and each has many possible conformations in which it might bind to the receptor. Also, fast methods are needed since the goal is to screen many possibilities (a typical database size is 10^5, and this is a small portion of all possible designed inhibitors).

Success?

Examples of the design of successful inhibitors using these methods are still very rare (though there are now a few). In fact, even in cases where a ligand was successfully designed or predicted to be active, structural studies have often shown that the molecule binds to the target differently than predicted (oops!). Nevertheless, especially when coupled with experimental information in a feedback loop, structure-aided design can be a valuable part of the drug-discovery process.

Docking

Two approaches used for docking.

In docking there are three degrees of translational and three degrees of rotational freedom between the receptor and the ligand. As this is already a very large space to sample, most docking methods to date use a rigid body assumption in the docking process (that is, there is no conformational flexibility of the ligand or receptor considered). This is not a very good assumption, but it is made so that the problem will be somewhat tractable.

A docking method requires:

1. a representation of the receptor and ligand
2. a method for generating potential binding geometries
3. a way of scoring these geometries

The method may also involve some refinement of the initially generated conformations before or after scoring.

Method 1: the DOCK program (Kuntz, UCSF)

Molecular Representation

Both the ligand and the potential binding sites of the receptor are modeled using clusters of spheres. [2]

Begin by describing the surface of both ligand and receptor using a Connolly surface. This surface is generated by taking a spherical probe of some radius r (typically $r = 1.4 - 1.8$Å) and rolling it over the surface of the molecule. All points where the probe touches the molecule once are included on the surface, along with points which lie on the surface of the probe when the probe touches the molecule in more than 1 point. The Connolly surface basically smoothes out invaginations in the molecular surface. Rather than a smooth analytical surface, a grid of points on this surface is commonly used.

Given the Connolly surface grid, a series of spheres are generated such that:

1. each sphere touches the Connolly surface at two grid points with the sphere center on the normal from one of the points
2. spheres lie outside the surface of the receptor, inside the surface of the ligand

The spheres thus describe the ligand (key) and the negative image of the receptor (key hole). The ligand and any potential binding sites on the receptor will be defined in this way as clusters of overlapping spheres. In a perfectly-complementary complex, the centers of the ligand and receptor spheres will be coincident.

Generating receptor/ligand matches [3]

Size of space

For n receptor spheres and m ligand spheres (typically hundreds of each), there are $\frac{n!}{(n-m)!}$ ways to map the ligand points onto the receptor. However, it only takes a mapping of three points of the ligand onto the receptor to define a rigid body transformation. (A fourth point is needed to discriminate two possible reflections through the plane of the first three points). Thus, we need to consider roughly n^4m^4 possible configurations. This number can be further reduced by breaking spheres into pre-defined subgroups (e.g. corresponding to potential binding sites) and only allowing interactions where all points are within the cluster.

Acceptable matches

DOCK generates matches between receptor and ligand spheres such that internal distances (intra-ligand and intra-receptor) are equivalent, within a tolerance, between matched points.

GOAL: find a set of matched pairs $\{(r_i, l_i)\}$, r_i = receptor point; l_i = ligand point such that $\forall i, j \; |d(r_i, r_j) - d(l_i, l_j)| < \epsilon$ where $d(x, y) = |x - y|$ = distance between points x, y.

Matching using a bipartite graph

This method involves systematic generation of possible matches between ligand and receptor. The receptor and ligand are represented as separate graphs, and edges connect matched pairs of points. Correspondences between edges are found as follows by successive trial and backtracking (see Figure 1 of [3]).

LOOP 1 From the $n \times m$ possible pairs of (r_i, l_j) points choose one: (r_1, l_1). This is the seed edge. (Loop over all possible seed edges.)

LOOP 2 For each potential second edge (r_2, l_2) evaluate:

$$
\begin{aligned}
|d(r_1, r_2) - d(l_1, l_2)| &\geq \epsilon \rightarrow \text{reject match pair, back up and choose again} \\
&< \epsilon \rightarrow \text{accept match pair}
\end{aligned}
$$

LOOP 3 For each accepted (r_2, l_2) choose each possible (r_3, l_3) and evaluate:

$$|d(r_2, r_3) - d(l_2, l_3)| \geq \epsilon \text{ OR } |d(r_1, r_3) - d(l_1, l_3)| \geq \epsilon \rightarrow \text{reject, choose again}$$

$$|d(r_2, r_3) - d(l_2, l_3)| < \epsilon \text{ AND } |d(r_1, r_3) - d(l_1, l_3)| < \epsilon \rightarrow \text{accept match pair}$$

Continue until you have generated all possible sets of 3 matched pairs which give a unique transformation (except for the chirality which can be checked separately).

This is not particularly elegant or speedy. Heuristics can be used to speed it up, though they compromise the exhaustive search. One such heuristic is to use distance binning (pre-sort points by their distance from the points used in the seed edge) so that only potential edges compatible with the seed edge will be considered. See Figure 7 of [4] for a picture of how this works. A better way to avoid the backtracking required when bad matches are considered is to do...

Matching using a single docking graph

In the graph used for this search the receptor and ligand points are combined to give $n \times m$ nodes where each node represents a potential match pair (r_i, l_i). Edges exist between nodes with compatible distances.

Edge if $|d(r_i, r_j) - d(l_i, l_j)| < \epsilon$

The problem is then defined as finding all unique, completely connected subgraphs containing at least 3 nodes. This is related to clique detection but subgraphs need not be maximal so the problem is not as hard.

The graph is pre-computed as an Adjacency Matrix, A. A is an n^2m^2 Boolean matrix defined as:

$$
\begin{aligned}
A_{ij} &= 1 \text{ if } |d(r_i, r_j) - d(l_i, l_j)| < \epsilon \\
Aij &= 0 \text{ if } |d(r_i, r_j) - d(l_i, l_j)| \geq \epsilon
\end{aligned}
$$

The matrix A is sparse (ca. 1% occupied), so the whole matrix is not stored.

With prior knowledge (stored in A) of what pairs of points match, no backtracking need be done. The match generation proceeds to exhaustively generate all matches consistent with the distance criteria (see Figures 2 and 3 of [3]). The search time is proportional to the number of allowed solutions. This method is approximately $2\times$ faster than the previous method using heuristics (which was not exhaustive). It requires about 10 to 1,000 seconds per molecule depending on what tolerances are used.

After matches are generated:

1. The ligand is transformed according to the unique transformation specified by the match.
2. The match is checked for steric clashes of points other than match points.
3. The match may be scored using some potential function (e.g. might include some surface complementarity term, electrostatic term, polar/non-polar term, etc.).

Method 2: Geometric Hashing

(Fischer, Lin, Wolfson, Nussinov, Tel Aviv University and NCI Maryland)

Geometric Hashing is a method adopted from computer vision. In vision applications where the goal is to recognize objects in a scene, given a database of known objects, you want to be able to:

1. recognize translated/rotated objects.
2. recognize objects that are partially occluded or in the presence of non-database objects.
3. avoid having to retrieve every object in turn and compare it to every part of the scene.

There are similarities between this problem and the docking problem, where you want:

1. to recognize suitable ligands that are translated/rotated relative to the receptor binding pocket.

2. to recognize good matches between part of a ligand and part of the receptor.
3. to avoid having to retrieve every ligand in turn and compare it to every part of the receptor.

The idea is to store information about ligands in a way such that it is easy to retrieve the ones appropriate for binding to a given receptor

Molecular Representation

The representation used by Fischer *et al.* is different than that used in DOCK. The goal is to have an accurate, sparse, uniquely-defined description of the molecular surface. [5]

Again start with the Connolly surface. The Connolly surface can be broken up into three different kinds of faces:

1. convex, where a probe touches the surface once
2. concave - where a probe touches the surface three times
3. saddle-like - where probe touches surface twice

The idea is to choose one point on the surface per face, which will be called a critical point. The three types of critical points are called caps (for convex faces), pits (for concave faces) and belts (for saddle-like faces). Each critical point has associated with it a surface normal and a size, which reflects how much of the surface it covers.

The main advantages of this representation are that

1. the shape of the whole surface is described
2. there are only a few points per atom (typically six, but this is often pruned to less than 1 by various considerations)
3. the location of the points is independent of the density of points chosen (not true for Connolly surface, for example)

Generating receptor/ligand matches [6]

This occurs in two phases.

Preprocessing of ligand or ligand database

This only needs to be done once, as follows:

1. For every pair of *cap* critical points on the ligand, cp_{l1} and cp_{l2}, generate a unique reference frame using the two points and the average of their associated normals (given m ligand points there will be m^2 such reference frames).
2. In each reference frame in turn, for every other cap critical point, cp_{li}, that meets some distance criterion, record the coordinates of this point in the chosen frame (gives (x,y,z)).
3. Build a hash table to hold this information where the address is the coordinates of the critical point and the entry is: (the identity of the critical point, the ligand reference frame in which it is described). The table will have $m^2 \times m = m^3$ entries.

Example:

Address	Entry
(x_a, y_a, z_a)	(cp_1, cp_{l1}, cp_{l2})
(x_b, y_b, z_b)	(cp_2, cp_{l1}, cp_{l2})
(x_c, y_c, z_c)	(cp_3, cp_{l1}, cp_{l2})
$\vdots$	$\vdots$
$(x_{a'}, y_{a'}, z_{a'})$	(cp_1, cp_{l2}, cp_{l3})
$(x_{b'}, y_{b'}, z_{b'})$	(cp_2, cp_{l2}, cp_{l3})
$\vdots$	$\vdots$
$(x_{a''}, y_{a''}, z_{a''})$	(cp_1, cp_{li}, cp_{lj})
$\vdots$	$\vdots$

Notice that each ligand critical point will occur many times in this table, indexed by its coordinates in many different reference frames.

What this is doing is storing information about triples of points found in the ligand in the hash table. The idea is that when we find a receptor that needs three ligand atoms in a certain orientation, we will know where to look.

Recognizing Matches

Process the receptor in a way similar to the ligand:

1. For every pair of *pit* critical points on the receptor, cp_{r1} and cp_{r2}, generate a unique reference frame using the 2 points and the average of their associated normals.

2. In each reference frame in turn, for every other pit critical point, cp_{ri}, that meets some distance criterion, note the coordinates of this point in the chosen frame (gives (x, y, z)).

3. Look in the hash table under the (x, y, z) entry and (after checking a few geometry requirements) cast a vote for the ligand reference frame that is found there.

4. For each receptor frame, tally the votes from different ligand frames and use the best as potential matches.

For a given receptor reference frame, every ligand reference frame that got K votes comes from a ligand that has at least $K+2$ critical points that can be superimposed on the receptor. The two reference frames (ligand and receptor) can be used to compute the transformation which accomplishes the superposition.

After potential matches are found the ligand is transformed, and the match is checked for steric clashes and can be scored, etc., as described for DOCK above.

Advantages of Geometric Hashing

1. pre-processing of database of ligands is done only once

2. relatively efficient lookup $O(m^3)$ for well-balanced hash table

3. inherently parallel

A disadvantage (perhaps) is the possibility of collisions in the hash table (more than one entry with same address slows down lookup). This could potentially be a problem since indices describe interatomic distances which are not really random.

Geometric Hashing and DOCK both have the potential to include requirements of chemical complementarity between matched points. They are also both translation/rotation invariant, and do pretty well at reproducing the correct geometry of known complexes when starting with the X-ray structure of the components. However, they are bad at picking the best solution out of the others generated, and they are bad when they do not have the correct structures to dock together. Since conformational changes upon binding are common, this is a big problem.

The main challenge for docking methods in general these days is to incorporate some amount of flexibility into the ligand and/or receptor. Methods examined for receptor flexibility include allowing pre-determined hinge motions, performing sidechain repacking calculations (using DEE and A*) during docking or docking to an "average" structure that is derived from several known conformations of the receptor. The most obvious way to include ligand flexibility is to consider explicitly multiple conformations of each molecule being docked. Other possibilities include using some amount of molecular dynamics or Monte Carlo search in the procedure, or using analogies to robotics to describe articulated motions of docking ligands.

[1] Leach, A. R. Molecular Modelling Principles and Applications, Addison Wesley Longman (1996).

[2] Kuntz *et al. J. Mol. Bio. 161* (1982) 269-288.

[3] Ewing Kuntz, *J. Comp. Chem. 18* (1997) 1175-1189.

[4] Shoichet, Bodian Kuntz, *J. Comp. Chem. 13* (1992) 380-397.

[5] Lin *et al. Proteins 18* (1994) 94-101.

[6] Fischer *et al. J. Mol. Bio. 248* (1995) 459-477.