Efficient Database Screening for Rational Drug Design Using Pharmacophore-Constrained Conformational Search

Steven M. LaValle
Department of Computer Science
Iowa State University
Ames, IA, USA

Paul W. Finn
Pfizer Central Research
Sandwich, Kent, UK*

Lydia E. Kavraki
Department of Computer Science
Rice University
Houston, TX, USA

Jean-Claude Latombe
Department of Computer Science
Stanford University
Stanford, CA, USA

Abstract

Computational tools have greatly expedited the pharmaceutical drug design process in recent years. One common task in this process is the search of a large library for small molecules that can achieve both a low-energy conformation and a prescribed pharmacophore. The pharmacophore expresses constraints on the 3D structure of the molecule by specifying relative atom positions that should be maintained to increase the likelihood that the molecule will bind with the receptor site. This paper presents a pharmacophore-based database screening system that has been designed, implemented, and tested on a molecular database. The key ingredient in this system is a simple, randomized conformational search technique that attempts to simultaneously reduce energy and maintain pharmacophore constraints. This enables efficient identification of molecules in a database that are likely to dock with a given protein, which can serve as a powerful aid in the search for better drug candidates.

1 Introduction

The development of a pharmaceutical drug is a long, incremental process, typically requiring years of research and experimentation. The goal is to find a relatively small molecule (*ligand*), typically comprising a few dozen atoms, that docks with a receptor cavity in a specific protein; Figure 1 shows an illustration. Protein-ligand docking can stimulate or inhibit some biological activity, ultimately leading to the desired pharmacological effect. The problem of finding suitable ligands is complicated due to both energy considerations and the flexibility of the ligand. In addition to satisfying structural considerations, factors such as synthetic accessibility, drug pharmacology and toxicology greatly complicate and lengthen the search for the most effective drug molecules.

*Author's current address: Prolifix Ltd., 91 Milton Park, Abingdon, Oxfordshire, UK

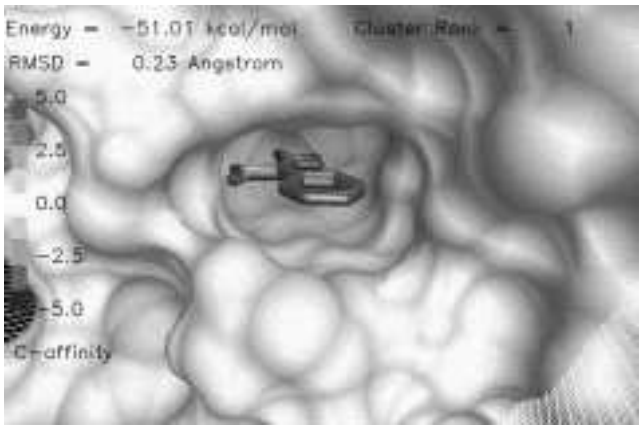

Figure 1: A 3D model of protein-ligand docking.

In their search for a new drug, chemists often construct a *pharmacophore*, which serves as a template for the desired ligand. The pharmacophore is expressed as a set of *features* that an effective ligand should possess and a set of *spatial constraints* among the features. The features can be specific atoms, centers of benzene rings, positive or negative charges, hydrophobic or hydrophilic centers, hydrogen bond donors or acceptors, and others. The spatial arrangement of the features represents the relative 3D placements of these features in the docked conformation of the ligand. The pharmacophore encapsulates a prevailing assumption in drug design that ligand binding is due primarily to the interaction of some features of the ligand to "complementary" features of the receptor. The interacting features are included in the pharmacophore and are key for searching for new drugs. The rest of the ligand atoms merely provide a scaffold for holding the pharmacophore features in their spatial positions. Figure 2 offers an illustration.

The Problem This paper deals with the following problem. Given a pharmacophore and a database of flexible ligands, identify those ligands that can achieve a low-energy spatial conformation that matches some of their features to the features of the pharmacophore. We will say that the selected ligands can 'satisfy' the pharmacophore.

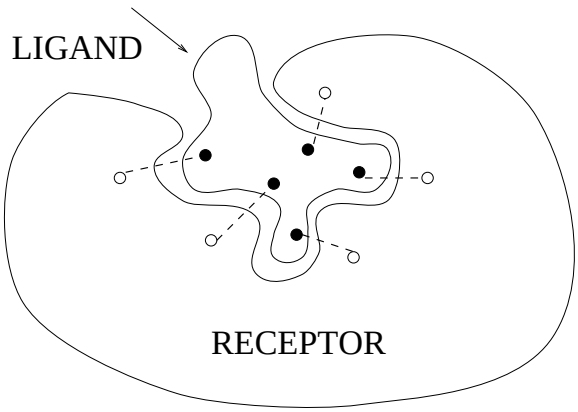

Figure 2: The black dots on the ligand represent the features of the pharmacophore. They interact with complementary features of the receptor.

We view the problem primarily as that of obtaining a constrained conformation of a known kinematic structure (the ligand). The 3D positions of certain parts of the structure are predetermined by the pharmacophore model. We compute the "folding" of the rest of the structure in a way that it preserves all the pharmacophore matches while respecting all structural constraints (i.e., bond lengths), all kinematic constraints (i.e., torsional degrees of freedom and their allowed values), and all energy constraints (i.e., the energy of the ligand should be below a threshold).

We model the kinematics of the ligand using techniques common in robotics. Although powerful analytical techniques exist for searching the solution spaces of similar structures in robotics, such techniques are impractical for handling high degrees of freedom. Our ligands have many torsional degrees of freedom; thus, we focus on randomized solutions to the conformational search problem. The need for efficiency has also motivated randomized search techniques in robotics [16, 17]. We use our experience with these methods to develop a randomized conformational search technique, which simultaneously reduces energy and maintains the pharmacophore constraints.

Significance of Our Work Major pharmaceutical companies maintain extensive databases of chemical compounds that have been synthesized in previous research efforts. If chemists are able to efficiently screen these databases for ligands that can satisfy the pharmacophore model, great amounts of effort and expense can be spared by exploiting the results of past experimental efforts [23]. In several cases, many properties of the ligands have been systematically documented in the database. In other cases, the search may reveal ligands with diverse chemical compositions that can still satisfy the same pharmacophore. The comparative analysis and modification of such ligands can lead to better drug candidates. An overview of a typical drug design cycle is given in Figure 3.

The identification of a pharmacophore is a challenging and speculative task and is beyond the scope of this paper. We will only mention that computational chemists construct pharmacophores both when the 3D structure of the receptor is known and when it is not. Several Three-Dimensional Quantitative Structure-Activity Relationships, or 3D QSAR [21] theories have been developed to capture some of the underlying chemical activity in a pharmacophore. If the 3D

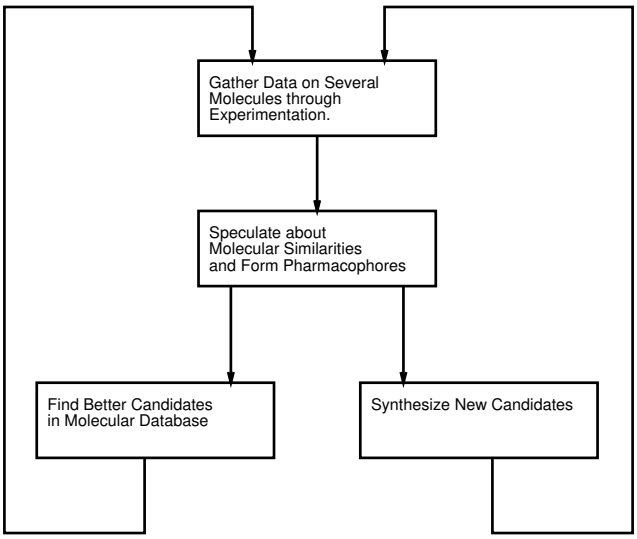

Figure 3: An overview of a typical drug design cycle.

structure of the docking site is known, the pharmacophore is based on the properties of atoms on or close to the surface of the receptor and the ligands that are known to dock in that site. Often the 3D structure of the receptor can not be obtained using techniques such as X-ray crystallography or NMR. In that case the only information available to the chemist is a set of molecules that interact with the specific receptor and hence exhibit the pharmacophore in their docked conformations. Each of these ligands, however, has many torsional degrees of freedom making the identification of the pharmacophore an extremely difficult task. Computational techniques that automatically construct a pharmacophore from a set of molecules have been developed [10, 13, 20, 21]. Both in the case of known and unknown receptor structure, efficient database screening techniques, such as the one presented in this paper, are powerful tools in the drug development process.

Related Work It is widely recognized that the simple problem of matching a single flexible molecule to a pharmacophore is a difficult problem which is currently poorly addressed [26]. Distance geometry, systematic search, randomized search, and genetic algorithms have been tried but have produced slow algorithms [1, 4, 5, 11]. One of the most efficient existing techniques for flexible matching is the "Directed Tweak Method" [15, 24]. The method minimizes a pseudo-energy 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 pharmacophore query. Unfortunately, the pseudo-energy function contains a large number of local minima, and conformations having high energy are frequently returned [5]. Our work differs from previous work in the sense that it rigorously treats the kinematics of the molecule while guiding the molecule into low energy conformations.

2 Problem Formulation

The Molecule Model A *molecule* is characterized by a pair (A, B), in which A represents a collection of *atoms*, and B represents a collection of *bonds* between pairs of atoms. An underlying graph can be considered for the molecule, in

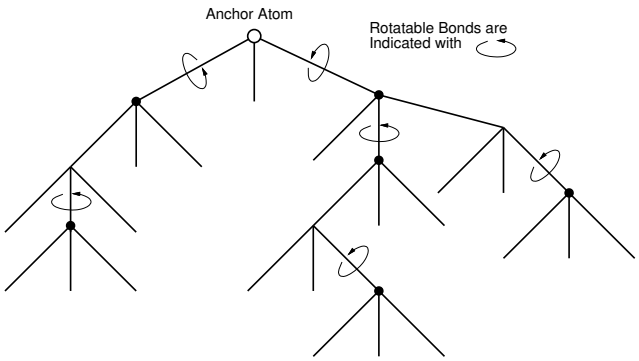

Figure 4: The molecule is considered as a tree that is rooted at the anchor atom. Some of the bonds are considered rotatable. For each rotatable bond, the structure below the black dot rotates about the bond's main axis.

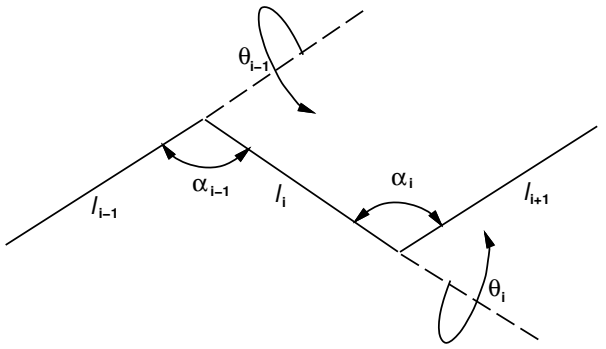

Figure 5: The assignment of α, l, and θ parameters along the kinematic chain.

which atoms represent vertices and edges represent bonds. Thus, usual graph-theoretic concepts such as connectedness, paths, trees, and cycles can be applied to molecules. It will be convenient to choose one atom, $a_{anch} \in A$, as the *anchor* for the molecule (or the root of the corresponding graph). Figure 4 shows an example. We assume that the underlying graph structure is a tree (i.e., no flexible rings). We represent rigid rings by considering the entire ring as a special "atom" that is attached by a "bond" to the rest of the molecule. Our assumption about rings remains valid for a large set of molecules that are of interest in drug design. Our general approach could be extended to cyclic molecules by exploiting computational algebra techniques that obtain kinematic solutions to cyclic chains [8], or by using techniques for large cyclic chains in [25].

Information used for kinematic and energy computations is associated with each of the atoms and bonds. Each atom carries standard information, such as its van der Waals radius. Three pieces of information are associated with each bond, $b_i \in B$: (i) the *bond length*, l_i; (ii) the *bond angle*, α_i, is the angle between b_i and the previous bond, in the direction toward a_{anch}; (iii) the set of possible *torsion angles*, $\theta_i \subseteq [0, 2\pi)$, which represents the ability of the bond to rotate about its own axis. The part of the molecule that is attached to b in the direction away from the anchor will also undergo rotation about this axis. If θ_i must remain constant, the bond is *fixed*; otherwise, it is considered *rotatable*. In most molecular studies [1, 26], bond lengths and bond angles are considered fixed, while torsions are allowed to vary. We follow this assumption in our work. From now on we represent the conformation of the ligand as m-dimensional vector of torsion angles $\boldsymbol{\theta}$, in which each component of $\boldsymbol{\theta}$ corresponds to a rotatable bond.

The Pharmacophore Model A pharmacophore is defined in terms of a finite set of features. Typically, there are between three and six features. A feature usually corresponds to an atom in A; however, there are other possibilities, such as the center of a rigid ring. Even "dummy atoms" can be defined, which are not strictly part of the molecule, but have positions that are determined by defining artificial bonds and atoms. These extensions can be useful in some pharmacophore models (i.e., for modeling hydrogen bond donors and acceptors). It will be assumed from this point onward that every feature is an atom. If other kinds of features are needed, the molecule can be appropriately extended using fictitious atoms and bonds.

The pharmacophore model also includes constraints on the relative positions between features. Suppose for convenience that one of the features is designated as a_{anch}, which lies at the origin of a global xyz coordinate system. For each remaining feature, the corresponding atom is constrained to lie near a specified position in this new coordinate system. Any set of torsion angles that is chosen for the molecule must place each of the feature atoms within a small neighborhood of its prescribed position. Another coordinate frame is attached to the molecule at a_{anch}. The molecule coordinate frame and the feature coordinate frame can become misaligned by rotations; this will be handled shortly. For the purpose of assigning bond angles for bonds that are attached to a_{anch}, assume that a_{anch} is attached to a fictitious "bond" that connects $(0, 0, -1)$ and $(0, 0, 0)$.

The Kinematic Model Molecular kinematics give the positions of all of the atoms of the ligands in terms of the torsion angles. The bond lengths, bond angles, and torsion angles can be conveniently used as parameters in the Denavit-Hartenburg representation for spatial kinematic chains [7, 14]. This representation is useful for determining the appropriate rigid-body transformation to apply to any link in a series of attached links. For the molecule, suppose that a local coordinate frame is attached at the beginning of each link (or atom center). If a bond b_i follows a bond b_{i-1} in the chain, then the coordinate frame of b_i is related to that of b_{i-1} by the homogeneous (both rotation and translation are performed) transformation

$$T_i = \begin{bmatrix} c\theta_i & -s\theta_i & 0 & 0 \\ s\theta_i c\alpha_{i-1} & c\theta_i c\alpha_{i-1} & -s\alpha_{i-1} & -l_i s\alpha_{i-1} \\ s\theta_i s\alpha_{i-1} & c\theta_i s\alpha_{i-1} & c\alpha_{i-1} & l_i c\alpha_{i-1} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (1)$$

in which θ_i represents the torsion angle, $c\theta_i$ is $cos\theta_i$ and $s\theta_i$ is $sin\theta_i$. Figure 5 depicts the quantities that appear in (1). The fictitious bond is used to define α_0. If b_i is not rotatable, then θ_i is a constant; otherwise, θ_i is a conformation parameter, included in $\boldsymbol{\theta}$.

The position of any atom in the molecule can be determined by chaining matrices of the form (1). For example, suppose $b_i, b_{i-1}, \ldots, b_1$ represents the sequence of bonds in the path from a particular atom, $a \in A$, to a_{anch}. The xyz

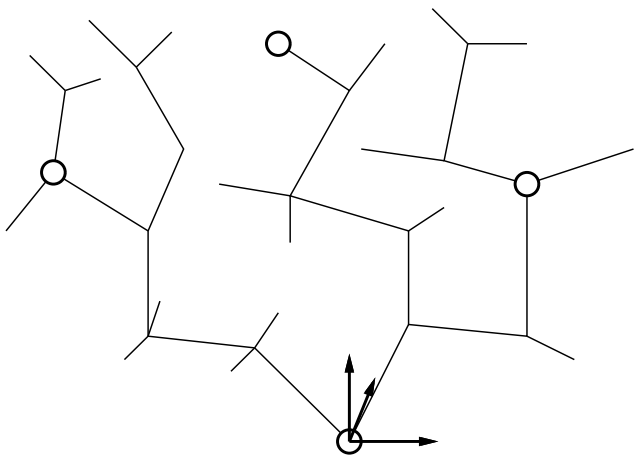

Figure 6: Four features are shown for a hypothetical molecule. The coordinate frame is attached to one of the features, which is designated as a_{anch}. The anchor atom, a_{anch}, must be allowed to rotate to preserve the original freedom of the molecule.

position of a is given by

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = T_1 T_2 \cdots T_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{2}$$

Before the expression of the kinematics is complete, there is one additional transformation that must be defined because we assume, without loss of generality, that a_{anch} is a feature. The definition of the features requires that the position of a_{anch} is at the origin, but it does not impose any constraints on the orientation of a_{anch}. Thus, it is possible that a coordinate frame attached to the molecule could be rotated with respect to the global coordinate frame for the pharmacophore feature positions. It is therefore necessary to allow the frame attached to the molecule at a_{anch} to achieve any orientation with respect to the global frame. See Figure 6.

The space of 3D rotations can be parameterized using Euler angles, γ, ϕ, ψ, for which $0 \leq \gamma < \pi$, $0 \leq \phi < 2\pi$, and $0 \leq \psi < 2\pi$ [7]. The parameterized rotation matrix can be placed into homogeneous form by extending one row and column to obtain

$T_R(\gamma, \phi, \psi) =$

$$\begin{bmatrix} c\phi s\gamma c\psi & -c\phi c\gamma s\psi - s\phi c\psi & c\phi s\gamma & 0 \\ s\phi c\gamma c\psi + c\phi s\psi & -s\phi c\gamma s\psi + c\phi c\psi & s\phi s\gamma & 0 \\ -s\gamma c\psi & s\gamma s\psi & c\gamma & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Taking into account the anchor orientation, the position of a particular atom, $a \in A$, at the end of a path, $b_i, b_{i-1}, \ldots, b_1$, to a_{anch} is given by

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = T_R(\gamma, \phi, \psi) T_1 T_2 \cdots T_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{4}$$

The Kinematic Error Function Using the kinematic expressions, an error function can be specified to express how closely the pharmacophore constraints are satisfied. The transformation (4) is expressed in terms of several constants and variables. The constants are the bond angles, α_i, bond lengths, l_i, and torsion angles for non-rotatable bonds. The variables are γ, ϕ, ψ, and the torsion angles for rotatable bonds which are given in $\boldsymbol{\theta}$.

Suppose that there are N features in the pharmacophore model. Let G^k denote the prescribed position for the k^{th} feature, for each $k \in \{0, 1, \ldots, N-1\}$. Let $G_0 = (0, 0, 0)$ represent the feature that corresponds to a_{anch}. Let $g^k(\boldsymbol{\theta}, \gamma, \phi, \psi)$ represent the xyz position of the atom that corresponds to the k^{th} feature. This position is expressed only in terms of the variables in the kinematic formulation (4).

Given $\boldsymbol{\theta}$, γ, ϕ, and ψ, the total amount of error between the prescribed feature positions and the actual feature positions can be measured as

$$d(\boldsymbol{\theta}, \gamma, \phi, \psi) = \sum_{i=1}^{N-1} \| G^i - g^i(\boldsymbol{\theta}, \gamma, \phi, \psi) \|, \tag{5}$$

in which $\| \cdot \|$ denotes the Euclidean norm.

The Energy Function In a sense, the energy function measures the likelihood that the molecule will achieve a conformation in nature (lower energy states are more likely to occur). It is common in molecular modeling [2] to use an empirical energy function; we use the SYBYL system (Tripos Inc [24]) energy function:

$$\begin{aligned} e(\boldsymbol{\theta}) = \quad & \textstyle\sum_{bonds} \frac{1}{2} K_b (R - R')^2 + \sum_{ang} \frac{1}{2} K_a (\alpha - \alpha')^2 + \\ & \textstyle\sum_{torsions} K_d [1 + \cos(n\theta - \theta')] + \\ & \textstyle\sum_{i,j} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \right\}. \end{aligned} \tag{6}$$

In the above, the first sum is taken over all bonds, the second over all bond angles, the third over all rotatable bonds, and the last sum of is taken over all pairs of atoms. K_b, K_a, and K_d are force constants, ϵ is the dielectric constant, and n is a periodicity constant. R, α, and θ are the measured values of the bond lengths, bond angles, and torsional angles in conformation $\boldsymbol{\theta}$, while R', α', and θ' are equilibrium (or preferred) values for these bond lengths, bond angles, and torsional angles. r_{ij} measures the distance of atom centers in $\boldsymbol{\theta}$. 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.

The expression for the energy of a molecule may appear quite complicated, especially due to interactions between each pair of atoms; however, the energy depends only on the conformation, $\boldsymbol{\theta}$, because the anchor orientation parameters are only used for pharmacophore purposes. Notice also that the first and the second term of the energy function are constant with our assumptions.

The General Task The general task is to find conformations that satisfy both pharmacophore and energy considerations. For a given molecule in the database and a given pharmacophore, two important questions are asked

- **Question 1:** Can this molecule achieve a low-energy conformation that satisfies the given pharmacophore?

- **Question 2:** What are the distinct low-energy conformations that satisfy the pharmacophore?

Question 1 decides whether the molecule is worth considering as a candidate in the drug design process. Using the concepts defined in this section, this question can be formulated as determining whether there exist values for $\boldsymbol{\theta}$, γ, ϕ, and ψ, such that $d(\boldsymbol{\theta}, \gamma, \phi, \psi)$ and $e(\boldsymbol{\theta})$ are below some fixed thresholds.

The answer to Question 2 provides additional information for chemists. There might be several low-energy conformations that satisfy the pharmacophore, but each could place the other non-feature atoms in very different locations. These distinct conformations may be used to refine a pharmacophore model. A new, hypothesized feature might only be satisfied in some of these conformations. By looking at multiple conformations for several molecules, it might be possible to select a single conformation from each molecule that also satisfies the new feature. The main difficulty in providing a set of distinct conformations is that the notion of "distinct" is not well-defined. The difference between two conformations can be compared by defining a metric, such as the RMS distance of the atomic displacements between the two configurations. In Section 4, a clustering technique is described that ensures that the reported clusters sufficiently differ according to this metric.

3 Randomized Conformational Search with Constraints

The Approach In this section we focus our attention on Question 1 above. In Section 4 we show how to integrate the approach with a database screening system that can answer both Questions 1 and 2. Figure 7 indicates the three main steps involved in the computation. The first step generates a random conformation, $\boldsymbol{\theta}$, and anchor orientation, given by γ, ϕ, and ψ. The second step attempts to reduce the kinematic error (5) by performing a randomized gradient descent. If the second step is successful, the third step is reached; otherwise, the first step is repeated. The third step attempts to reduce the energy while keeping the kinematic error within an acceptable range. If the energy falls below a prescribed threshold, then the method reports success. Otherwise, the first step is repeated. In practice, a limit is set of the maximum number of allowable failures before the algorithm terminates. It is important to note that if the algorithm terminates after any number of failures, it cannot be concluded with certainty that the molecule does not admit a low-energy conformation that satisfies a given pharmacophore. This is the tradeoff that is typically made for an efficient, randomized algorithm, as opposed to a costly, systematic approach that carefully considers all possibilities. Randomized techniques are popular for conformational search problems that do not involve pharmacophore constraints; see for example [3, 4, 9, 12, 22].

An alternative way to approach the constrained search problem would be to obtain an explicit characterization of the set of conformations that satisfy the pharmacophore. This generally involves characterizing the solutions to the inverse kinematics problem. The equations of the form (4) that express the pharmacophore can be converted to a polynomial system. Each trigonometric function can be replaced by a ratio of polynomials by using stereographic projection, and the problem conformations that satisfy the equations lie in an algebraic variety [6]. Efficient elimination techniques from computational algebraic geometry have been developed

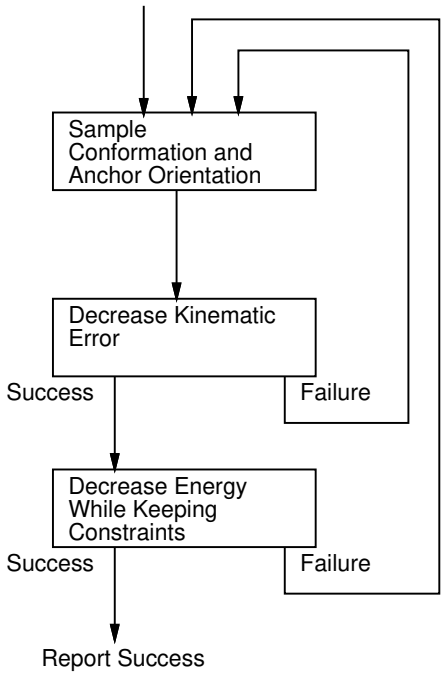

Figure 7: An overview of constrained conformational search.

for numerating the solution set for problem in which there are a finite number of solutions [8, 19]. For the problem discussed in this paper, the system of equations is generally underconstrained, which leads to a complicated, multi-dimensional solution set. Although it is straightforward to obtain a parametric representation of some algebraic varieties (such as a sphere or a torus), it not generally possible to employ elimination techniques to find a parameterization of any algebraic variety [6]. This consideration, and the need for efficiency, led to the choice of a numerical, randomized technique, as opposed to performing symbolic computations with polynomial systems.

Distance Minimization The algorithms for reducing the kinematic error and for decreasing the energy are shown in Figures 8 and 9, respectively. The first algorithm, DECREASE_KINEMATIC_ERROR, resembles the minimization technique used in the UNITY-3D package (which is included in SYBYL, from Tripos Inc. [24]). The algorithm accepts an initial conformation and anchor orientation parameters, and iteratively adjusts these until the total kinematic error falls below some acceptable limit, d_{tol}. Our discussion in Section 2 shows how to compute the kinematic error d. In each iteration, the RANDOM_NEIGH function slightly perturbs each of the conformation and orientation parameters. The size of this neighborhood was determined experimentally in advance. There are two ways in which this algorithm can fail: (i) the counter k reaches its maximum value, k_{max}, which means that too many iterations have been executed, or (ii) the counter f reaches its maximum value f_{max}, which means that too many failures to reduce the error have occurred. This second failure essentially detects that the minimization is trapped in a local minimum. Rather than try to escape this minimum, a new conformation is sampled.

DECREASE_KINEMATIC_ERROR($\boldsymbol{\theta}, \gamma, \phi, \psi$)

```
1   k ← 0;  f ← 0;  d_min ← ∞;
2   while k < k_max and f < f_max and d_min > d_tol do
3       (θ', γ', φ', ψ') ←RANDOM_NEIGH(θ, γ, φ, ψ);
4       if d(θ', γ', φ', ψ') < d_min then
5           f ← 0;  d_min ← d(θ', γ', φ', ψ');
6           (θ, γ, φ, ψ) ← (θ', γ', φ', ψ');
7       else
8           f ← f + 1;
9       k ← k + 1;
10  if d_min ≤ d_tol then
11          Return (θ, γ, φ, ψ)
12  else
13          Return FAILURE
```

Figure 8: This algorithm iteratively attempts to reduce the kinematic error.

DECREASE_ENERGY($\boldsymbol{\theta}, \gamma, \phi, \psi$)

```
1   k ← 0;  f ← 0;  g ← 0;  d_min ← ∞;  e_min ← ∞;
2   while (k < k_max and f < f_max
          and g < g_max and e_min > e_thr) do
3       (θ', γ', φ', ψ') ←RANDOM_NEIGH(θ, γ, φ, ψ);
4       if d(θ', γ', φ', ψ') ≤ d_tol then
5           f ← 0;  e_min ← ∞;
6           if e(θ', γ', φ', ψ') < e_min then
7               g ← 0;  e_min ← e(θ', γ', φ', ψ');
8               (θ, γ, φ, ψ) ← (θ', γ', φ', ψ');
9           else
10              g ← g + 1;
11      else
12          f ← f + 1;
13      k ← k + 1;
14  if e_min ≤ e_thr then
15          Return (θ, γ, φ, ψ)
16  else
17          Return FAILURE
```

Figure 9: This algorithm iteratively attempts to reduce the energy while keeping the kinematic error within tolerance limits.

Energy Minimization If the kinematic error is successfully reduced, the resulting conformation and base orientation are passed to DECREASE_ENERGY. This algorithm proceeds in the same manner as the previous one, except that two different criteria must be monitored. It attempts to reduce the energy, while ensuring that the kinematic error distance, d, does not increase beyond d_{tol}. For this algorithm there are three counters, each of which can halt the algorithm if its limit is reached. The counter k records the total number of iterations, f records the number of consecutive failures to maintain the pharmacophore without reducing energy, and g records the number of consecutive failures to reduce the energy. e_{thr} is the desired energy threshold.

This algorithm exploits that fact that some tolerance is allowed for matching the pharmacophore. Figure 10(a) illustrates that a perturbation of the conformation and anchor orientation parameters only slightly moves the feature atoms. If each of the feature atoms remains within acceptable tolerance, then the new parameters are evaluated based on energy. The speed of the algorithm is directly related

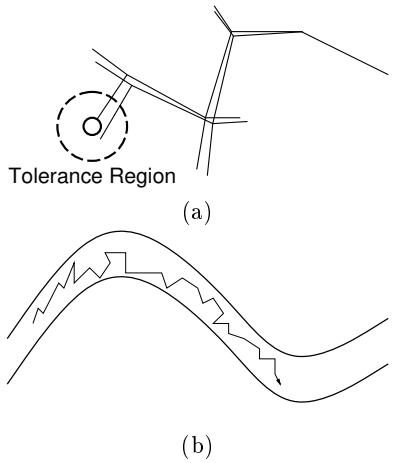

Figure 10: a) After perturbing the molecule, it must be determined whether the feature atoms remain within acceptable tolerance from their prescribed positions; b) the search technique can be considered as a randomized traversal inside of a thick surface.

to the feature tolerance. If the tolerance is larger, then larger variations can be considered in RANDOM_NEIGH. This causes the molecule to move more quickly toward an energy minimum. As shown in Figure 10(b), the "thickness" of the set of acceptable parameters greatly facilitates a randomized traversal. In the current implementation the neighborhood size in RANDOM_NEIGH is an empirically-chosen constant.

Note that the choice of anchor atom prohibits the specification of a tolerance for the feature at the origin. This can be overcome by defining three positional offset variables for the anchor (in addition to the rotation parameters); however, this would cause some loss of computational performance.

4 An Integrated Database Screening System

A database screening system has been implemented and tested on a small molecular database. The principle module in this system is the constrained conformational search technique described in Section 3.

Iterative Sweeping Across the Database Suppose that one would like to use the approach described in Figure 7 to search a database for molecules that have a low-energy conformation that satisfies a given pharmacophore. Initially, non-geometric constraints can be used to quickly discard any molecules that do not contain all of the feature atoms. Our work pertains to the remaining set of molecules. The primary question to answer is how many iterations of our approach should be applied to each molecule? For a single molecule, each time a new sample conformation and anchor orientation is chosen and fails to lead to success, the likelihood that the molecule will ever succeed is decreased. However, after any number of iterations, it is impossible to conclude that the molecule will never succeed.

One sensible way to use this method is to sweep across the database, using only one iteration of the approach in Figure 7 for each molecule. Once the last molecule has been

	Themolysin Inhibitors		Ace Inhibitors	
Molecule	Atoms	Rot. Bonds	Atoms	Rot. Bonds
mol1	69	10	48	8
mol2	66	11	50	8
mol3	22	3	31	7
mol4	42	8	47	8
mol5	64	13	45	6
mol6	63	12	30	3

Figure 11: The thermolysin and ace inhibitors used in our experiments.

processed, a second sweep can be made across the database. The sweeps across the database can be repeated until some predetermined criteria are met (for example, a certain number of successes have been found or the maximum number of sweeps has been reached). Each iteration of this method is completely independent; therefore, there is no difficulty in switching frequently between molecules. Furthermore, the whole process can be easily parallelized. Because the likelihood that a molecule will succeed decreases with each iteration, it makes sense to avoid focusing multiple iterations on a single molecule before continuing.

The result is an "any-time" algorithm in the sense that the solutions gradually improve over time, and there is no natural termination point. Typically, the first sweep might turn up a small number of successful molecules. These results could be analyzed while the program continues to identify other successful molecules over time. If the goal is only to identify successful molecules, then a successful molecule is removed from the search set before the next sweep. The computation will gradually focus in this case on the more difficult molecules.

Conformation Clustering The goal might alternatively be to characterize the set of possible solutions for each of the successful molecules. (This is Question 2 from Section 2.) In this case, clustering can be performed incrementally for each solution that is generated [18]. A metric $m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ can be defined that quantifies the difference between two conformations. In the database screening system, $m(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is defined as the RMS of the displacements of the atoms between the two conformations. A threshold, m_{thr}, is set as the maximum distance allow for two different conformations to be considered as part of the same cluster.

The incremental clustering approach proceeds are follows. The "Report Success" step in Figure 7 is replaced by an operation that updates the cluster record using the new conformation (the anchor rotation is also retained). Each cluster is represented by a single conformation that has the lowest energy compared with any other known conformation within distance m_{max}. If the cluster record is empty, then the first update generates a single cluster for the given conformation. Suppose the cluster record contains several conformations when an update is requested for a new conformation. If there are any other conformations that are within a distance m_{max} that have lower energy, then the new conformation is discarded (a better, similar conformation already exists). Otherwise, the new conformation is added to the cluster record. Any existing clusters that are within a distance m_{max} and have higher energy are deleted. This has the effect of making the "representative" of each cluster the conformation that has the least energy.

Experiments The database screening system was implemented in Gnu C on an SGI Indy and on a Pentium Pro 200Mhz PC running Linux. A database of molecules was provided from Pfizer Limited. This database includes 6 different inhibitors of thermolysin and 6 different inhibitors of ace. The table in Figure 11 reports the number of atoms and the number of rotatable bonds for each of these molecules. All molecules are very flexible as they contain 3 to 13 torsional degrees of freedom. The inhibitors of thermolysin are shown in Figure 12, and the inhibitors of ace are shown in Figure 13. Each molecule appears in a random configuration in our database. For both sets of molecules, the docked conformations are known by previous pharmacophore identification studies. We use subsets of the true pharmacophore as queries in our database for testing purposes.

For the thermolysin inhibitors we tried a query with 4 features of the known pharmacophore and for the ace inhibitors we run a query with 3 features of the pharmacophore. In both cases we let our program complete 20 iterations and set the cluster distance m_{max} to 1.5 Å. The maximum feature tolerance was set to 0.5 Å. It was generally found that a single iteration of the method in Figure 7 could be performed in about 5-10 seconds. A sufficient clustering record for a single molecule required about 5-20 minutes. Our results are shown in Table 14 for the thermolysin inhibitors and in Table 15 for the ace inhibitors. Column 1 of these tables shows our molecules. We report in columns 2 and 3 the number of clusters found and the minimum energy conformation in all of these clusters as a proof of the fact that the conformations that we find are indeed of low energy. In general, the energies of the conformations found are within 2-7 Kcals/Mol of the energy of the known docked conformation.

Our clustering scheme ensures that we record conformations that are significantly different from each other. In Figure 16 we show the representatives of the clusters that are generated for mol5 of the thermolysin inhibitors. The pharmacophore features are circled in white. Notice how well these are matched in contrast with the rest of the atoms in the molecule. Clusters for mol5 of the ace inhibitors are shown in Figure 17. A similar observation holds here.

Our implementation allows us to answer the two questions posed at the end of Section 2. Question 1 is answered efficiently by our randomized constrained minimization technique. Question 2 is answered by iterating our constrained minimization and by clustering the results. In each iteration of the algorithm a new random conformation of the molecule is generated and subsequently minimized. As discussed at the end of Section 2, the position of non-feature atoms may help the chemist select among possible ligands or modify existing ligand in search for a better drug.

Our previous work with randomized techniques has shown [10, 16] that if we continue iterating our algorithm we increase our chances of covering the conformational space of the molecule and hence our chances of providing exhaustive information about the constrained conformations of the molecule. As a proof of concept, in the examples used in this section we obtained conformations that are fairly close to the known docked conformations. For the 5 thermolysin inhibitors, we obtained conformations whose RMS distances from the corresponding docked conformation were 0.50, 2.96, 0.59, 0.81, 2.40, 2.56 Å correspondingly. In the ace case, we obtained conformations whose RMS distances from the corresponding docked conformation were 1.26, 1.79, 0.94, 2.03, 1.87, 1.98 Å. Notice that the more features we use, the closer we can get to the docked conformation (4 features

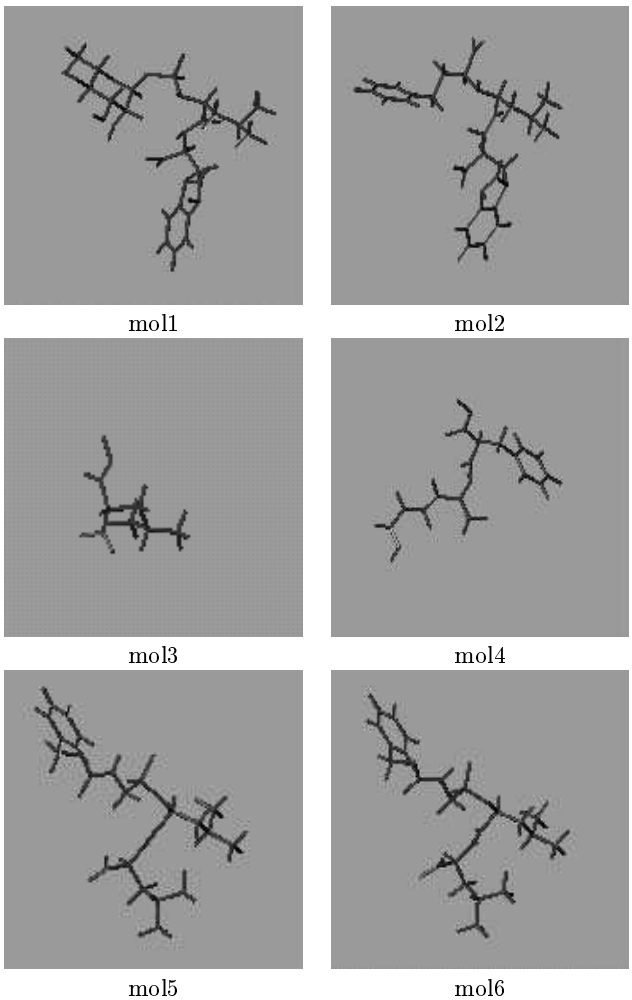

Figure 12: The inhibitors of theormolysin used. Each molecule is at a random conformation.

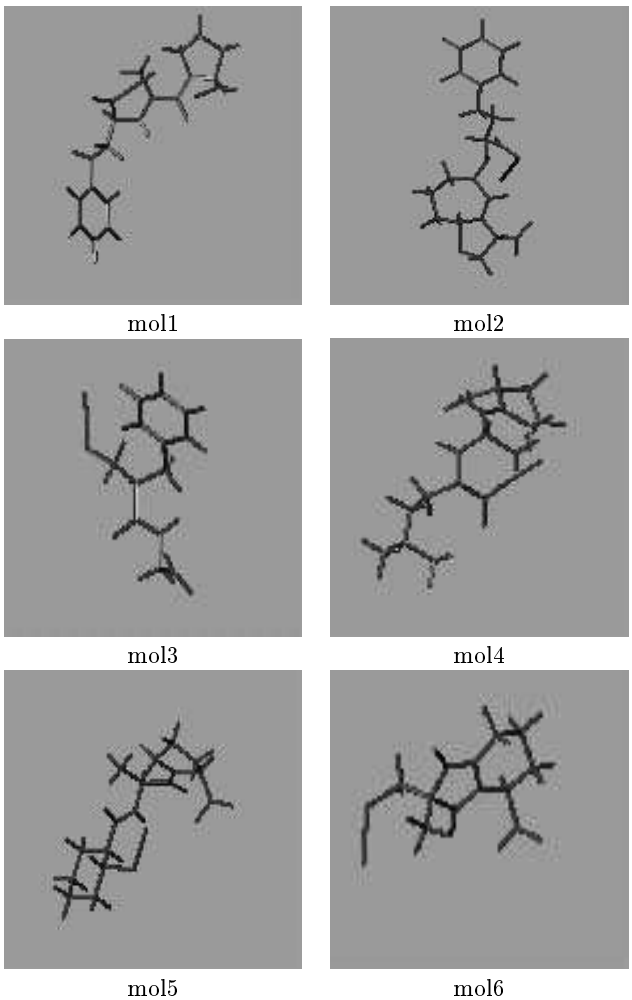

Figure 13: The inhibitors of ace used. Each molecule is at a random conformation.

for the thermolysin inhibitors versus 3 features for the ace inhibitors). In certain cases, we get very close to the docked configuration (certain thermolysin inhibitors), which is an indication of good conformational space coverage. Without doubt, more extensive experiments are needed to fully evaluate the system, but our preliminary results are encouraging.

5 Discussion

A database screening system has been presented that can help expedite the drug design process. The system identifies molecules that are able to satisfy a given pharmacophore, and are therefore reasonable candidates for further investigation. The key to this screening system is a randomized conformational search approach that considers both the kinematic error imposed by the pharmacophore constraints and the energy. The simplicity and efficiency of the approach should enable straightforward extensions to other classes of molecules, such as those with flexible rings.

A difficulty with the current approach is the selection of the step size for the random neighborhoods in the kinematic error descent and the energy descent. Although the same value was used over a wide variety of molecules, it seems that performance can be greatly improved by giving more careful attention to this selection. For example, larger step sizes might be appropriate if the conformation is near a kinematic singularity. This could compensate for the fact that large displacements near a singularity lead to small displacements of the feature atom. A quaternion parameterization of the anchor orientation, instead of a Euler angle parameterization, might also improve performance for similar reasons. Performance improvement might also be obtained by constraining random neighborhood samples to lie in the tangent space to the constraints.

Acknowledgments

This work started in the context of a grant with Pfizer Limited and Stanford University. At that time, Dr. Paul Finn was with Pfizer Limited and Steve LaValle was at Stanford University. This work was continued and completed under different funding. Steve LaValle is currently funded by an NSF CAREER Award. He can be contacted via e-mail at lavalle@iastate.edu. Lydia Kavraki is currently

Thermolysin Inhibitors	Num of Clusters	Min energy found (Kcal/Mol)
mol1	17	13.44
mol2	14	20.30
mol3	3	20.94
mol4	7	25.44
mol5	12	27.72
mol6	14	26.84

Figure 14: The results of a 4-feature thermolysin search.

Ace Inhibitors	Num of Clusters	Min energy found (Kcal/Mol)
mol1	13	42.42
mol2	11	43.12
mol3	15	37.88
mol4	13	42.25
mol5	5	39.40
mol6	2	36.38

Figure 15: The results of a 3-feature ace search.

partially supported by NSF CAREER Award IRI-970228 and SA1728-21122NM. The authors are especially grateful to David Hsu, Rajeev Motwani, Suresh Venkatasubramanian for many useful discussions.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms.* Springer-Verlag, Berlin, 1992.

[7] J. J. Craig. *Introduction to Robotics.* Addison-Wesley, Reading, MA, 1989.

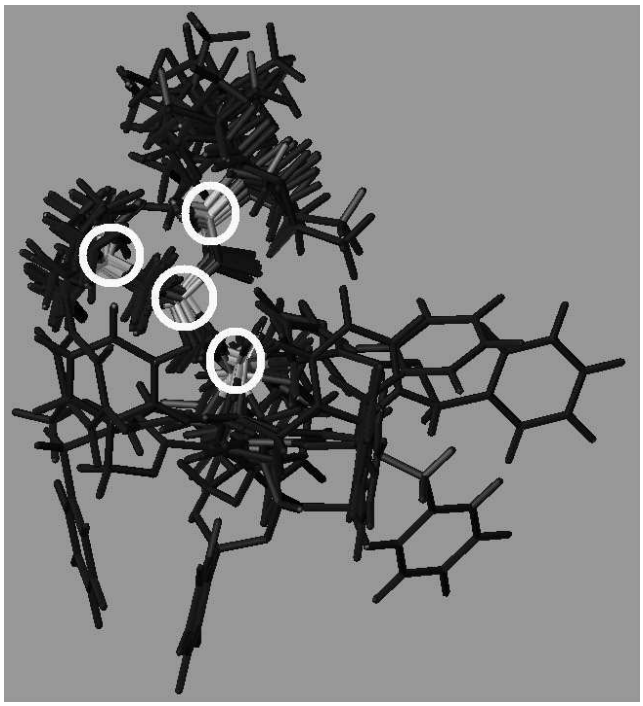

Figure 16: 12 clusters were obtained for mol5 of the thermolysin inhibitors.

[8] I. Z. Emiris and B. Mourrain. Computer algebra methods for studying and computing molecular conformations. Technical report, INRIA, Sophia-Antipolis, France, 1997.

[9] E. Fadrna and J. Koca. Single-coordinate-driving method coupled with simulated annealing. an efficient tool to search conformation space. *J. Phys. Chem. B*, 101:7863–7868, 1997.

[10] P. W. Finn, D. Halperin, L. E. Kavraki, J.-C. Latombe, R. Motwani, C. Shelton, and S. Venkatasubramanian. Geometric manipulation of flexible ligands. In *Applied Computational Geometry (Lecture Notes in Computer Science, 1148)*, pages 67–78. Springer-Verlag, Berlin, 1996.

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

[12] 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.

[13] A. K. Ghose, M. E. Logan, A. M. Treasurywala, H. Wang, R. C. Wahl, B. E. Tomczuk, M. R. Gowravaram, E. P. Jaeger, and J. J. Wendoloski. Determination of pharmacophoric geometry for collagenase inhibitors using a novel computational method and its verification using molecular dynamics, NMR, and X-ray crystallography. *J. Am. Chem. Soc.*, 117:4671–4682, 1995.

[14] R. S. Hartenburg and J. Denavit. A kinematic notation for lower pair mechanisms based on matrices. *J. Applied Mechanics*, 77:215–221, 1955.

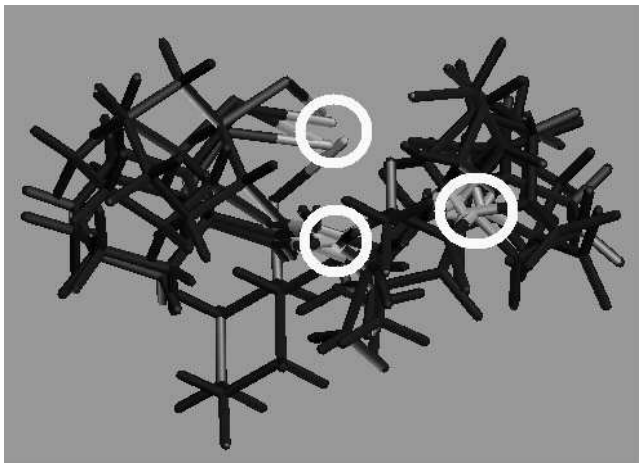

Figure 17: 5 clusters were obtained for mol5 of the ace inhibitors.

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

[16] L. E. Kavraki. *Random Networks in Configuration Space for Fast Path Planning.* PhD thesis, Stanford University, 1994.

[17] J.-C. Latombe. *Robot Motion Planning.* Kluwer Academic Publishers, Boston, MA, 1991.

[18] A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Computational Chemistry*, 13(6):730–748, 1992.

[19] D. Manocha and J. Canny. Real time inverse kinematics of general 6R manipulators. In *IEEE Int. Conf. Robot. & Autom.*, pages 383–389, Nice, May 1992.

[20] Y. C. Martin, M. G. Bures, E. A. Danaher, and J. DeLazzer. New strategies that improve the efficiecy of the 3D design of bioactive molecules. In C.-G. Wermuth, editor, *Trans in QSAR and Molecular Modeling*, pages 20–27. ESCOM, Leiden, 1993.

[21] T. I. Oprea and C. L. Waller. Theoretical and practical aspects of three-dimensional quantitative structure-activity relationships. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, pages 127–182. John Wiley and Sons, New York, 1997.

[22] 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.

[23] N. F. Sepetov, V. Krchnak, M. Stankova, S. Wade, K. S. Lam, and M. Lebl. Library of libraries: Appraoch to synthetic combinatorial library design and screening of "pharmacophore" motifs. *Proc. Natl. Acad. Sci. USA*, 92:5426–5430, June 1995.

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

[25] C. Wang. An efficient algorithm for conformational search of macrocyclic molecules. *J. Computational Chemistry*, 13(6):730–748, 1992.

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