Computational Approaches to Drug Design

P. W. Finn* L. E. Kavraki†

Abstract

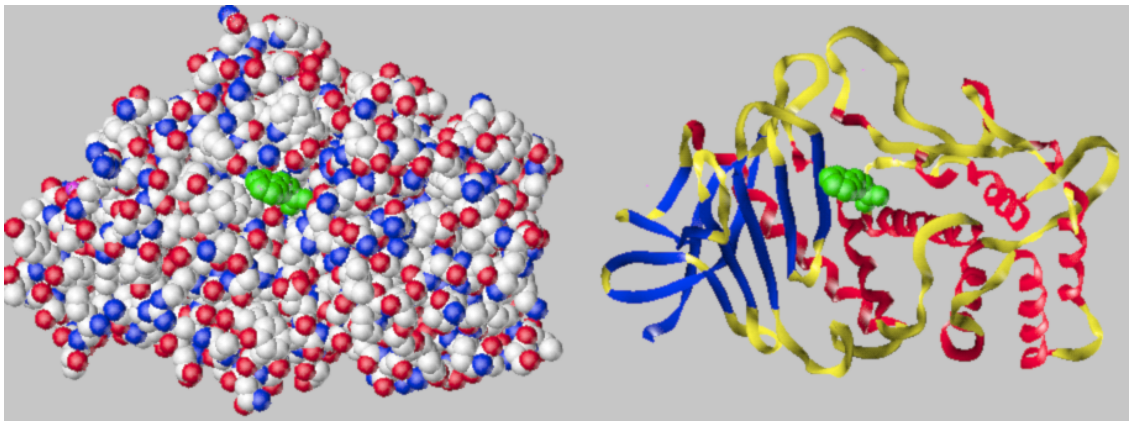
The rational approach to pharmaceutical drug design begins with an investigation of the relationship between chemical structure and biological activity. Information gained from this analysis is used to aid the design of new, or improved, drugs. Primary considerations during this investigation are the geometric and chemical characteristics of the molecules. Computational chemists who are involved in rational drug design routinely use an array of programs to compute, among other things, molecular surfaces and molecular volume, models of receptor sites, dockings of ligands inside protein cavities, and geometric invariants among different molecules that exhibit similar activity. There is a pressing need for efficient and accurate solutions to the above problems. Often, limiting assumptions need to be made, in order to make the calculations tractable. Also, the amount of data processed when searching for a potential drug is currently very large and is only expected to grow larger in the future. This paper describes some areas of computer-aided drug design that are important to computational chemists but are also rich in algorithmic problems. It surveys recent work in these areas both from the computational chemistry and the computer science literature.

1 Introduction

The design of pharmaceutical drugs is an extremely complex and still not completely understood process [7]. 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

*Pfizer Central Research, Sandwich, UK. Author's current address: Prolofix Ltd., 91 Milton Park, Milton, Oxfordshire, UK.

†Department of Computer Science, Rice University, Houston, TX 77005, USA.



(a) (b)

Figure 1: (a) The protease thermolysin with one of its known inhibitors (CDP), (b) the folding of thermolysin and the CDP inhibitor.

drug design [16], and is becoming more relevant as the understanding of molecular activity improves and the amount of available experimental data that requires processing increases.

The Framework. A fundamental assumption for rational drug design is that drug activity is obtained through the molecular binding of one molecule (the ligand) to the pocket of another, usually larger, molecule (the receptor, which is commonly a protein). In their active, or binding, conformations, the molecules exhibit geometric and chemical complementarity, both of which are essential for successful drug activity [7, 84]. By binding to these macromolecules, drugs may modulate signal pathways, for example by altering sensitivity to hormonal action, or by altering metabolism, for example by interfering with the catalytic activity of the enzyme. Most commonly, this is achieved by binding in the specific cavity of the enzyme (the active site) which catalyses the reaction, thus preventing access of the natural substrate(s). Figure 1(a) shows the protease thermolysin and one of its inhibitors. Thermolysin is the large molecule shown in the picture, while the inhibitor (a carboxymethyl dipeptide, CDP) is drawn near the center of the molecule. The 3D structure of the complex has been obtained by X-ray crystallography and can be retrieved from the Brookhaven protein data bank (code 1TMN) [10]. The protein is a linear polymer of amino acids which fold into a compact globular state. Figure 1(b) shows the folding of the polymer chain of thermolysin as a ribbon and the inhibitor as a set of spheres.

The Modeling. The modeling of molecular structure is a complex task, in particular because most molecules are flexible, being able to adopt a number of different conformations that are of similar energy. Quantum mechanics provide a detailed description of molecules in terms of the positions of atomic nuclei and the electron distribution among them. However, quantum mechanical calculations cannot be used to treat even small molecules because of high computational demands. The modeling of the binding process is also a difficult task, as 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 idealized conditions (e.g. in a vacuum) in order to simplify energy calculations. It is clear that the more accurate the model used, the better the chances chemists stand in predicting molecular interactions. Nevertheless, a large number of predictions made with approximate models have been confirmed with experimental observations [16, 55]. This has encouraged researchers to build tools that use approximate models and investigate the extent to which these tools can be useful. As discussed later in this article, these approximate models pose difficult algorithmic questions. More accurate molecular modeling, gained through better theoretical understanding or increased computational power, can only improve the techniques developed with simpler models.

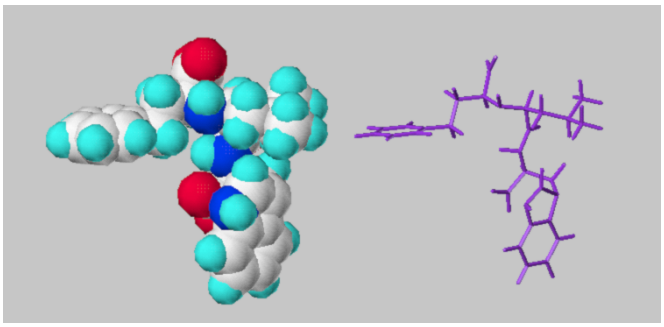
Two classes of problems. Depending on whether the chemical and geometric structure of the receptor is known or not, the problems arising can be classified into two broad categories. If the receptor is known, chemists are interested in finding if a ligand can be placed inside the binding pocket of the receptor in a conformation that results in a low energy for the complex. This problem is referred to as the *docking problem*. It has several variations: an accurate description of the binding interaction may be desired, or an approximate estimate may be sought of which ligands, from those contained in a huge database, are likely to fit inside the receptor.

Very often the binding pocket is unknown. In fact, the 3D structure of relatively few large molecules (or macromolecules) has been determined by X-ray crystallography or NMR techniques, although this number is increasing rapidly. In this case, indirect approaches must be adopted. These approaches use a number of ligands that have been experimentally found to interact with that specific receptor. 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 in a specific 3D arrangement contained in all the active conformations of the considered molecules. A prevailing hypothesis is that the pharmacophore is the part, or parts, of the molecule that is responsible for drug activity, while the rest of the molecule is a scaffold for the pharmacophore's features. If the pharmacophore is determined, by examining the different activities, relative shapes, and chemical structures of the initial molecules, chemists can use it to design a more potent pharmaceutical drug [46]. One common strategy is to design rigidified molecules that must present the correct binding geometry. Other things being equal, these will bind more strongly than their flexible counterparts. This is particularly interesting when two of the pharmacophoric groups would have a natural tendency to associate [121]. This strategy can be challenging, however, as rigidifying into even a slightly incorrect geometry can cause dramatic deceases in binding - geometric complementarity is very important in this context.

The Methods. The techniques that have been used so far in computer-aided drug design include graphics algorithms (visualization of molecules), geometric calculations (surface computation), numerical methods (energy minimization), graph theoretic methods (invariant identification), randomized algorithms (conformational search), computer vision methods (docking), and a variety of other techniques like genetic algorithms and simulated annealing. A number of tools for performing complex geometric and energy calculations are now available and the success of these computer-aided methods is under evaluation [7, 16].

This paper surveys some of the computational approaches to rational drug design. Four core areas which are rich in algorithmic problems have been identified: (a) surface and volume calculations, (b) conformational search, (c) docking, and (d) pharmacophore identification. We define the problems in each of these areas and discuss recent work in the computational chemistry and computer science communities. The discussion reveals the wealth and diversity of the algorithmic issues that arise in the domain of computer-aided pharmaceutical drug design. The interested reader is also referred to other recent surveys on topics in molecular biology and bioinformatics [84, 102]. An earlier version of this paper appeared in [67].



(a) (b)

Figure 2: The hard-sphere model and stick diagram of CDP.

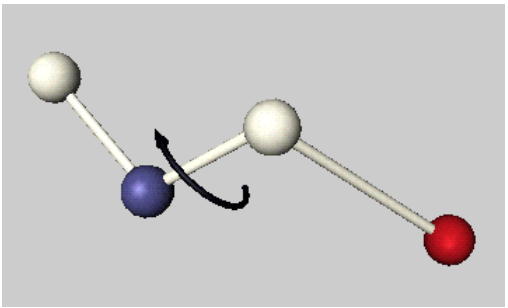


Figure 3: Illustration of torsional angles.

2 Some Background on Molecular Modeling

The *hard-sphere* model of CDP, 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* [15]. 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*. Figure 3 illustrates the torsional angles and the rotation

that they allow.

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 central 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 [47, 55]. 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. A conformation of a molecule is obtained by assigning values to all the DOF of the molecule.

Standard geometries are commonly used to construct reasonable models of molecules. For example, tables have been compiled of "preferred values" for bond lengths and these depend on the kind of atoms participating in these bonds [15]. Preferred values have also been tabulated 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 [5] databases. Although it is true that there is variability in the geometric data in these repositories, the information gathered provides a reasonable approximation of reality [15, 55].

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 [88]. Frequently, terms that model solvation effects are also included. Interactions of the molecule with the solvent in which it is dissolved are very important but also difficult to model accurately [15, 55]. 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:

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

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. Defining a force field parameterization within this framework that reproduces geometries and energies accurately is a time-consuming process, as is the addition of new parameters to the force field. 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 [88], 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 [12, 15, 55].

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 medicinal chemists but is out of the scope of this paper [30, 53].

3 Molecular Surfaces and Volume

We begin our survey with the computation of molecular surfaces and volume. Although it can be argued that molecules do not have surfaces and volume analogous to our perception of macroscopic surfaces and volume, this abstraction is more than helpful for visualizing molecules and their interactions [95]. It is also useful in calculations for molecular recognition and docking

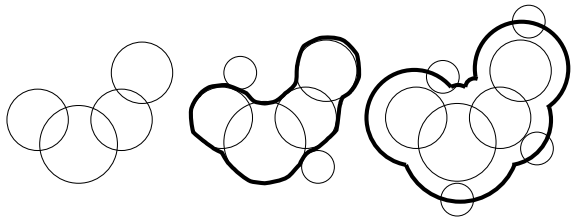


Figure 4: (a) van der Waals, (b) smooth molecular, and (c) solvent accessible surface.

[24], in computations of the energy of a molecule in solution [40], and in the process of pharmacophore identification, since atoms that are buried or little exposed are not likely to participate in a pharmacophore.

3.1 The Problem

The surfaces that are of interest to chemists include the *van der Waals* surface (defined in Section 2), the *smooth molecular* surface, and the *solvent accessible* surface [82, 111]. Figure 4 illustrates these different surfaces in two dimensions. The smooth molecular surface and the solvent accessible surface are defined with the help of a solvent molecule which is a sphere of radius r. In particular, the smooth 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. Borrowing our terminology from robotics, it is the union of the Minkowski sums of each molecular atom sphere and the sphere of the solvent [57].

In all the above cases what is required is the calculation and accurate representation of the considered surface and volume. Other useful information is the contribution of each atom to the surface, the connectivity of the surface patches, as well as the location of voids, pockets, and canyons in the molecule.

3.2 The Methods

Both numerical and analytical methods have been developed for the computation of surfaces and volume of molecules. A survey of early techniques is given in [95] and many references to recent work can be found in [86].

Most numerical or approximation techniques involve some kind of discretization which is in general a polyhedral decomposition of the space occupied by the molecule, or a covering of surfaces with a large number of dots. All these techniques trade accuracy for speed of calculations.

As far as analytical methods are concerned, closed expressions have been obtained for the different patches that compose the molecular surfaces described in the previous subsection. This line of work has not led to robust software implementations, a fact partly due to the complexity of the calculations involved with highly intersecting atoms. Techniques from computational geometry have also been applied. These techniques include alpha-shape theory and results from arrangements. Dealing with degeneracies and arithmetic errors in this context is imperative as surface and volume computations are applied to larger and larger molecules.

3.3 Literature Survey

One of the most widely used approximate methods is described in [110]. This approach uses bisector planes to divide the space of the internal atoms into polyhedra. Atoms on the surface are divided in the same way by considering virtual solvent atoms placed next to them. The area and volume are calculated through these polyhedra. Another approximate technique that enjoys widespread use is discussed in [25]. This method produces a set of dots on the surface of the molecule. The dots are calculated as follows. A probe sphere is placed tangent to each atom, each pair of neighboring atoms, and each triple of neighboring atoms. Points that are on the molecule-facing surface of the probe became part of the smooth molecular surface when the probe does not intersect any other atoms. The probe is moved at small angular increments around each atom and pair of atoms. It is clear that dot spacing influences the accuracy of the solution of this approach. Other methods that approximately compute molecular surfaces and volume include [52, 98, 119, 123].

Analytical calculation of surfaces was first done in [23]. The output of this algorithm consists of a set of curved regions of spheres and tori, joined together at circular arcs. Closed forms for

volume have also been derived in the same work [23]. To simplify calculations the authors did not consider intersections of 4 or more atoms, which can result in significant miscalculations [86]. Intersections that involve more than 4 atoms were later treated in [44, 45]. Other analytical methods have been presented in [75, 93, 105]. Unfortunately there are only very few implementations of the above methods.

In recent work, computational geometry techniques have given rise to methods for surface and volume calculations that overcome many of the limitations of previous methods [32, 33, 42, 57, 86]. In [57] it was 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 was 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 smooth molecular surface. The work in [58] addresses the degeneracies and precision problems inherent in computing spherical arrangements while using floating point arithmetic.

Alpha shape theory also proved a useful tool in accurately computing the surface and volume of molecules [31]. 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 and w denotes the van der Waals radius 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, smooth 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 alpha complex permits the identification of voids and canyons in the molecule [32, 33, 86]. Comparing shapes of different resolution it is possible to identify the cavities on the surface of the protein responsible for ligand binding [106]. Alpha shapes have also been used in [117] 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.

Although algorithms that compute molecular surfaces have been widely investigated, little has been done for their dynamic maintenance. Dynamic maintenance would, however, have interesting applications. For example, in calculation of binding energies, 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 [56].

4 Conformational Search

4.1 The Problem

Given a ligand and its degrees of freedom (see Section 2), the conformational search problem is to find a set of conformations of the ligand whose energy is below a threshold and which are geometrically distinct [79].

A particular case of the conformational search problem is the protein folding problem. It is believed that proteins have "unique" 3D shapes (the "native" state), which correspond to the global minima of their total energy and which are specified only by the chemical composition of the molecules. This is supported by the ability of proteins to refold to the native state following unfolding by altered salt concentration or temperature. Finding these conformations is by no means an easy task and involves several hundreds of DOF. For a survey of protein folding techniques see [29].

For small ligands, finding the conformation with the minimum energy is of little interest. In the process of computer-aided pharmaceutical drug design, several low energy conformations of a ligand are used in docking [96] and pharmacophore identification [91]. These represent the conformations of the molecule that are present in solution, and thus those that are available for binding to the receptor.

Frequently additional constraints among the atoms of a molecule are imposed. For example, "distance constraints" may specify the "desired" relative positions of two or more atoms (or features) of the molecule. When constraints are specified, the output of conformational search is a set of conformations that are geometrically distinct, are of low energy, and satisfy the constraints. We refer to this problem as the constrained conformational search problem to contrast it with the previously defined unconstrained conformational search. Tools that can produce conformations that satisfy known constraints have additional applications in database screening [92].

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 as fixed, choose torsional angle values from predefined distributions, and simplify the energy model used [47, 55]. Frequently the molecule is assumed to be in a vacuum. An external potential can be considered with most conformational search methods but may result in longer computation times [55].

4.2 The Methods

When the degrees of freedom (in most cases the torsional angles) of the molecule are more than 4 and no pruning is done, any kind of systematic search method is impractical. Systematic search discretizes each of the degrees of freedom of the molecule and explores all possible combinations of them minimizing the energy of every single combination.

Most implemented unconstrained search procedures employ Monte-Carlo randomized techniques to limit the number of conformations explored. For example, the search may apply random increments to each of the degrees of freedom of the molecule, or choose the values of the degrees of freedom according to a distribution that fits experimental data of "preferred" values of the corresponding degree. An extended bibliography of recent search methods can be found in [43, 79]. A number of important questions are raised in the above framework, namely how to ensure the diversity of the sample and when to stop the search procedure. Another question that is closely linked with unconstrained conformational search in this framework is how to reliably identify and eliminate from the output conformations that are very similar in both geometric and energy terms. Most frequently clustering techniques are used for grouping similar conformations into sets and then a representative per set is selected. Recent work focuses on developing robust methods for clustering possibly thousands of 3D conformations.

As far as constrained conformational search is concerned, it clear that if the ligand molecule is flexible, a blind search has very few chances of producing, in a reasonable amount of time, a ligand that respects the specified constraints. Here techniques from robotics have been used. Typically, part of the molecule is considered as a linear manipulator with its tip (an atom of the molecule) at the desired location. The values of the degrees of freedom of the manipulator are computed to accommodate the positioning of the tip. The solution of inverse kinematics is by no means a simple task (see [90] for some pointers to recent work).

4.3 Literature Survey: Unconstrained Search

Systematic search methods sample each torsional DOF of the ligand at regularly spaced intervals and were among the first to be developed and used [87]. The discretization of the torsional values is typically as coarse as 30^o or 60^o [79]. Even with such a resolution the number of conformations that are generated with a systematic search can be very large [43]. Furthermore, 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 in an effort to enhance the diversity of the sample. For example, in [115] a technique called poling is introduced. This technique uses a penalty energy function that forces similar conformations away from each other. In [80] clustering is interleaved with conformational search and new conformations are created only if they are dissimilar to already produced cluster representatives.

A variety of randomized methods are currently under investigation: conformations are obtained by applying random increments to the torsional DOF of the molecule starting from a user-specified initial conformation [43], or from a previously found low-energy conformation [20]. Techniques that use molecular dynamics trajectories [19], simulated annealing [103], or distance geometry [27] for obtaining distinct conformations have also been tried. Recent articles, which attempt to compare different methods, emphasize the superior quality of the results obtained with randomized methods [43].

The random sampling method for exploring the conformation space of small molecules described in [36, 37] works as follows. Initially a large number of conformations are generated at random. A random conformation is obtained by selecting each degree of freedom from its allowed range according to a user-specified distribution. This distribution is frequently the uni-

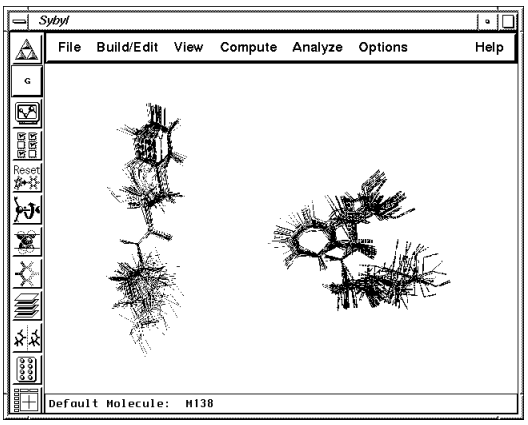


Figure 5: Two clusters of CDP.

form distribution. However, if some a priori information is available about the preferred values of a particular degree of freedom [70], then the corresponding values are selected according to a distribution that reflects this information (i.e. a Gaussian distribution). The energy of the resulting conformation is minimized with an efficient technique [107]. To obtain a representative set of conformations from the sample, the method partitions the conformations into sets that reflect geometric similarity as captured by the distance measure DRMS. DRMS($\mathbf{c}_i$,$\mathbf{c}_j$) is the square root of the mean of the squared distances of the corresponding atoms of $\mathbf{c}_i$ and $\mathbf{c}_j$, after $\mathbf{c}_i$ is transformed to $\mathbf{c}_j$. This transformation is computed using a basis of three predefined atoms a_1, a_2, and a_3. The clustering performed minimizes the maximum inter-cluster distance [49]. Figure 5 shows two of the clusters obtained with the randomized approach of [36, 37] for CDP. At the end of the clustering step, a representative per cluster can be retained. The method described above borrows ideas from randomized techniques for path planning in high-dimensional configuration spaces [66]. Experimental observations show that it is very effective in discovering the local minima of the irregular energy landscape of ligands with 5 to 15 degrees of freedom.

Unconstrained conformational search thus 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 or clustering can be done efficiently. Incremental clustering techniques can be very interesting in the context of conformational search. For some recent results see [21].

Improvements in conformational search software directly affect the quality of solutions ob-

tained for the problems where these conformations are actually used (i.e. docking and pharmacophore identification). A discussion on the quality of the conformational coverage required for the above problems can be found in [114].

4.4 Literature Survey: Constrained Search

Most of the techniques described in the previous section will produce poor results when distance constraints are imposed in the 3D structure of the molecule. Distance constraints arise frequently in practice. For example, chemists may be interested in conformations that keep two atoms (features) of the molecule at specific positions in space because these two atoms belong to a pharmacophore. Database queries are usually specified by a 3D graph whose nodes correspond to specific features and whose edges correspond to diatomic distances. 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 (and also that bond angle constraints are satisfied).

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).

Work done in robotics for computing inverse kinematics of manipulators, is exploited in [90] to find valid conformations for small molecular chains. In particular, the case of a serial manipulator with 6 consecutive 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 $SE(3)$ [26]). Symbolic manipulation of the equations of Raghavan and Roth [108] transforms the inverse kinematics problem into one of computing the eigenvalues and eigenvectors of a matrix, which in turn can be done very efficiently [89]. 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 [90] are very efficient when computing conformations that maintain ring closure and local deformations of small protein chains. It is worth mentioning that algebraic equations in 6 unknowns were also derived in [47] 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. In contrast with the above algebraic methods, the work in [9] computes the conformations of large cyclic molecules with purely geometric techniques.

Other kinds of methods, like distance geometry [27], have also been 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 [27]. NMR spectroscopy produces distances between pairs of atoms of a macromolecule and the problem is to construct a conformation of the molecule that satisfies 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 [90].

Constrained conformational search is frequently encountered in the domain of database searching, as mentioned earlier. Queries in many chemical database systems usually specify a set of features and and their pairwise distances. The result of the search is typically a set of molecules that contain these features and can assume a low-energy conformation that satisfies the pairwise distances. For a detailed discussion on database queries and the definition of similarity in this context see [17, 92, 112, 122]. The importance of fast methods for database searching can only be emphasized: database searching is becoming a basic tool for finding novel bioactive compounds [35].

Although, it is fairly simple to do an initial screening of a database with one million compounds and isolate, for example, molecules with a particular desired set of features, it is difficult to narrow down the results at later stages [122]. The main reason for the above is that 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. Constrained conformational search should be performed very fast in this domain [79, 92, 122]. So far, distance geometry techniques, systematic and randomized search, and genetic algorithms have been tried but have produced slow algorithms [11, 20, 22, 39]. Heuristic methods are the ones which can produce a solution fast, although this may not be a good solution. One of the most efficient heuristic techniques for flexible searching is the "Directed Tweak Method" [59, 116]. The method minimizes a pseudo-energy function which is the sum of the squares of the deviations of the distances found in the

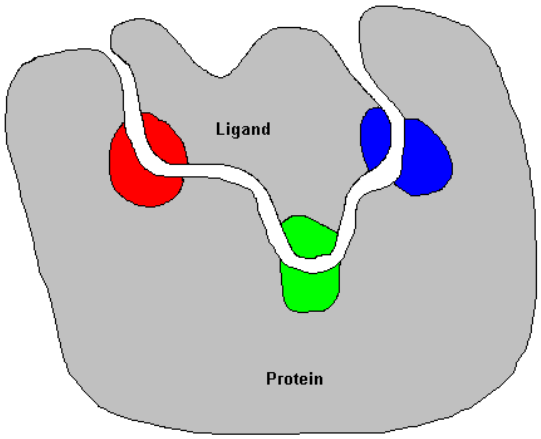


Figure 6: Docking of a ligand to a receptor. Shading indicates schematically the energetically most favourable interactions.

molecular structure to the distances expressed in the database query. Unfortunately the pseudo-energy function contains a large number of local minima and conformations having high energy are frequently returned [22]. Fast techniques that can produce low-energy geometries that avoid these local minima are clearly needed [122].

5 Docking

5.1 The Problem

Given two interacting molecules of known geometry the docking problem consists of finding their relative positions during the interaction. This paper is mostly concerned with docking small ligands into receptors that are of larger size, typically proteins. However docking problems are also studied when the molecules involved are proteins, or proteins and macromolecules like DNA or RNA [64, 85]. In the case of protein-ligand docking, information about the geometry of the receptor is obtained by X-ray crystallography or NMR techniques. It is generally assumed that the receptor molecule is rigid [81] and that the geometric and chemical complementarity at the binding site are important (see Figure 6). Figure 7 shows the pocket of the protein thrombin. Thrombin is an enzyme with an important function in blood coagulation; its inhibitors have potential application as anti-coagulants. The approximation of the rigid receptor is justified by

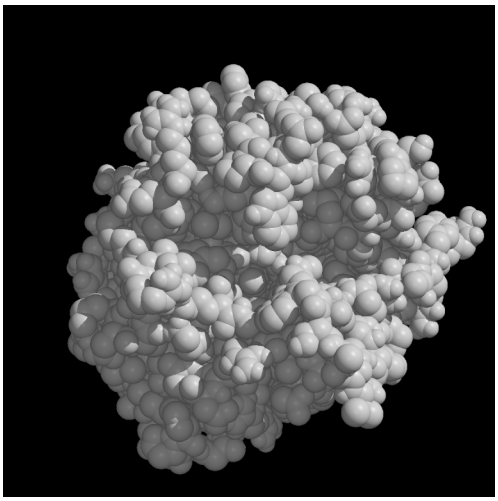


Figure 7: The binding pocket of thrombin. The cavity runs across from left to right, with the residues responsible for catalysis, and an overhanging surface group, at the center of the cleft.

experimental data i.e. crystals of the molecule with and without the ligand, but exceptions have also been noted [99]. As the understanding of the docking process improves, flexible models for the receptor may be considered. For the ligand however, it is essential to address its flexibility. The quality of the docking is measured in energy terms: the binding energy of the complex must be as low as possible [85].

5.2 The Methods

Assume for the moment that the receptor and the ligand are rigid. An obvious approach is to explore the 6D conformational space of the ligand inside the receptor's cavity, but this can be time consuming. An alternative approach is to represent the cavity with a number of sparse points and match these points against corresponding points of the ligand. Hence in this case the docking problem can be viewed as a point matching problem. The techniques used for this matching problem include graph-theoretic methods, computer vision techniques (i.e. geometric hashing), the use of the Fast Fourier Transform for computing optimal translations combined with rotational sampling, as well as heuristic techniques and evolutionary algorithms like genetic algorithms or simulated annealing.

When the ligand is flexible, then one could perform first a conformational search of the ligand and then dock the conformations obtained in the receptor assuming that these conformations

are rigid. Of course it is also possible to treat the flexibility of the ligand inside the receptor's cavity with randomized methods or any of the other methods that have been developed for conformational search, as this case can be seen as conformational searching with an additional potential. Another popular approach is to break down the ligand in smaller pieces and to dock these pieces separately inside the cavity. The problem in this case is treated with a combination of geometric techniques (shape matching) and efficient energy minimization procedures. An extensive bibliography on several recent docking methods is given in [85].

A central question for all the above cases 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 also a very important issue to be addressed in docking. Numerous approaches have been taken to this problem ranging from the highly empirical to more rigorous methods, such as free-energy perturbation (FEP) techniques. A detailed discussion is out of the scope of the present paper (see also [71]), but it is worth pointing out that a major component in many of the empirical schemes is a measure of the geometric complementarity of the ligand and receptor. This can be expressed as the buried surface area of the ligand [4, 13, 14], or the contact area of the two molecules [83]. Note that the FEP methods have also highlighted the importance of geometric complementarity in some cases [97].

5.3 Literature Survey: 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 [94]. A lower resolution approach has been developed to speed the search and implicitly allow for small conformational changes [61]. Alternatively, the Fast Fourier Transform (FFT) can be used to compute the possible translations of the ligand for a specific orientation, and this process can be repeated for multiple orientations [65]. The FFT essentially computes the Minkowski sum of the pocket and the ligand. Most recent methods, however, adopt the following 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" [50], "essential points" [96], or "match probes" [120].

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 favourable sites for certain functional groups [50, 96]. 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 [76, 113], 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 when matching ligand and receptor points [113]. This method has recently been extended to consider multiple conformations of the receptor, where these are available [72].

As noted above, 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. For example, the CLIX method [78], which uses minima in interaction energy maps [50] as centers to which it tries to fit ligands, tries all possible combinations. At each step a match is sought between two interaction points and corresponding features in the ligand. This initial orientation is then optimized, and those with good geometric and chemical complementarity are retained. All pairs of points are tried in the process.

Simple heuristics can be used to narrow the search of possible matchings. 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 and 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 build a "docking graph" using the receptor and ligand matching points [74]. 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 [41] but the method is reported to work well in practice [74].

The matching problem that arises in docking clearly has analogies with the geometric match-

ing performed for model-based shape recognition [34]. These analogies are extensively discussed in [102]. 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 [54] or geometric hashing [77, 109], 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 [6, 100]. 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.

5.4 Literature Survey: Flexible Ligand

For flexible ligands, a common approach is to consider different low-energy conformations of the ligand, obtained by a conformational search procedure. These conformations are tried against the receptor cavity using a technique developed for docking a rigid ligand to a rigid receptor [96]. To facilitate such docking approaches, several molecular databases now store a set of geometrically distinct conformations per ligand [68]. Recent work advocates that a good approximation to the docked position of the ligand can be found by storing only a small number of conformers in 3D chemical databases [114]. It is clear that if the active conformation is not close to the conformations considered, these methods will fail to produce a docking close to the optimal one.

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 [51]. Genetic algorithms have also been applied [101, 118]. In particular, the genetic algorithm in [63] includes limited flexibility of the receptor in addition to full flexibility of the ligand. 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 [28, 109, 120].

The placement of the fragments raises difficult combinatorial and geometric questions. 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 [70].

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 good placements for these fragments may help in pruning 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 [12].

6 Pharmacophore Identification

6.1 The Problem

When the 3D structure of the target macromolecule is unknown, the identification of a pharmacophore is key to the rational development of new pharmaceutical drugs [91]. The pharmacophore is a set of features in a specific 3D arrangement that is present in all (or most) of the active conformations of a set of ligands that exhibit similar activity. As outlined in Section 1, 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 8). Thus, if a pharmacophore has been isolated, chemists can use it as a template to build more potent drugs [46]. Note, however, that other approaches have been developed which highlight the requirement for shape complementarity between the ligands [104].

As an example, consider the four different inhibitors of the protease thermolysin of Figure 1. These inhibitors are shown in Figure 9. The ligands have 5 to 11 torsional degrees of freedom and each of them can assume a few hundreds of distinct low-energy conformations [36, 79]. Each conformation gives rise to a 3D point set (typically 5-20 points or features per conformation). The pharmacophore, whose size is in general small (3-5 features), should be congruent with a subset of the features of at least one conformation of each considered molecule. The problem is further complicated by the fact the exact congruence is usually not achieved. The matching is done within some tolerances to reflect the uncertainties in the position of the

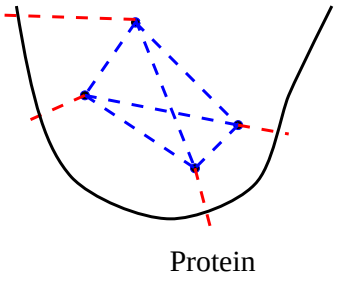


Figure 8: The features of the pharmacophore interact with features of the receptor cavity.

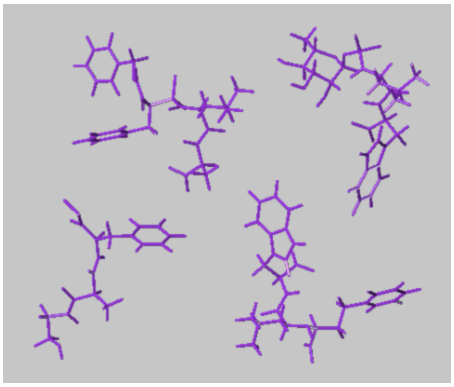


Figure 9: Four different inhibitors of thermolysin.

features. Such uncertainties arise from variability tolerated by the receptor, imperfect modeling, and the approximations done during conformational search. Note that in contrast with docking, the active conformations of the ligands may have no other features in common except the pharmacophore.

6.2 The Methods

Given two 3D point sets, finding a subset of the first which is congruent to a subset of the second is in itself a difficult problem. Geometric optimization techniques and primarily graph theoretic techniques have been used. Geometric hashing and other techniques from computer vision developed for model based recognition are again applicable. The fact that the matching should be done within small tolerances further complicates the problem.

When multiple conformations per ligand are considered there is an inherent combinatorial explosion in the process of identifying the pharmacophore. One issue that arises is that two set

congruence should be decided quickly and a clever scheme is needed for eliminating invariants as soon as possible. Randomized methods have recently been considered. Hashing schemes are also relevant in encoding the large number of "intermediate" invariants that are discovered in the process - the invariants that are found only in proper subsets of the molecules and are at some later stage eliminated.

6.3 Literature Survey

The problem of determining the congruence of point-sets is studied in [1, 2, 3, 69] using graph-theoretic methods. Determining the congruence in $\Re^3$ is tractable [1, 3] in the absence of complications such as noise. Undoubtly, invariant identification is more closely related to the problem of identifying the *largest common point set* (LCP). Unfortunately, the LCP problem turns out to exceedingly difficult; in fact, even for m collections of n points on the real line, the LCP cannot be approximated to within an n^{ϵ} factor unless $P = NP$, and only weak positive results are known [2, 69]. Of course the problem is polynomially solvable when $m = 2$ [3]. However in the case of pharmacophores, there is noise present in the data and the above methods have limited applicability.

DISCO [91], one of the most popular algorithms for pharmacophore identification, uses clique detection to identify invariants. Initially the program considers a pair of conformations $\mathbf{c}_1$ and $\mathbf{c}_2$ belonging to different molecules. A "correspondence graph" G is constructed and this graph is similar to the "docking graph" described in Section 5.3. The nodes of G are again all node pairs of $\mathbf{c}_1$ and $\mathbf{c}_2$. 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 [18] is then used to find cliques in G. These correspond to invariants in $\mathbf{c}_1$ and $\mathbf{c}_2$ and thus to candidate pharmacophores. The algorithm seems to work well in practice [91, 122]. Generalization of the above approach to n conformations (corresponding to n different molecules) is straightforward by considering one of the conformations as a reference and comparing it with all other $n-1$ conformations. If multiple conformations per molecule are available, all possible combinations should be tried. 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 [8]. This is the main reason that different approaches are under development. One idea is to start with small invariants (2-3 features) and gradually expand them [8]. Another idea is to use optimized geo-

metric hashing techniques from computer vision [38]. Another idea is to use genetic algorithms. A genetic algorithm has been described in which conformational flexibility is considered during the pharmacophore identification process. The chromosome encodes the torsional angles of the rotatable bonds and the feature mappings [62]. The fitness function is a weighted combination of the feature overlap (the pharmacophore), the common volume occupied by the ligands, and energy of the ligands (to prevent the identification of pharmacophores which use high-energy conformations). The method has been shown to work well in test cases.

Yet another idea is to use randomized techniques when searching for invariants [37]. Randomization can be introduced both when solving the two point set problem and when considering multiple conformations per molecule. In [37] two methods for determining congruence are presented. One just selects three random points from the first set and matches them with points in the other set that have approximately the same pairwise distances. Then a transformation is defined to overlap the triples and the rest of the points in the sets are checked for congruence. This process is repeated and the maximal matching is retained. The second approach randomly selects two subsets A and B of size $1/\alpha$ from the first set of points (α is the estimated size of the pharmacophore), and a subset C of size $1/\alpha$ from the second set of points. For every triangle (a, b, p) with $a \in A, b \in B$, and $p \in \overline{(A \cup B)}$, it finds matching triplets (c, p_1, p_2) with $c \in C$ and $p_1, p_2 \in \overline{C}$. The first triplet is transformed to the second and the matches are counted. Again the process is repeated and the maximal matching is retained. It is shown in [37] that the second scheme is more efficient than the first because of the distribution of distances among features that is encountered in molecules. The question of matching with noise is further treated in [60] using techniques from combinatorial geometry. The bounds given depend on the diameter of the point set. The problem of matching under noise is still open.

7 Discussion

Computed-assisted methods for rational drug design combine a number of different techniques. This article shows that methods from graph-theory, geometry, randomized algorithms, computer vision and graphics have useful applications in this domain. It is clear that there is a need for robust algorithms that provide good performance guarantees and allow the chemists to test their hypothesis. It is also clear that there is a need to further develop these algorithms to deal with noisy data.

Undoubtedly, the molecular representations and methods discussed here are only one part of the picture of rational drug design. Others include, for example, incorporating into the design appropriate selectivity for the desired target (so as to reduce side-effect profiles) and resistance to metabolic degradation. Nonetheless, software tools that consider molecular geometries and perform simple energy calculations, and in particular conformational search, docking and pharmacophore identification methods, can help in the early stages of drug development [7, 12, 16, 122]. The increased use of such tools can contribute to the development of improved models of the relationship between chemical structure and biological activity [48, 73], making drug design a more efficient and effective process. Last but not least, the amount of data that is now available in molecular databases makes such tools indispensable to medicinal chemists. From a theoretical point of view, the computational problems that arise in drug design, even when simple energy models are assumed, are truly challenging.

Acknowledgment

Work in this paper was partially supported by Pfizer Central Research. L. Kavraki is also supported by Rice start up funds. This work started when L. Kavraki was with the Robotics Laboratory at Stanford University and participated in a joint project between the Stanford Robotics Laboratory and Pfizer Central Research. Professor Jean-Claude Latombe, who led the Stanford team, explored many of the topics discussed here in the course of the above project. The authors are grateful to him for many discussions and comments that made this paper possible and also for his encouragement during the writing of the paper. The authors would also like to thank Danny Halperin, Steven LaValle, Rajeev Motwani, Christian Shelton, and Suresh Venkatasubramanian for many helpful discussions and comments. S. Venkatasubramanian produced Figure 3. An earlier version of this article appeared in [67].

References

[1] T. Akutsu. On determining the congruence of point sets in higher dimensions. *Lecture Notes in Computer Science*, 834:38–55, 1994.

[2] T. Akutsu and M. Halldórsson. On the approximation of largest common subtrees and largest common point sets. *Lecture Notes in Computer Science*, 834:405–413, 1994.

[3] T. Akutsu, H. Tamaki, and T. Tokuyama. Distribution of distances and triangles in a point set and algorithms for computing the largest common point set. In *Proceeding of the ACM Symposium on Computational Geometry*, pages 314–322, Nice, France, 1997.

[4] A. Alex and P. W. Finn. Fast and accurate predictions of relative binding free energies. To appear in Journal of Molecular Structure.

[5] F. H. Allen, J. E. Davies, J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith, and D. G. Watson. The development of versions 3 and 4 of the cambridge structural database system. *Journal of Chemical Information and Computer Science*, 31:187–204, 1991.

[6] 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.

[7] 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.

[8] D. Barnum, J. Greene, A. Smellie, and P. Sprague. Identification of common functional components among molecules. *Journal of Chemical Information and Computer Science*, 36:653–571, 1997.

[9] M. Benedetto, P. Lucibello, S. Sangiovanni-Vincentelli, and K. Yamaguchi. Chain closure: A problem in molecular cad. In *31st Design and Automation Conference*, pages 497–502, 1994.

[10] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, F. Meyer, M. D. Bryce, J. R. Rogers, O. Kennard, T. Shikanouchi, and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542, 1977.

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

[12] 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.

[13] R. S. Bohacek and C. McMartin. Definition and display of steric, hydrophobic and hydrogen bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: validation of a high-resolution graphical tool for drug design. *Journal of Medicinal Chemistry*, 35:1671–1684, 1992.

[14] H.-J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, 8:243–256, 1994.

[15] 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.

[16] D. 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.

[17] A. Brint and P. Willett. Algorithms for the identification of three dimensional maximal common substructures. *Journal of Chemical Information and Computer Science*, 27:152–158, 1987.

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

[19] D. Byrne, J. Li, E. Platt, B. Robson, and P. Weiner. Novel algorithms for searching conformational space. *Journal of Computer-Aided Molecular Design*, 8:67–82, 1994.

[20] G. Chang, W. Guida, and W. Still. An internal coordinate monte-carlo method for searching conformational space. *Journal of the American Chemical Society*, 111:4379–4386, 1989.

[21] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering. In *Proceedings 29th Annual ACM Symposium on Theory of Computing*, 1997.

[22] D. Clark, G. Jones, P. Willett, P. Kenny, and R. Glen. Pharmacophoric pattern matching in files of three-dimensional chemical structures: Comparison of conformational searching algorithms for flexible searching. *Journal of Chemical Information and Computer Science*, 34:197–206, 1994.

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

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

[25] M. Connolly. The molecular surface package. *Journal of Molecular Graphics*, 11:138–141, 1993.

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

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

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

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

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

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

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

[33] 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.

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

[35] P. W. Finn. Computer-based screening of compound databases for the identification of novel leads. *Drug Discovery Today*, 1:363–370, 1996.

[36] P. W. Finn, D. Halperin, L. E. 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*, pages 67–78. Springer-Verlag, 1996.

[37] P. W. Finn, L. E. Kavraki, J.-C. Latombe, R. Motwani, C. Shelton, S. Venkatasubramanian, and A. Yao. Rapid: Randomized pharmacophore identification. In *Proceedings of the International Symposium on Computational Geometry*, pages 324–333, Nice, France, 1997.

[38] D. Fischer, R. Nussinov, and H.Wolfson. 3-d substructure matching in protein molecules. In *Proceedings Combinatorial Pattern Matching*, pages 136–150, 1992.

[39] E. Fontain. Applications of genetic algorithms in the field of constitutional similarity. *Journal of Chemical Information and Computer Science*, 32:748–752, 1992.

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

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

[42] M. Gerstein and M. Levitt. The volume of atoms on the protein surface: Calculated from simulation, using voronoi polyhedra. *Journal of Molecular Biology*, 249:955–966, 1995.

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

[44] K. Gibson and H. Scheraga. Exact calculation of the volume and surface area of fused hard-sphere molecules with unequal atomic radii. *Molecular Physics*, 62:1247–1265, 1987.

[45] K. Gibson and H. Scheraga. Surface area of the intersection of three spheres with unequl radii: A simplified analytic formula. *Molecular Physics*, 62:641–644, 1988.

[46] 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 $5 - HT_{1D}$ receptor: Discovery of compounds with potential anti-migraine properties. *Journal of Medicinal Chemistry*, 38:3566–3580, 1995.

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

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

[49] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

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

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

[52] S. M. L. Grand and K. M. Merz. Rapid approximation of molecular surface area via the use of boolean logic and look-up tables. *Journal of Computational Chemistry*, 14:349–352, 1993.

[53] J. Greene, S. Kahn, H. Savoj, P. Sprague, and S. Teig. Chemical function queries for 3D database search. *Journal of Chemical Information and Computer Science*, 34:1297–1308, 1994.

[54] 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.

[55] 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. *Journal of Computational Chemistry*, 13(2):214–228, 1992.

[56] D. Halperin, J.-C. Latombe, and R. Motwani. Dynamic maintenance of kinematic structures. In J.-P. Laumond and M. Overmars, editors, *Algorithms for Robotic Motion and Manipulation*, pages 155–170. A K Peters, MA, 1997.

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

[58] D. Halperin and C. Shelton. A perturbation scheme for spherical arrangements with application to molecular modeling. In *ACM Conference on Computational Geometry*, pages 183–192, Nice, France, 1997.

[59] T. Hurst. Flexible 3D searching: The directed tweak method. *Journal of Chemical Information and Computer Science*, 34:190–196, 1994.

[60] P. Indyk, R. Motwani, and S. Venkatasubramanian. Geometric matching under noise: Combinatorial bounds and algorithms. Manuscript.

[61] F. Jiang and S. H. Kim. Soft docking: matching of molecular surface cubes. *Journal of Molecular Biology*, 219:79–102, 1991.

[62] G. Jones, R. C. Glen, and P. Willett. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *Journal of Computer-Aided Molecular Design*, 9:532–549, 1995.

[63] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267:727–748, 1997.

[64] N. Kasinos, G. A. Lilley, N. Subbarao, and I. Haneef. A robust and efficient automated docking algorithm for molecular recognition. *Protein Engineering*, 5:69–75, 1992.

[65] E. Katchalski-Katzir, I. Shavir, M. Eisenstein, A. Friesem, C. Aflalo, and I. Vakser. Molecular surface recognition: determination of geometric fit between proteins and ligands by correlation techniques. *Proceedings of the National Academy of Sciences*, 89:2195–2199, 1992.

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

[67] L. E. Kavraki. Geometry and the discovery of new ligands. In J.-P. Laumond and M. Overmars, editors, *Algorithms for Robotic Motion and Manipulation*, pages 435–448. A. K. Peters, 1997.

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

[69] S. Khanna, R. Motwani, and F. F. Yao. Approximation algorithms for the largest common set. Technical Report STAN-CS-95-1545, Stanford University, 1995.

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

[71] R. Klebe. Structure correlation and ligand/receptor interactions. In H. B. Burgi and J. Dunitz, editors, *Structure Correlation*, pages 543–561. VCH Weiheim, 1996.

[72] R. M. Knegtel, I. D. Kuntz, and C. M. Oshiro. Molecular docking to ensembles of protein structures. *Journal of Molecular Biology*, 266:424–440, 1997.

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

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

[75] C. Kundrot, J. Ponder, and F. Richards. Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. *Journal of Computational Chemistry*, 12:402–409, 1991.

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

[77] 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.

[78] 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.

[79] 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.

[80] A. Leach. An algorithm to directly identify a molecule's most different conformations. *Journal of Chemical Information and Computer Science*, 34:661–670, 1994.

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

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

[83] P. Lehnof. New contact measures for the protein docking problem. In *First Annual International Conference on Computational Molecular Biology*, pages 182–189. ACM, Albuquerque, NM, 1997.

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

[85] T. Lengauer and M. Rarey. Computational methods for biomolecular docking. *Current Opinion in Structural Biology*, 6:402–406, 1996.

[86] 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.

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

[88] T. Lybrand. Computer simulation of biomolecular 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.

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

[90] 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.

[91] 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 *Journal of Computer-Aided Molecular Design*, volume 7, pages 83–102, 1993.

[92] Y. C. Martin, M. G. Bures, and P. Willett. 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.

[93] N. Max and E. Getzoff. Spherical harmonic molecular surfaces. *IEEE Computer Graphics and Applications*, pages 42–50, 1998.

[94] E. Meng, D. Gschwend, J. Blaney, and I. Kuntz. Orientational sampling and rigid-body minimization in molecular docking. *Proteins: structure, Function and genetics*, 17:266–278, 1993.

[95] 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.

[96] 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. *Journal of Computer-Aided Molecular Design*, 8:153–174, 1994.

[97] S. Miyamoto and P. A. Kollmann. Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins: Structure, Function and Genetics*, 16:226–245, 1993.

[98] J. Muller. Calculation of scattering curves for macromolecules in solution and comparison with results of methods using effective atomic scattering factors. *Journal of Applied Crystalography*, 16:74–82, 1983.

[99] M. Nicklaus, S. Wang, J. Driscoll, and G. Milne. Conformational changes of small molecules binding to proteins. *Bioorganic and Medicinal Chemistry*, 3(4):411–428, 1995.

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

[101] C. M. Oshiro, I. D. Kuntz, and J. S. Dixon. Flexible ligand docking using a genetic algorithm. *Journal of Computer-Aided Molecular Design*, 9(2):113–130, 1995.

[102] 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.

[103] T. D. Perkins and P. M. Dean. An exploration of a novel strategy for superposing several flexible molecules. *Journal of Computer-Aided Molecular Design*, 7(2):155–172, 1993.

[104] T. D. Perkins, J. E. Mills, and P. M. Dean. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *Journal of Computer-Aided Molecular Design*, 9(6):479–490, 1995.

[105] G. Perrot, B. Cheng, K. Gilson, K. Palmer, A. Nayeem, B. Maigret, and H. Scheraga. Mseed : a program for the rapid analytical calculation of accessible surface and their derivatives. *Journal of Computational Chemistry*, 13:1–11, 1992.

[106] K. P. Peters, J. Fauck, and C. Frommel. The automated search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of Molecular Biology*, 256:201–213, 1996.

[107] D. Pierre. *Optimization theory with applications*. Dover, NY, 1986.

[108] 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.

[109] 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.

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

[111] F. Richards. Areas, volumes, packing, and protein structures. *Annual Reviews of Biophysics and Bioengineering*, 6:151–176, 1977.

[112] R. Sheridan, M. Miller, D. Underwood, and S. Kearsley. Chemical similarity using geometric atom pair descriptors. *Journal of Chemical Information and Computer Science*, 36:128–136, 1996.

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

[114] A. Smellie, S. Kahn, and S. Teig. Analysis of conformational coverage: 2. applications of conformational models. *Journal of Chemical Information and Computer Science*, 35:295–304, 1995.

[115] A. Smellie, S. Teig, and P. Towbin. Poling: Promoting conformational variation. *Journal of Computational Chemistry*, 16(2):171–187, 1995.

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

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

[118] G. M. Verkhivker, P. A. Rejto, D. K. Gehlhaar, and S. T. Freer. Exploring the energy landscapes of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of hiv-1 protease and fkbp-12 complexes. *Proteins*, 25:342–353, 1996.

[119] H. Wang and C. Levinthal. A vectorized algorithm for calculating the accessible surface area of macromolecules. *Journal of Computational Chemistry*, 12:868–871, 1991.

[120] W. Welsh, J. Ruppert, and A. Jain. Hammerhead: Fast fully automated docking of flexible ligands to protein binding sites. *Chemistry and Biology*, 3:449–462, 1996.

[121] R. A. Wiley and D. H. Rich. Peptidomimetics derived from natural products. *Medicinal Research Reviews*, 13:327–384, 1993.

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

[123] S. Wodak and J. Janin. Analytical approximation to the accessible surface area of proteins. *Proceedings of the National Academy of Science (USA)*, 77:1736–1740, 1980.