Algorithmic Foundations of Robotics (Proc. of WAFR 96), J.-P. Laumond and M. Overmars Eds., A K Peters, 1996

Geometry and the Discovery of New Ligands

Lydia E. Kavraki *Stanford University, Stanford, CA 94305, USA*

Computer-aided drug design is a significant component of the rational approach to pharmaceutical drug design. Chemists now consider the geometric and chemical characteristics of molecules early in the design process in an effort to quickly identify ligands that have good chances of becoming potent pharmaceutical drugs. Computer assistance is not only helpful but also necessary to narrow down the search for potential ligands. Depending on the level of accuracy desired to model drug action, detailed quantum mechanical methods or approximate molecular mechanics methods are used. Even when simple approximations are made, efficient approaches are needed to compute, among other things, molecular surfaces and molecular volume, models of receptor active sites, reasonable dockings of ligands inside protein cavities, and geometric invariants among different ligands that exhibit similar activity. This paper surveys several problems and approaches in the area of computer-aided pharmaceutical drug design and draws analogies with problems from robotics and computational geometry.

1 Introduction

The design of pharmaceutical drugs is an extremely complex and still not completely understood process [2]. Computational chemists combine their knowledge of molecular interactions and drug activity together with visualization techniques, detailed energy calculations, geometric considerations, and data filtered out of huge databases, in an effort to narrow down the search for potent pharmaceutical drugs. Computer-aided drug design is a significant component of rational drug design [6], and is becoming more relevant as the understanding of molecular activity improves and the amount of available experimental data that requires processing increases.

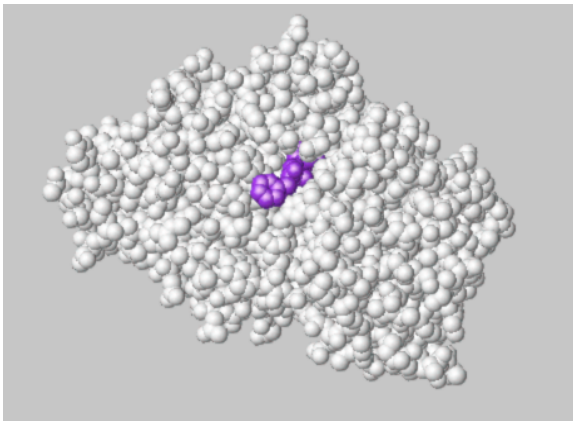

Figure 1: *The protease thermolysin with one of its known inhibitors (1TMN)*

A fundamental assumption for rational drug design is that drug activity, or pharmacophoric activity, is obtained through the molecular recognition and binding of one molecule (the ligand) to the pocket of another, usually larger, molecule (the receptor). In their active, or binding, conformations, the molecules exhibit geometric and chemical complementarity, both of which are essential for successful drug activity [2, 49]. There is no simple way to explain how drugs achieve their desired effects. It is known, however, that several pharmaceutical drugs are inhibitors, i.e. inhibit reactions that would take place without their presence. For example, if a cavity of a molecule provides a favorable environment for a reaction, a ligand that fills that cavity in an energetically stable conformation can prevent this reaction from happening. Figure 1 shows the protease thermolysin and one of its inhibitors. Thermolysin is the large molecule shown in the picture, while the inhibitor (1TMN) is drawn in a darker color near the center of that picture. The 3D structure of the complex has been obtained by X-ray crystallography and can be retrieved from the Brookhaven protein data bank.

The modeling of molecular structure is a complex task. Quantum mechanics provide a detailed description of molecules in terms of atomic nuclei and electron distribution among them. However, quantum mechanical calculations cannot be used to treat large systems because of high computational demands. The modeling of the binding process is also a difficult task. The characteristics of the receptor, the ligand, and the solvent in which these are found have to be taken into account. Although chemists strive to obtain models that are as accurate as possible, several approximations have to be made in practice. Molecules are thus visualized to have surfaces and volume similar to our perception of surfaces and volume of macroscopic objects, or are considered under ideal conditions (i.e. in vacuum). It is clear that the more accurate the model used, the better the chances chemists stand in predicting molecular interactions. Nevertheless, there is a large number of predictions made with approximate models which have been confirmed with experimental observations [6, 34]. This has encouraged researchers to build tools that use approximate models and investigate the extent to which these tools can be useful. More accurate molecular modeling, gained through better understanding of drug activity or increased computational power, can only improve the techniques developed with simpler models.

Depending on whether the chemical and geometric structure of the receptor is known or not, the problems arising can be classified in two broad categories. If the receptor is known, chemists are interested in finding if a ligand can be placed inside the binding pocket of the receptor in a conformation that results in low energy for the complex. This problem is referred to as the *docking problem*. It has several variations: an accurate description of the binding may be desired, or an approximate estimate of which ligands from a huge database are likely to fit inside a receptor may be sought. Very often the binding pocket is not known. In fact, the 3D structure of few large molecules (or macromolecules) has been determined by X-ray crystallography or NMR techniques. When the receptor is not known, what is usually known is a number of ligands that interact with that specific receptor. These ligands have been discovered mainly by experiments. Using the geometric structure and the chemical characteristics of these molecules, chemists attempt to infer information about the receptor. In particular, chemists are interested in identifying the *pharmacophore* present in these ligands. The pharmacophore is a set of features at a specific 3D arrangement contained in all the active conformations of the considered molecules. A prevailing hypothesis is that the pharmacophore is the part of the molecule that is responsible for any observed drug activity, while the rest of the molecule is a scaffold for the pharmacophore's features. If the pharmacophore is isolated, chemists can use it to design a more potent pharmaceutical drug by examining the different activities, relative shapes, and chemical structures of the initial molecules [27].

The techniques that have been used so far in computer-aided drug design include geometric calculations (surface computation), numerical methods (energy minimization), randomized algorithms (conformational search), and a variety of other techniques like genetic algorithms and simulated annealing (docking). The machines used for these calculations range from desktop workstations to supercomputers. It is only recently that chemists have tools for complex geometric and energy calculations and the success of these computer-aided methods is currently being evaluated [2, 6].

This paper describes some of the computational problems arising in rational drug design. It surveys recent work on surface and volume calculations, conformational search, docking, pharmacophore generation, and database searching. The discussion reveals the wealth and diversity of the problems that arise in the domain of computer-aided pharmaceutical drug design. Analogies with problems from robotics and computational geometry are also drawn.

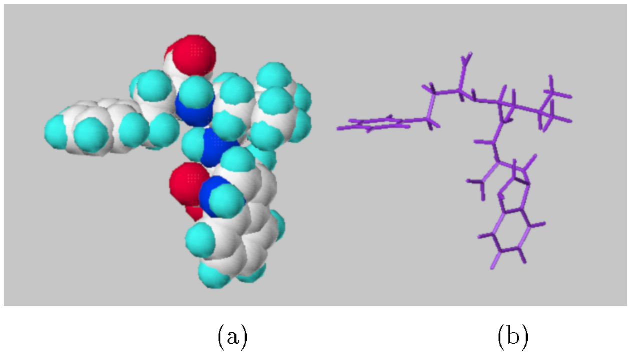

Figure 2: *The hard-sphere model and stick diagram of 1TMN*

2 Molecular Modeling

The *hard-sphere* model of 1TMN, the inhibitor of thermolysin of Figure 1, is drawn in Figure 2(a). This model is an abstraction frequently used by chemists to approximate the volume of a molecule. A sphere is drawn around the center of every atom of the molecule. The radius of each sphere reflects the space requirements of the corresponding atom and has been determined by a combination of experimental observations and quantum mechanical calculations. A set of radii that are commonly used are the *van der Waals radii* [7]. If the van der Waals radii are used, the envelope surface of the hard-sphere model is called the *van der Waals surface*.

The *stick diagram* of a molecule (Figure 2(b)) draws a line segment for each chemical bond. The angle between two consecutive bonds is called the *bond angle* and the angle formed by the first and the third of three consecutive bonds, when one looks along the axis of the second bond, is called the *dihedral* or *torsional angle*.

A priori, all bond lengths, bond angles, and torsional angles are degrees of freedom (DOF) of the molecule. Because of their chemical characteristics certain bonds cannot rotate about themselves and, as a result, all the torsions in which they participate as middle bonds are fixed. Bond lengths and bond angles tend not to exhibit large variations in their values. It is fairly common to consider bond lengths and bond angles constant in calculations [28, 34]. Torsional angles, however, vary significantly and this affects the 3D shape of the molecule. When bond lengths and bond angles are considered fixed and only torsions vary, a molecular chain with n torsions can be viewed as an articulated mechanism with n revolute joints.

Standard geometries are commonly used to construct reasonable models of molecules. For example, there exist tables that show “preferred values” for bond lengths and these depend on the kind of atoms participating in these bonds [7]. Preferred values have also been calculated for bond angles and torsional angles, and again depend on the types of atoms linked by the corresponding bonds. The exact values used are obtained from statistical analysis of structural data in X-ray databases, like the Brookhaven or the Cambridge databases. Although it is true that there is variability in the geometric data in these depositories, the information gathered provides a reasonable approximation of reality [7, 34].

As far as calculations of energy are concerned, empirical force fields are used in practice instead of more detailed methods like quantum mechanics. A typical empirical force field includes terms for bond-stretch, bond-angle and torsional-angle deformations, and terms for van der Waals and Coulomb potentials [52]. Frequently, terms that model solvation effects are also included. Interaction of the molecule with the solvent in which it is dissolved is very important but also difficult to model accurately [7, 34]. An example of how the energy of conformation $\mathbf{c}$ can be calculated with empirical force fields when the molecule is considered in vacuum is given below:

$$\begin{aligned} \mathrm{E}(\mathbf{c}) = \; & \textstyle\sum_{bonds} \frac{1}{2} K_b (R - R_0)^2 + \\ & \textstyle\sum_{angles} \frac{1}{2} K_a (\theta - \theta_0)^2 + \\ & \textstyle\sum_{torsions} K_d [1 + \cos(n\phi - \gamma)] + \\ & \textstyle\sum_{i,j} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \right\}. \end{aligned}$$

In the above K_b, K_a, and K_d are force constants, ϵ is the dielectric constant, and n is a periodicity constant. R, θ, and ϕ are the measured values of the bond lengths, bond angles, and torsional angles in conformation $\mathbf{c}$, while R_0, θ_0, and γ are equilibrium (or preferred) values for these bond lengths, bond angles, and torsional angles. r_{ij} measures the distance of atom centers in $\mathbf{c}$. The parameters $\sigma_{ij}, \epsilon_{ij}$, and q_i are the Lennard-Jones radii, well depth, and partial charge for each atom in the system. All parameters and constants above are derived by a combination of quantum mechanics, vibrational methods, and experimental data. Once the values of bond lengths, bond angles and torsional angles of a conformation are known, obtaining the energy of a molecule with an empirical force field is a straightforward task. Minimization of this energy is not easy however, since force fields are non-linear functions and may contain a large number of local minima.

Calculations of energy are very important in the molecular world. In nature, molecules are usually found in low-energy conformations. Protein-ligand complexes are stable when the binding energy of the system is low. It should be emphasized that the exact calculation of molecular and binding energies is by no means a simple task [52], and that empirical force fields offer only an approximation. Nevertheless, as noted above, there are several cases where reasoning with these approximations has produced meaningful results [5, 7, 34].

Before describing specific problems let us also define the concept of molecular *features*. Chemists group atoms according to their chemical characteristics and use a label to refer to these groups. Given a molecule there are rules that identify the hydrophilic and hydrophobic parts of that molecule, the hydrogen-bond donors and acceptors, the charged centers, etc. These features are used, for example, to define pharmacophores, or to specify database queries that will retrieve ligands with certain characteristics. The accurate definition of features is a difficult task for the chemist, but is out of the scope of this paper [17, 32].

3 Molecular Surfaces & Volume

Computing surfaces and volume of molecules analogous to our perception of macroscopic surfaces and volume has attracted considerable attention [58]. This information is useful in calculations for molecular recognition and docking [12], or computations of the energy of a molecule in solution [24]. Surface computation is also useful in pharmacophore identification, since atoms that are buried or little exposed are not likely to participate in a pharmacophore.

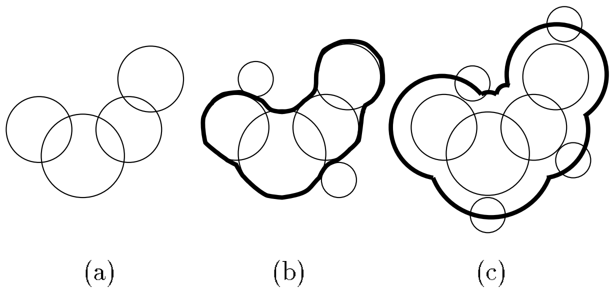

Figure 3: *(a) van der Waals, (b) molecular, and (c) solvent accessible surface*

3.1 Types of Surfaces

Surfaces that are of interest to chemists include *the van der Waals* surface (defined in Section 2), the *molecular* surface and the *solvent accessible* surface [48, 66]. Figure 3 illustrates these different surfaces.

The molecular and the solvent accessible surfaces are defined with the help of a solvent atom which is a sphere of radius r. In particular, the molecular surface is defined by the front of the solvent sphere when this is rolling around the van der Waals surface. The solvent accessible surface is defined by the center of the solvent when this is rolling around the van der Waals surface of the molecule. In other words, the solvent accessible surface is the boundary of the free placements of the center of the sphere of the solvent, when this is moving among the atom spheres of the molecule. It can thus be computed, using configuration space techniques from robotics, as the union of the Minkowski sums of each molecular atom sphere and the sphere of the solvent [36].

3.2 Methods

Approximation and analytical methods have been used for the computation of surfaces of molecules. A survey of early techniques is given in [58]. Two widely used methods are Richards' method [65], which approximates the volume of a molecule by polyhedra, and Connolly's approach [11], which analytically computes molecular surface patches. More recently, methods from computational geometry are being employed to efficiently compute surfaces and volume without the limitations of previous approaches [19, 36, 50, 70].

Halperin and Overmars [36] observed that the complexity of the arrangement defined by n atomic spheres of a molecule is $\Theta(n)$, as opposed to $O(n^3)$ for a general arrangement of spheres in space. The complexity of an arrangement is defined as the overall number of cells in that arrangement. In the same paper it is shown that the arrangement of atomic spheres can be decomposed into an arrangement of simple cells whose total complexity is $O(n)$. As a result, it is possible to construct a hashing data structure that uses $O(n)$ space and can answer intersection queries for spheres of comparable radii to the atomic spheres in constant time. Computation of surfaces and volume follows nicely from this data structure. In particular, the van der Waals surface of a molecule can be constructed in $O(n \log n)$ time. Similar results can be obtained for the solvent accessible and the molecular surface.

Edelsbrunner uses alpha shape theory to accurately compute the surface and volume of molecules [18]. The alpha shape is the space occupied by the simplices of an alpha complex. These simplices are constructed in such a way that they are always a subset of the simplices defined by the weighted Delaunay triangulation of the molecule. In the above model, α is a parameter that regulates the radius $d = \sqrt{w^2 + \alpha}$ of atomic spheres, where w denotes the van der Waals sphere of an atom. If α is increased from its least possible value (a negative value) to zero, the shape of a molecule grows from a set of points to its van der Waals shape. Appropriate simplices are maintained as α changes, and when $\alpha = 0$ the set of constructed simplices, the alpha complex, contains important information about atom intersections and the topology of the molecule. The alpha complex can be computed in $O(n \log n)$ time and then it is possible to quickly identify the atoms on the surface of the molecule, and compute the van der Waals, molecular, and solvent accessible surfaces. The volume of the alpha complex can be combined with the volume of the surface atoms to compute the volume of the molecule. Furthermore, the topological structure of the apha complex permits the identification of voids and canyons in the molecule [19, 20, 50]. Alpha shapes have also been used by Varshney el al [70] for molecular modeling. This work has produced a parallelizable algorithm that scales linearly with the number of atoms in a molecule for computing molecular surfaces.

3.3 Dynamic Maintenance

Although algorithms that compute molecular surfaces have been widely investigated, little has been done for their dynamic maintenance. For calculation of binding energies, to give an example, it is interesting to know how the surface that one particular atom contributes to the outer van der Waals surface changes, as the shape of the molecule changes. Work on dynamic data structures is useful in this respect [35].

4 Conformational Search

Conformational search is a fundamental problem in molecular biology. Perhaps the most well-known conformational search problem is the protein folding problem. It is believed that proteins have "unique" 3D shapes which correspond to global minima of their total energy and which are specified only by the chemical composition of the molecules. Finding these conformations is by no means an easy task and involves several hundreds of DOF [16].

For small ligands, finding the conformation with the minimum energy is of little interest. What is interesting is to find a set of conformations whose energy is below a threshold and which are geometrically distinct [46]. Such conformations are used in docking [59] and pharmacophore identification [55]. Low-energy conformations of a molecule that also respect certain "distance constraints" (i.e. have certain features at specific positions in 3D space) are also of interest to computational chemists. Tools that can produce such conformations have applications in database screening [56].

Several approximations are made during conformational search depending on the level of detail required. For example, it is usual to consider bond lengths and bond angles almost fixed, choose torsional angle values from predefined distributions, and simplify the energy model considered [28, 34]. Frequently the molecule is assumed to be in vacuum. An external potential can be considered with most conformational search methods but may result in longer computation times. Depending on whether distance constraints are imposed when conformational search is performed, we distinguish conformational methods into unconstrained and constrained techniques.

4.1 Unconstrained Search

A wide variety of methods for searching conformational space have been described in recent years (for a survey see [46]). Systematic search methods sample each torsional DOF of the ligand at regularly spaced intervals and were among the first to be developed and used [51]. The discretization of the torsional values is typically as coarse as 30^o or 60^o [46]. Even with such a resolution the number of conformations that are generated with systematic search can be very large. Typically the energy of all generated conformations is minimized which is an expensive operation. Several heuristics have been used to quickly prune down conformations that are close to previously generated conformations [68] in an effort to enhance the diversity of the sample.

A variety of randomized methods are also under investigation: conformations are obtained by applying random increments to the torsional DOF of the molecule starting from a user-specified initial conformation [26], or from a previously found low-energy conformation [9]. Recent articles, which attempt to compare different methods, emphasize the superior quality of the results obtained with stochastic methods [26].

Randomized techniques have been proven useful for high-dimensional search problems [38] and this direction deserves to be further explored in the context of conformational search. A random sampling method for exploring the conformation space of small molecules has been recently developed in [22]. This method borrows ideas from randomized techniques for planning in high-dimensional configuration spaces [38]. The approach is divided into three steps: generation of random conformations, minimization of these conformations, and grouping or clustering of the minimized conformations. Initially, a large number of conformations, frequently tens of thousands, are generated at random over the conformational space of the molecule. The generation of these conformations is done by selecting each torsional DOF of the molecule uniformly from its allowed range. The selection can also be done according to a distribution that reflects preferred values for each torsional DOF, if such information is available. The resulting structure is stored only if it avoids intersections of the spheres of non-bonded atoms. Subsequently an efficient minimizer is used to obtain a conformation at a low energy minimum. At this step only conformations below a user-defined energy threshold are retained. Experimental observations have shown that the number of these conformations can be very large. Since only conformations that are geometrically distinct are interesting, it is necessary to partition the low-energy conformations into clusters of similar conformations. Figure 4 shows two of the clusters obtained

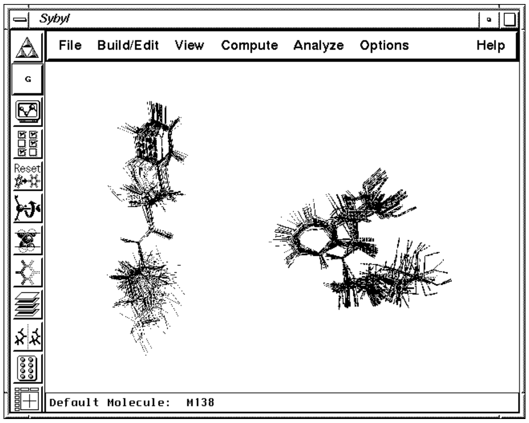

Figure 4: *Two clusters of 1TMN*

with the randomized approach of [22] for 1TMN. At the end of the clustering step, a representative per cluster can be retained.

Conformational search raises a number of interesting issues. Many open questions remain on what are good minimization techniques for the energy models that are available for small molecules, what are reasonable similarity measures for conformations, and how partitioning can be done efficiently. Improvements in each of these domains can affect the performance of conformational search software and the quality of solutions obtained for the problems where these conformations are actually used (i.e. docking and pharmacophore identification).

4.2 Constrained Search

Most of the techniques described in the previous section will produce poor results when distance constraints are imposed in the structure of the molecule. Distance constraints arise frequently in practice. For example, chemists may be interested in conformations that keep two atoms of the molecule at specific positions in space because these two atoms belong to a pharmacophore. Ring structures impose distance constraints by their own nature: maintaining ring closure when a torsional angle in the ring changes, requires the atom at the beginning and the atom at the end of the chain to be at a bond's length distance from each other.

The constrained conformational search problem has a direct analog in robotics, namely the problem of *inverse kinematics*. If the bond lengths and bond angles in a single molecular chain are considered fixed, then the chain can be viewed as a serial manipulator with revolute joints (these joints correspond to the torsional DOF of the chain).

Manocha et al [54] exploit the work done in robotics for computing inverse kinematics of manipulators to find valid conformations for small molecular chains. In particular, the case of a serial manipulator with 6 revolute DOF has been extensively studied (6 is the minimum number of DOF for a robot to be able to span a full-rank subset of $SE(3)$ [13]). Symbolic manipulation of the equations of Raghavan and Roth [63] transforms the inverse kinematics problem into one of computing the eigenvalues and eigenvectors of a matrix, which in turn can be done efficiently [53]. In a similar way, the inverse kinematics of a serial molecular chain with 6 torsional DOF can be computed by finding the eigenvalues and eigenvectors of appropriately defined matrices. For chains with $n > 6$ torsions, 6 torsions are considered free while the rest $n - 6$ are assigned discrete values and this procedure is repeated for different values of the $n - 6$ "fixed" torsions. The techniques in [54] are very efficient when computing conformations that maintain ring closure and for local deformations of small protein chains. It is worth mentioning that algebraic equations in 6 unknowns were also derived in [28] for finding the permissible conformations of a single-loop molecule when only 6 torsional angles are considered free, and solutions in limited cases were obtained.

Other kinds of methods, like distance geometry [14], are also being tested for constrained conformational search problems. Distance geometry exploits the fact that lower and upper values on interatomic distances can be derived from the restriction that atoms belong to a 3D chemical structure. These distances are used to refine 3D models of molecules by a variety of constraint propagation and "bounds smoothing" techniques. Distance geometry can also deal with large scale constrained conformational search problems like the ones arising from NMR data [14]. NMR produces distances between atoms of a macromolecule and chemists seek to reconstruct the 3D conformation of the molecule that produced these distances. The drawback of distance geometry methods is that they may fail to converge to a solution and can be relatively slow in practice [54].

Note finally that constrained optimization techniques, which minimize the energy of a conformation while observing distance constraints, can be used to obtain more stable conformations starting with conformations produced by algebraic or distance geometry techniques. The speed of constrained conformational procedures is crucial if these procedures are used to screen large databases [46].

5 Receptor is Known: Docking

Surface calculations and results of conformational search are used when trying to find a "reasonable" docked position of a ligand inside a known receptor. Information about the geometry of the receptor is obtained by X-ray crystallography or NMR techniques. For docking, it is generally assumed that the receptor molecule is rigid [47]. This approximation is justified by experimental data i.e. crystals of the molecule with and without the ligand, but exceptions have also been noted [60]. For the ligand however, it is essential to address its flexibility.

A central question for the docking problem is how to represent the geometry of the cavity, and how to compare it to the geometry of the ligand. The computation of the binding energy of the complex is a very important issue to be addressed in docking, but his problem is out of the scope of the present paper.

5.1 Rigid Ligand

If the ligand is considered rigid, it is possible to systematically search its six-dimensional configuration space for possible placements inside the binding pocket, but such a process can be time consuming. Several recent methods adopt a different approach: they try to match points (features) of the binding pocket to points (features) of the ligand. The points inside the pocket are referred to as "hot spots" [30], "essential points" [59], or "matchprobes" [71].

The definition of matching points in the receptor and the ligand varies widely with the method used. Some approaches use energy calculations to define these points. They describe, for example, the chemical environment of the pocket using a 3D grid, and define matching points as energetically favorable sites for certain functional groups [30, 59]. When the ligand molecule is placed in the grid region, the interaction energy can be efficiently calculated using precomputed data. Other docking approaches use only the geometry of the receptor and the ligand to define matching points. DOCK [43, 67], one of the earliest methods for docking, generates spheres inside the binding site in a way that they touch the surface of the pocket in two points and have their centers along the surface normal at one of these points. The centers of these typically overlapping spheres are the receptor's matching points. Spheres are created in a similar way inside the ligand and their centers are the matching points of the ligand. The description of the binding pocket by the spheres described above is not unique and may seem arbitrary, but several successful predictions have been reported [67].

After essential points have been identified in the pocket and the ligand, the docking problem reduces to a matching problem. All possible combinations of ligand-receptor points can be tried if their number is small [45]. Simple heuristics can be used to narrow the search. DOCK, for example, selects a pair of receptor points and measures their distance. Then a pair of ligand points that are at approximately the same distance with the receptor points is found. A third receptor point is chosen its distances with the previously selected receptor points are used to identify a third point of the ligand. This process continues until a specified number of pairs is found or until no possible matches can be found. In that case the algorithm backtracks. At least four points are necessary to define an unambiguous orientation of a ligand inside a receptor. Other approaches [42] build a "docking graph" using the receptor and ligand matching points. The graph has a node for all pairs of receptor-ligand points, and an edge between two nodes, if the pairs corresponding to the nodes can be matched at the same time. A maximal clique in this graph will produce a maximal matching between the receptor and the ligand. It is well known that this problem is NP-hard [25] but the method is reported to work well in practice [42].

The matching problem that arises in docking, has analogies with the geometric matching performed for model-based shape recognition [21]. These analogies are extensively discussed in [62]. In geometric matching, a 3D model of an object is known. Given a set of 3D points which may lie on the surface of that object, a rigid transformation is sought to align these points to the model. In the context of molecular docking the ligand provides the model, and the receptor provides the set of 3D points that are checked against the model.

Techniques developed for model-based recognition, like interpretation trees [33] or geometric hashing [44], are thus applicable to the docking problem. In fact, geometric hashing has already been used for molecular docking in protein-ligand and protein-protein studies [1, 61]. In geometric hashing, a hash table for the ligand is computed and this is a transformation invariant representation of the molecule. Given a set of points in the receptor, matches can be detected through a voting scheme. An advantage of this approach is that the hash table for the ligand can be computed off-line, and after that it is possible to dock the ligand to multiple receptors.

5.2 Flexible Ligand

To address conformational flexibility, a widely used approach has been to consider different low-energy conformations of the ligand. These conformations, which are frequently obtained by a conformational search procedure, are tried against the receptor cavity using a technique developed for docking a rigid ligand to a rigid receptor [59]. To facilitate such docking approaches, several molecular databases now store a set of geometrically distinct conformations per ligand [39]. It is clear that if the active conformation is not one of the conformations considered, these methods will fail to produce the optimal docking.

Conformational flexibility has also been addressed directly by simulated annealing techniques. In that case, the torsional DOF of the molecule are changed inside the receptor's cavity [31]. One could also imagine using randomized sampling techniques instead of simulated annealing to find low-energy conformations of the ligand inside the binding pocket. Matching points defined inside the binding pocket are again useful when flexible ligands are considered. In this case however, fragments of the ligand are docked independently and the fragments are later joined into conformations which are in turn refined and ranked with appropriate scoring functions [15, 64, 71]. The idea of "building" a ligand inside a binding pocket is also popular with methods that suggest unsynthesized compounds or add functionality to a known inhibitor [40].

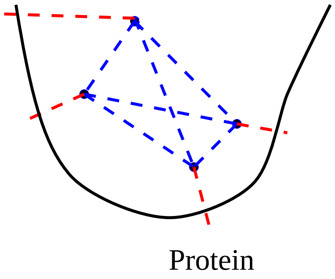

Figure 5: *The features of the pharmacophore interact with features of the receptor cavity*

Allowing for ligand flexibility is a challenging and still unsolved problem in protein-ligand docking. Efficient geometric techniques that can exclude placements of fragments that are in collision with the rest of pocket, or can suggest reasonable placements for these fragments may help prune the number of placements that are subjected to rigorous energy calculations. Researchers have also stressed the need for more accurate scoring functions for characterizing the energy of the binding. The development of such functions remains a difficult and poorly understood problem [5].

6 Receptor is Unknown: Pharmacophores

When the 3D structure of the target macromolecule is not known, the identification of a pharmacophore is key to the development of new pharmaceutical drugs [55]. A prevailing assumption in rational drug design is that if different ligands exhibit similar activity with a receptor, this activity is due largely to the interaction of the features of the pharmacophore to "complementary" features of the receptor (see Figure 5). Thus, if a pharmacophore has been isolated, chemists can use it as a template to build more potent drugs [27]. Given 5-10 ligands that are very flexible, finding a set of features that is present in the same 3D arrangement in the active conformation of these ligands is by no means a simple task. Figure 6 shows 4 different inhibitors of the protease thermolysin which was drawn in Figure 1. These ligands have 5 to 11 torsional DOF and each of these molecules can assume a large number of distinct low-energy conformations when these torsions are

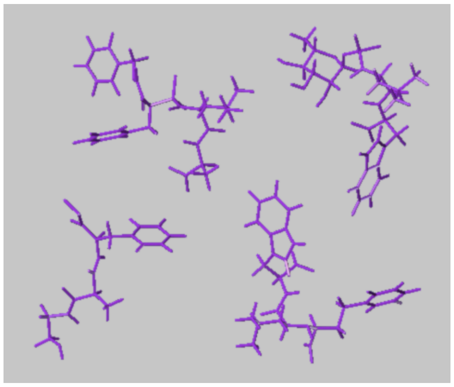

Figure 6: *Four different inhibitors of thermolysin*

varied [22, 46].

Initial approaches to pharmacophore identification searched simultaneously, in a systematic way, the conformational space of all molecules [57]. These approaches are now abandoned due to their prohibitive computational cost. The most popular recent algorithms start with a collection of distinct low-energy conformations per molecule, obtained by a conformational search procedure. They search for an invariant present in at least one conformation of most of the given molecules. Requiring that the invariant be present in all molecules may unnecessarily exclude solutions, since conformational search methods do not guarantee that all distinct low-energy conformations have been produced.

DISCO [55], one of the most popular algorithms for pharmacophore identification, uses clique detection to identify invariants. Initially the program considers a pair of conformations c_1 and c_2 belonging to different molecules. A "correspondence graph" G is constructed and this graph is similar to the "docking graph" described in Section 5.1. The nodes of G are again all node pairs of c_1 and c_2. An edge in G is created if the pairs in each of the connected nodes can be matched simultaneously. The Bron-Kerbosch clique detection algorithm [8] is then used to find cliques in G. These correspond to invariants in c_1 and c_2 and thus to candidate pharmacophores. The algorithm seems to work well in practice [55, 72]. Generalization of the above approach to n conformations is straightforward by considering one of the conformations as a reference and comparing it with all other $n-1$ conformations. Common parts of all pairwise invariants need to be computed in the end.

If a large number of conformations per molecule are considered, there can be a combinatorial explosion in the number of basic operations performed by algorithms like DISCO [3]. This is the main reason why different approaches are under development. One idea is to start with small invariants (2-3 features) and gradually expand them [3]. Another idea is to use randomized techniques when searching for invariants. When conformations c_1 and c_2 are compared in [22], a randomized sampling scheme is used to select atoms (features) in conformation c_1, and a hashing structure is built to find possible matchings of these atoms (features) in c_2. This process is repeated for all pairs of conformations of two molecules and produces several invariants. It is then checked if these invariants are present in the rest of the considered molecules with an elaborate hashing scheme.

7 Database Searching

Searching databases of 3D chemical structures for ligands with specific characteristics is becoming a basic tool in rational drug design [56, 72]. Although, it is fairly simple to do an initial screening of a database with one million compounds, it is difficult to narrow down the results at later stages [72]. Ligand flexibility can increase dramatically the number of cases that need to be examined before it is decided that a molecule does not match a query.

Queries in current database systems are usually specified by a 3D graph whose nodes correspond to specific features and whose edges correspond to diatomic distances. Formulating a query in this way is consistent with the definition of a pharmacophore. To find ligands with a specific pharmacophoric pattern in a database, a combination of the techniques described in this paper can be used. The efficiency requirements for these techniques are however increased considerably. For example, algorithms developed for surface computation or conformational search may need to be reevaluated in the context of database queries: it may be possible to find if a feature is on the surface of a conformation without computing the whole surface, or to produce a conformation which is very different from a given one without performing a large scale conformational search.

Many database queries result in a constrained conformational search problem which is currently poorly addressed [72]. Distance geometry, systematic/randomized search, and genetic algorithms have been tried but have produced slow algorithms [4, 9, 10, 23]. One of the most efficient existing techniques for flexible searching is the "Directed Tweak Method" [37, 69]. The method minimizes a pseudoenergy function which combines the energy of the molecule and the sum of the squares of the deviations of the distances found in the molecular structure to the distances expressed in the database query. Unfortunately the pseudoenergy function contains a large number of local minima and conformations having high energy are frequently returned [10]. Techniques that can produce low-energy geometries that avoid these local minima are clearly needed [72].

8 Discussion

Computed-assisted methods for rational drug design are likely to combine a number of different techniques like randomized search methods, efficient indexing schemes, algebraic techniques, constrained optimization, etc. Undoubtedly, the geometry of the ligands is only one part of the picture of rational drug design, the other being the energy and chemical properties of the molecules involved. Software tools that consider molecular geometries and perform simple energy calculations can help in the early stages of drug development [2, 5, 6, 72]. The increased use of such tools may also contribute to an improved understanding of drug action and to the development of models that can better explain drug activity [29, 41]. Last but not least, the amount of data that is now available in molecular databases makes such tools indispensable to medicinal chemists. From a computational point of view, the geometric problems that arise in drug design, even when simple energy models are assumed, are truly challenging.

Acknowledgment

The author is partially supported by a grant from Pfizer Central Research. Many of the problems and ideas described in this paper have been formulated in the course of a joint project between the Robotics Laboratory at Stanford University and Pfizer Central Research. The author is grateful to Prof. Jean-Claude Latombe of Stanford and Dr. Paul Finn of Pfizer Central Research who initiated and developed the project and to the rest of the people involved in this effort: Dan Halperin, Rajeev Motwani, Christian Shelton, and Suresh Venkatasubramanian.

References

[1] O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A computer-vision based technique for 3d sequence independent structural comparison of proteins. *Protein Engineering*, 6(3):279–288, 1993.

[2] L. Balbes, S. Mascarella, and D. Boyd. A perspective of modern methods in computer-aided drug design. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 5, pages 337–370. VCH Publishers, 1994.

[3] D. Barnum, J. Greene, A. Smellie, and P. Sprague. Identification of common functional components among molecules. To appear in J. Chem. Inf. Comput. Sci., 1996.

[4] J. Blaney, G. Crippen, A. Dearing, and J. Dixon. Dgeom: Distance geometry. Quantum Chemistry Program Exchange, 590, Dept. of Chemistry, Indiana Univ., IN.

[5] J. Blaney and S. Dixon. A good ligand is hard to find: Automated docking methods. *Perspectives in Drug Discovery and Design*, 1:301–319, 1993.

[6] B. Boyd. Successes of computer-assisted molecular design. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 1, pages 355–371. VCH Publishers, 1990.

[7] D. B. Boyd. Aspects of molecular modeling. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 1, pages 321–351. VCH Publishers, 1990.

[8] C. Bron and J. Kerbosch. Finding all cliques of an undirected subgraph. *Commun. ACM*, 16:575–577, 1973.

[9] G. Chang, W. Guida, and W. Still. An internal coordinate monte-carlo method for searching conformational space. *J. Am. Chem. Soc.*, 111:4379–4386, 1989.

[10] D. Clark, G. Jones, P. Willet, P. Kenny, and R. Glen. Pharmacophoric pattern matching in files of three-dimensional chemical structures: Comparison of conformational searching algorithms for flexible searching. *J. of Chem. Inf. Comput. Sci.*, 34:197–206, 1994.

[11] M. Connolly. Analytical molecular surface calculation. *J. of Applied Crystallography*, 16:548–558, 1983.

[12] M. Connolly. Shape complementarity at the hemoglobin alpha1-beta1 subunit surface. *Biopolymers*, 25:1229–1247, 1986.

[13] J. Craig. *Introduction to Robotics.* Addison-Wesley, Reading, MA, 1986.

[14] G. Crippen and T. Havel. *Distance Geometry and Molecular Conformation.* Research Studies Press, Letchworth, U.K., 1988.

[15] R. DesJarlais, R. Sheridan, J. Dixon, I. Kuntz, and R. Venkatarghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J. of Medicinal Chemistry*, 29:2149–2153, 1986.

[16] K. Dill. Folding proteins: Finding a needle in a haystack. *Current Opinion in Structural Biology*, 3:99–103, 1993.

[17] J. V. Drie, D. Weininger, and Y. Martin. Alladin: An integrated tool for computer-assisted molecular design and pharmacophore recognition, from geometric steric and substructure searching of three-dimensional molecular structures. *J. of Computer-Aided Molecular Design*, 3:225–251, 1989.

[18] H. Edelsbrunner. The union of balls and its dual shape. In *Proc. of the 9th Annual Symposium on Computational Geometry*, pages 218–231, 1993.

[19] H. Edelsbrunner, M. Facello, P. Fu, and J. Liang. Measuring proteins and voids in proteins. In *Proc. of the 28 Hawaii International Conf. on Systems Sciences*, pages 256–264, Wailea, Hawaii, 1995.

[20] H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. In *DIMACS Workshop on Computational Biology*, Rutgers, NJ, 1995.

[21] O. Faugeras. *Three-Dimensional Computer Vision.* MIT Press, Cambridge, MA, 1993.

[22] P. Finn, D. Halperin, L. Kavraki, J.-C. Latombe, R. Motwani, C. Shelton, and S. Venkatasubramanian. Geometric manipulation of flexible ligands. In M. Lin and D. Manocha, editors, *LNCS Series - 1996 ACM Workshop on Applied Computational Geometry.* Springer-Verlag, 1996.

[23] E. Fontain. Applications of genetic algorithms in the field of constitutional similarity. *J. Chem. Inf. Comput. Sci.*, 32:748–752, 1992.

[24] B. Freyberg, T. Richmond, and W. Braum. Surface area effects on energy refinement of proteins: a comparative study on atomic solvation parameters. *J. Molecular Biology*, 233:275–292, 1993.

[25] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, San Francisco, 1980.

[26] A. Ghose, J. Kowalczyk, M. Peterson, and A. Treasurywala. Conformational searching methods for small molecules: I. study of the sybyl search method. *J. of Computational Chemistry*, 14(9):1050–1065, 1993.

[27] R. Glen, G. Martin, A. Hill, R. Hyde, P. Wollard, J. Salmon, J. Buckingham, and A. Robertson. Computer-aided design and synthesis of 5-substituted tryptamines and their pharmacology at the $5-HT_{10}$ receptor: Discovery of compounds with potential anti-migraine properties. *J. of Medicial Chemistry*, 38:3566–3580, 1995.

[28] N. Go and H. Scherga. Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3(2):178–187, 1970.

[29] V. Golender and E. Vorpagel. Computer-assisted pharmacophore identification. In H. Kubinyi, editor, *3D QSAR in Drug Design*, pages 137–149. ESCOM, Leiden, 1993.

[30] P. Goodford. A computational procedure for determining energetically favored binding sites on biologically important macromolecules. *J. of Medicinal Chemistry*, 28:849–857, 1985.

[31] D. Goodsell and A. Olson. Simulated annealing and docking. *Proteins*, 8:195–202, 1990.

[32] J. Greene, S. Kahn, H. Savoj, P. Sprangue, and S. Teig. Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.*, 34:1297–1308, 1994.

[33] W. Grimson and T. Lozano-Pérez. Model-based recognition and localization from sparse range and tactile data. *The International Journal of Robotics Research*, 3(3):3–35, 1984.

[34] W. Guida, R. Bohacek, and M. Erion. Probing the conformational space available to inhibitors in the thermolysin active site using monter carlo/energy minimization techniques. *J. of Computational Chemistry*, 13(2):214–228, 1992.

[35] D. Halperin, J.-C. Latombe, and R. Motwani. Dynamic maintenance of kinematic structures. In J.-P. Laumond and M. Overmars, editors, *Algorithmic Foundations of Robotics*. A K Peters, MA, 1996.

[36] D. Halperin and M. Overmars. Spheres, molecules and hidden surface removal. In *Proc. 10th ACM Symposium on Computational Geometry*, pages 113–122, Stony Brook, 1994.

[37] T. Hurst. Flexible 3D searching: The directed tweak method. *J. Chem. Ing. Comp. Sci.*, 34:190–196, 1994.

[38] L. Kavraki. *Random Networks in Configuration Space for Fast Path Planning*. PhD thesis, Stanford University, 1995.

[39] S. Kearsley, D. Underwood, R. Sheridan, and M. Miller. Flexibases: A way to enhance the use of molecular docking methods. *J. of Computer-Aided Molecular Design*, 8:565–582, 1994.

[40] G. Klebe and T. Mietzener. A fast and efficient method to generate biologically relevant conformations. *J. of Computer-Aided Molecular Design*, 8:583–606, 1994.

[41] H. Kubinyi. *3D QSAR in Drug Design*. ESCOM, Leiden, 1993.

[42] G. Kuhl, G. Crippen, and D. Friesen. A combinatorial algorithm for calculating ligand binding. *J. of Computational Chemistry*, 5:24–34, 1984.

[43] I. Kuntz, J. Blaney, S. Oatley, R. Langridge, and T. Ferrin. A geometric approach to macromolecular-ligand interactions. *J. of Molecular Biology*, 161:269–288, 1982.

[44] Y. Lamdan and H. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *IEEE International Conference on Computer Vision*, pages 238–249, Tampa, FL, 1988.

[45] M. Lawrence and P. Davis. CLIX: A search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. *Proteins*, 12:31–41, 1992.

[46] A. Leach. A survey of methods for searching the conformational space of small and medium sized molecules. In K. Lipkowitz and D. Boyd, editors, *Reviews in Computational Chemistry*, volume 2, pages 1–47. VCH Publishers, 1991.

[47] A. Leach and I. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. of Computational Chemistry*, 13:730–748, 1992.

[48] B. Lee and F. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. of Molecular Biology*, 55:379–400, 1971.

[49] T. Lengauer. Algorithmic research problems in molecular bioinformatics. In *IEEE Proc. of the 2nd Israeli Symposium on the Theory of Computing and Systems*, pages 177–192, 1993.

[50] J. Liang, P. Sudhakar, H. Edelsbrunner, P. Fu, and S. Subramanian. Analytical shape computing of macromolecules: Molecular area and volume through alpha-shapes. In preparation.

[51] M. Lipton and W. Still. The multiple minimum problem in molecular modeling: Tree searching internal coordinate conformational space. *J. of Computational Chemistry*, 9(4):343–355, 1988.

[52] T. Lybrand. Computer simulation of biomelecular systems using molecular dynamics and free energy perturbation methods. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 1, pages 295–320. VCH Publishers, 1990.

[53] D. Manocha. *Algebraic and Numeric Techniques for Modeling and Robotics*. PhD thesis, University of California, Berkeley, 1992.

[54] D. Manocha, Y. Zhu, and W. Wright. Conformational analysis of molecular chains using nano-kinematics. *Computer Application of Biological Sciences (CABIOS)*, 11(1):71–86, 1995.

[55] Y. Martin, M. Bures, E. Danaher, J. DeLazzer, and I. Lico. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. In *J. of Computer-Aided Molecular Design*, volume 7, pages 83–102, 1993.

[56] Y. C. Martin, M. G. Bures, and P. Willet. Searching databases of three-dimensional structures. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 1, pages 213–256. VCH Publishers, 1990.

[57] D. Mayer, C. Naylor, L. Motoc, and G. Marshall. A unique geometry of the active site of angiotensin-converting-enzyme consistent with structure activities studies. *J. of Computer-Aided Molecular Design*, 1:3–16, 1989.

[58] P. G. Mezey. Molecular surfaces. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 1, pages 265–289. VCH Publishers, 1990.

[59] M. Miller, S. Kearsley, D. Underwood, and R. Sheridan. Flog: A system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. of Computer-Aided Molecular Design*, 8:153–174, 1994.

[60] M. Nicklaus, S. Wang, J. Driscoll, and G. Milne. Conformational changes of small molecules binding to proteins. *Bioorganic and Medicinal Chemistry*, 3(4):411–4128, 1995.

[61] R. Norel, D. Fischer, H. Wolfson, and R. Nussinov. Molecular surface recognition by a computer-based technique. *Protein Engineering*, 7(1):39–46, 1994.

[62] D. Parsons and J. Canny. Geometric problems in molecular biology and robotics. In *Intelligent Systems for Molecular Biology*, pages 322–330, Palo Alto, CA, 1994.

[63] M. Raghavan and B. Roth. Kinematic analysis of the 6r manipulator of general geometry. In *International Symposium of Robotics Research*, pages 314–320, Tokyo, 1989.

[64] M. Rarey, B. Kramer, and T. Lengauer. Time efficient docking of flexible ligands into active sites of proteins. In *International Conference on Intelligent Systems for Molecular Biology*, Cambridge, 1995.

[65] F. Richards. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J. of Molecular Biology*, 82:1–14, 1974.

[66] F. Richards. Areas, volumes, packing, and protein structures. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

[67] B. Shoichet, D. Bodian, and I. Kuntz. Molecular docking using shape descriptors. *J. of Computational Chemistry*, 13(3):380–397, 1992.

[68] A. Smellie, S. Kahn, and S. Teig. Analysis of conformational coverage: 1. validation and estimation of coverage. *J. Chem. Inf. Comput. Sci.*, 35:285–294, 1995.

[69] Tripos. *UNITY*. St. Louis, MO.

[70] A. Varshney, F. P. Brooks, Jr., and W. V. Wright. Computing smooth molecular surfaces. *IEEE Computer Graphics & Applications*, 15(5):19–25, September 1994.

[71] W. Welsh and A. Jain. Hammerhead: Fast fully automated docking of flexible ligands to protein binding sites. In preparation.

[72] P. Willet. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J. of Molecular Recognition*, 8:290–303, 1995.