A Randomized Kinematics-Based Approach to Pharmacophore-Constrained Conformational Search and Database Screening

STEVEN M. LAVALLE,[1] PAUL W. FINN,[2,*] LYDIA E. KAVRAKI,[3] JEAN-CLAUDE LATOMBE[4]

[1]*Department of Computer Science, 204 Atanasoff Hall, Iowa State University, Ames, Iowa 50011*
[2]*Pfizer Central Research, Sandwich, Kent, United Kingdom*
[3]*Department of Computer Science, Rice University, Houston, Texas*
[4]*Computer Science Department, Stanford University, Stanford, California*

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 article presents a pharmacophore-based database screening system that has been designed, implemented, and tested on a molecular database. The primary focus of this article is on a simple, randomized conformational search technique that attempts to simultaneously reduce energy and maintain pharmacophore constraints. This enables the 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. © 2000 John Wiley & Sons, Inc. J Comput Chem 21: 731–747, 2000

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. 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.

In the search for a new drug, a pharmacophore often 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.

THE PROBLEM

This article 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.

We view the problem primarily as that of obtaining a constrained conformation of a known kine-matic 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.[1, 2] Although powerful analytical techniques exist for searching the solution spaces of similar structures in robotics,[3] 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.[4, 5] We use our experience with these methods to develop a randomized conformational search technique, which simultaneously reduces energy and maintains the pharmacophore constraints.

In pharmaceutical chemistry, a variety of pharmacophore models and representations have been considered, and most of the current research addresses the development, evaluation, and selection of these models and representations. For example, as opposed to the 3D feature-matching model described so far, other possibilities include matching and agglomerating molecular fragments, and also defining enrichment factors that assess the similarity of entire known inhibitors. Instead of dealing with these issues, our work focuses primarily on the development of computational techniques that improve our understanding of geometric issues involved in drug design. For this reason, we chose to develop and demonstrate our methods for a particular pharmacophore model and representation. We believe that the techniques from this article can be adapted to other kinds of pharmacophores; however, such issues are beyond the scope of this article because they generally depend on the particular pharmaceutical application.

SIGNIFICANCE OF OUR WORK

Our work is expected to improve the ability to screen pharmaceutical databases for ligands that

can satisfy the pharmacophore model by delivering computational techniques that reason about the complicated geometric and kinematic constraints that arise from the matching problem. The techniques can be applied (in combination with other filters) to 3D structural databases, such as the Cambridge Structural Database, the Brookhaven Protein Databank, or proprietary pharmaceutical databases. As methods improve for screening databases, great amounts of effort and expense can be spared by exploiting the results of past experimental efforts.[6] 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. In the area of combinatorial chemistry, our techniques could be employed to assist the design of databases by screening enumerated structures (if pharmacophores are known).

The identification of a pharmacophore is a challenging and speculative task, and is beyond the scope of this article. Pharmacophore models are constructed for cases in which the 3D structure of the receptor is known and in which it is not. Several three-dimensional quantitative structure–activity relationships, or 3D QSAR[7] theories have been developed to capture some of the underlying chemical activity in a pharmacophore. If the 3D 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 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.[7–10] Both in the case of known and unknown receptor structure, database screening techniques, such as the one presented in this article, are powerful tools in the drug development process.

RELATED WORK

For over two decades, the drug degign process has been aided by database-searching techniques that perform matching based on 3D structure. A recent survey of database-searching techniques that consider flexible molecules appears in ref. 11. The earliest systems performed static matching of single conformations to queries.[12] There has been much interest in recent years to consider conformational flexibility in performing matches, which was first considered in ref. 13. There are generally two types of approaches to database-searching that consider conformational flexibility: similarity-based methods and pharamcophore-based methods. Many similarity-based approaches use superposition, such as the FlexS system, which computes a scoring function, and candidates are ranked and selected on the basis of enrichment factors.[14]

Pharmacophore-based methods represent the most common searching approach, and a continuing source of many computational challenges.[15] In addition to conformational flexibility, pharmacophores allow flexibility in the query through the specification of geometric features that only partially characterize the active site.[16] A wide variety of conformational search techniques have been developed and used with these methods,[17] including systematic search,[18] distance geometry,[19, 20] randomized search,[21–25] and genetic algorithms.[26, 27]

Several approaches to pharmacophore-based search consider conformational flexibility while the query is processed (as opposed to storing multiple conformations of each molecule).[28, 29] For example, the "Directed Tweak Method"[28, 30] minimizes a pseudoenergy function that 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 pseudoenergy function contains a large number of local minima, and conformations having high energy are frequently returned.[17] 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.

Problem Formulation

THE MOLECULE MODEL

A molecule is characterized by a pair (A, B), in which A represents a collection of atoms, and B
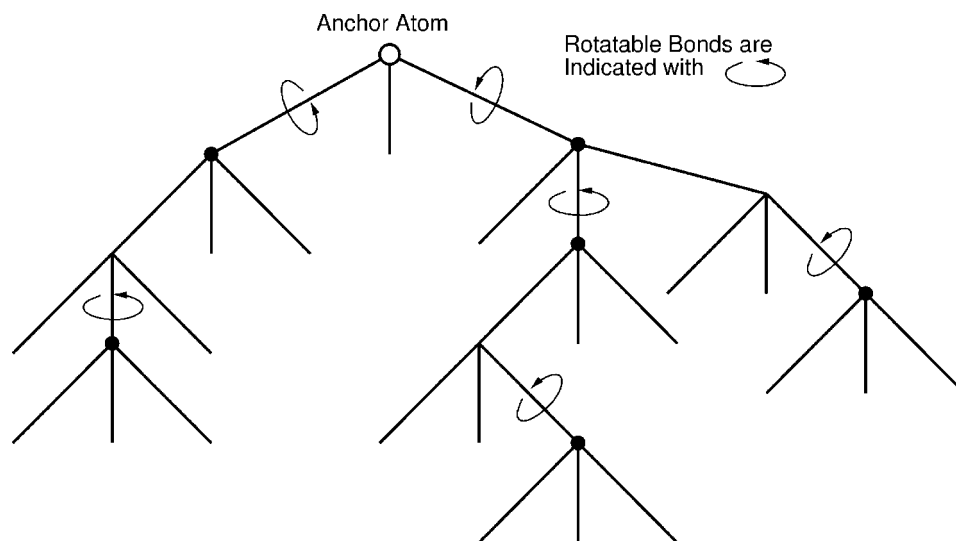
FIGURE 1. The molecule is considered as a tree that is rooted at the anchor atom. Some of the bonds are considered rotatable. For each rotatable bond, the structure below the black dot rotates about the bond's main axis.

represents a collection of bonds between pairs of atoms. An underlying graph can be considered for the molecule, in which vertices represent atoms 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, $a_{\text{anch}} \in A$, as the anchor for the molecule (or the root of the corresponding graph). Figure 1 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,[31] by using techniques for large cyclic chains as in ref. 32, or by breaking bonds and adding dummy constraints.

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,[15, 19] 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. In our system, a feature corresponds to an atom in A; however, it is straightforward to extend our implementation to incorporate other features, such as the center of a rigid ring. In general, the kinematic models, given in (1)–(4), can be extended to include "dummy atoms" by defining artificial bonds and atoms. These extensions are needed for some common pharmacophore models (e.g., for modeling hydrogen bond donors and acceptors). The result is an extended molecule, which can be handled using the same techniques as for the original molecule. The kinematic equations can be applied to the extended molecule, and our search methods will work on the extended feature set. Conformations for the extended molecule will lead directly to conformations of the original molecule. The conformation of the original molecule would be used for energy computations, and the conformation of the extended molecule would

be used for kinematics computations. In the remainder of this article we forego the consideration of an extended molecule, and assume that every feature corresponds to an atom in the original molecule; however, our general approach can be extended to consider additional types of features by replacing the original molecule with an extended molecule.

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.[1, 2] 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 2 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 position of a is given by

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = T_1 T_2 \ldots T_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \quad (2)$$

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, there-
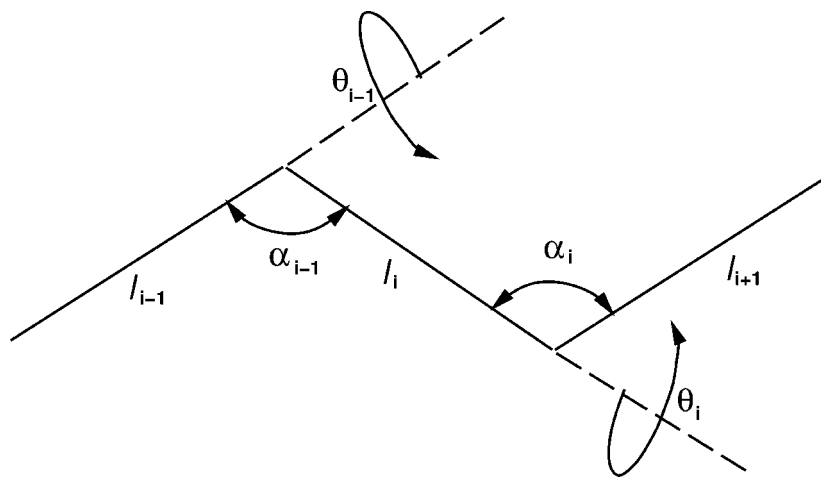

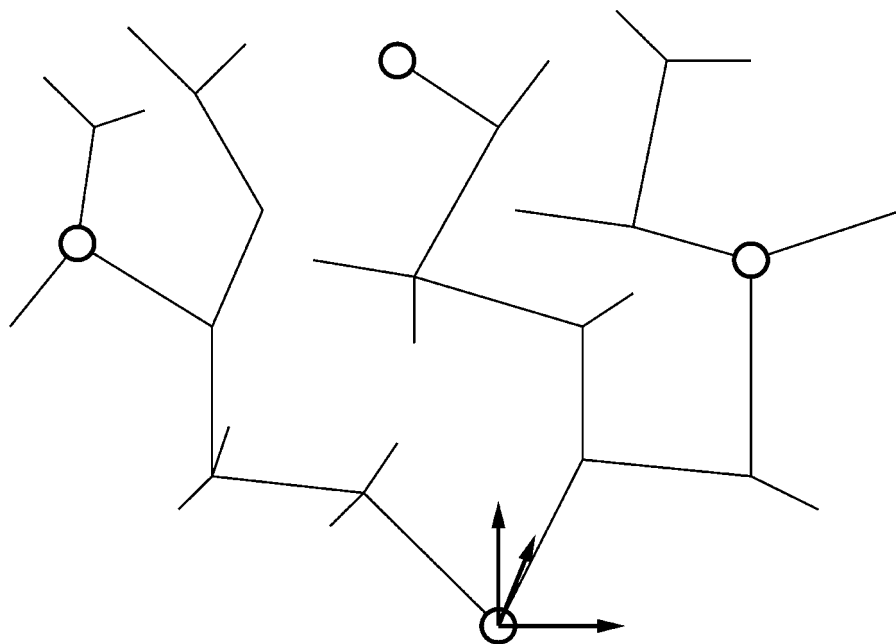**FIGURE 2.** The assignment of α, l, and θ parameters along the kinematic chain.

FIGURE 3. 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.

fore, necessary to allow the frame attached to the molecule at a_{anch} to achieve any orientation with respect to the global frame (see Fig. 3).

The space of the 3D rotations can be parameterized using Euler angles, γ, ϕ, ψ, for which $0 \leq \gamma < \pi$, $0 \leq \phi < 2\pi$, and $0 \leq \psi < 2\pi$.[1] The parameterized rotation matrix can be placed into homogeneous form by extending one row and column to obtain

$$
\begin{aligned}
&T_R(\gamma, \phi, \psi) \\
&= \begin{bmatrix}
c\phi s\gamma c\psi & -c\phi c\gamma s\psi - s\phi c\psi & c\phi s\gamma & 0 \\
s\phi c\gamma c\psi + c\phi s\psi & -s\phi c\gamma s\psi + c\phi c\psi & s\phi s\gamma & 0 \\
-s\gamma c\psi & s\gamma s\psi & c\gamma & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}.
\end{aligned}
\tag{3}
$$

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

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

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 nonrotatable 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 kth 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 kth 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

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

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[33] to

use an empirical energy function; we use the SYBYL system (Tripos Inc.[30]) energy function:

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

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.

Our algorithms require no assumptions on the particular form of the energy function. Thus, a variety of other energy models could be successfully substituted.

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 $\boldsymbol{\theta}, \gamma, \phi$, 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 pharmaceutical chemists. This is generally referred to as the conformational coverage problem. Techniques for analyzing the amount of coverage obtained from a small set of representative conformations appear in refs. 23–25. There might be several low-energy conformations that satisfy the pharmacophore, but each could place the other non-feature atoms in very different locations. In general, having alternative low-energy conformations is useful because in many cases it is not the lowest-energy conformation that results in docking, but another low-energy conformation.

These distinct conformations may also 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. One 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 conformations. In a later section, a clustering technique is described that ensures that the reported clusters sufficiently differ according to this metric.

Randomized Conformational Search with Constraints

THE APPROACH

In this section we focus our attention on Question 1 above. In the next section we show how to integrate the approach with a database screening system that can answer both Questions 1 and 2. Figure 4 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
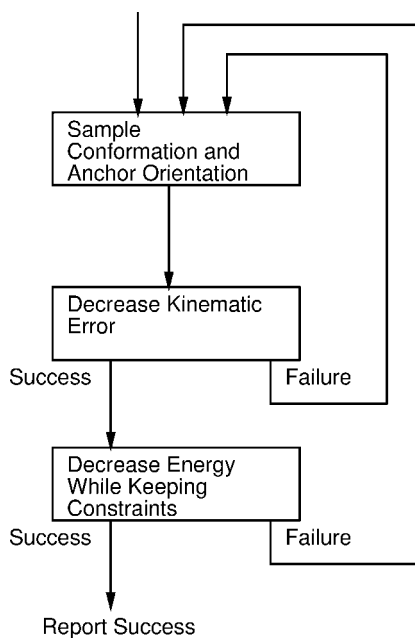
FIGURE 4. An overview of constrained conformational search.

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, e.g., refs. 21, 34–37).

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.[38] Efficient elimination techniques from computational algebraic geometry have been devel-

oped for numerating the solution set for problem in which there are a finite number of solutions.[31, 39] For the problem discussed in this article, the system of equations is generally underconstrained, which leads to a complicated, multidimensional 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.[38] 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 5 and 6, 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.[30]) 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 the last section 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: (1) the counter k reaches its maximum value, k_{max}, which means that too many iterations have been executed, or (2) 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.

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

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 5. This algorithm iteratively attempts to reduce the kinematic error.

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 7a 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,

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

FIGURE 7. (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.

then the new parameters are evaluated based on energy. The speed of the algorithm is directly related 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 7b, 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.

An Integrated Database Screening System

A prototype database screening system has been implemented and tested on a small molecular database. The principle module in our system is the constrained conformational search technique described in the last section. We have not performed extensive experimentation and evaluation on large databases because we are not evaluating pharmacophore models and representations, as is usually done in this area. Instead, our experiments are designed to illustrate the capabilities of our computational techniques on pharmacophore matching problems that are sufficiently complicated and diverse.

We assume that our techniques will be applied to a small set of molecules that were initially filtered from a database using 2D information. This could lead to small set of several hundred or fewer molecules that contains both successful candidates and many false positives. Standard indexing techniques are used to isolate these candidates from the full database. Our system is designed to further reduce the set of candidates by performing kinematics-based conformational search.

ITERATIVE SWEEPING ACROSS THE SET OF CANDIDATES

Suppose that one would like to use the approach described in Figure 4 to search a database for molecules that have a low-energy conformation that satisfies a given pharmacophore. Initially, 2D 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 (assuming that other constraints have already filtered out molecules that do not satisfy nongeometric constraints).

One sensible way to use this method is to sweep across the database, using only one iteration of the approach in Figure 4 for each molecule. Once the last molecule has been processed, a second sweep can be made across the database. The sweeps across the database can be repeated until some predetermined criteria are met (e.g., 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 it-

eration, 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. One tradeoff is that saving search information after each iteration could yield costly storage requirements, if a very large set of candidates is considered.

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 before.) In this case, clustering can be performed incrementally for each solution that is generated.[40] A metric $m(\theta_1, \theta_2)$ can be defined that quantifies the difference between two conformations. In the database screening system, $m(\theta_1, \theta_2)$ 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. An alternative metric, which measures distance in the conformation space (by comparing torsion angles), also yields reasonable performance.

The incremental clustering approach proceeds are follows. The "Report Success" step in Figure 4 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 six different inhibitors of thermolysin[41, 42] and six different inhibitors of ACE.[43, 44] Table I 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. Some of the inhibitors of thermolysin are shown in Figures 8 and 9, and some of ACE are shown in Figures 10 and 11. Each molecule appears in a random conformation 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 four features of the known pharmacophore, and for the ACE inhibitors we ran a query with three 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 4 could be performed in about 5–10 s. A sufficient clustering record for a single molecule required about 5–20 min. Our results are shown in Table II for the thermolysin inhibitors and in Table III 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

TABLE I.

The Thermolysin and ACE Inhibitors Used in Our Experiments.

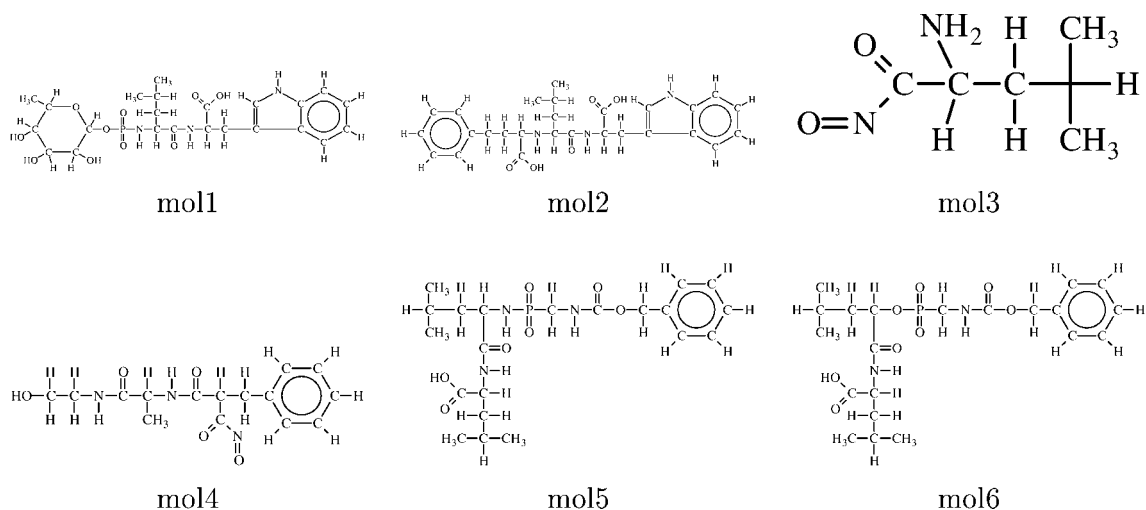| Molecule	Thermolysin Inhibitors		Ace Inhibitors	
	Atoms	Rot. Bonds	Atoms	Rot. Bonds
mol1	69	10	48	8
mol2	66	11	50	8
mol3	22	3	31	7
mol4	42	8	47	8
mol5	64	13	45	6
mol6	63	12	30	3

FIGURE 8. The inhibitors of theormolysin used in our experiments.

we find are indeed of low energy. In general, the energies of the conformations found are within 2–7 kcal/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 12 we show the representatives of the clusters that are generated for mol5 of
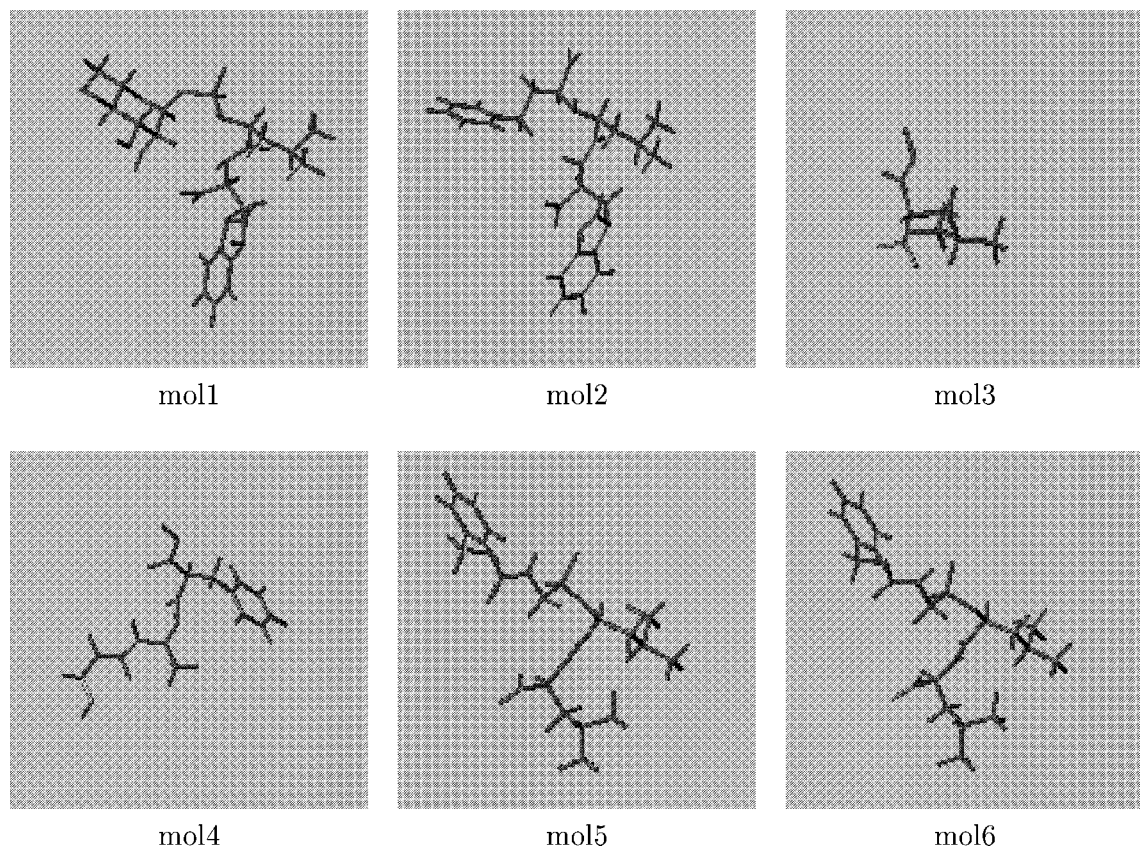

FIGURE 9. The 3D models of the inhibitors of theormolysin. Each molecule is at a random conformation.
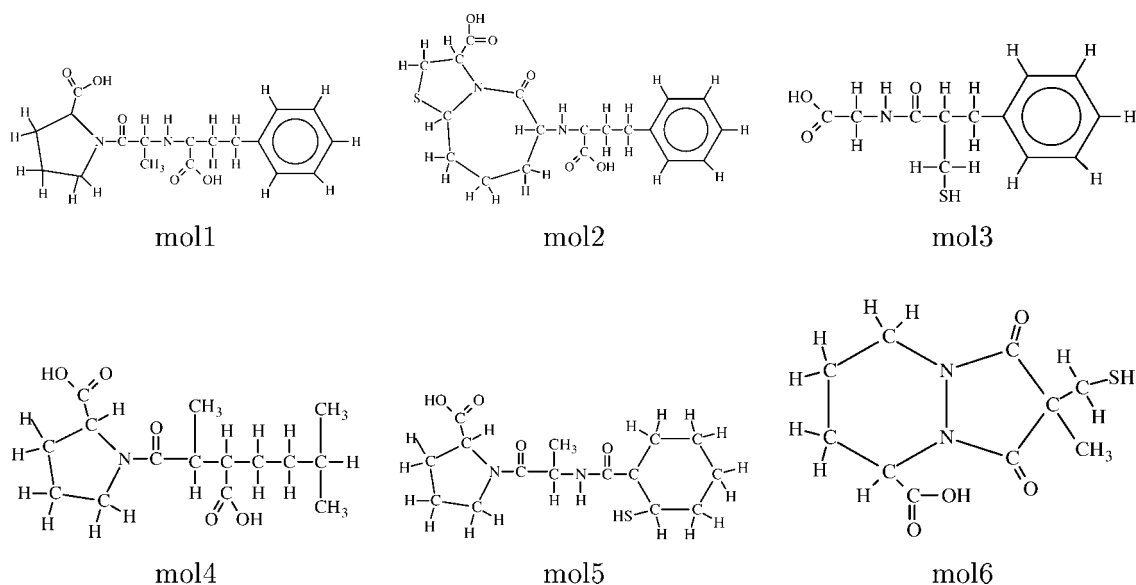
FIGURE 10. The inhibitors of ACE used in our experiments.

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 13. A similar observation holds here.
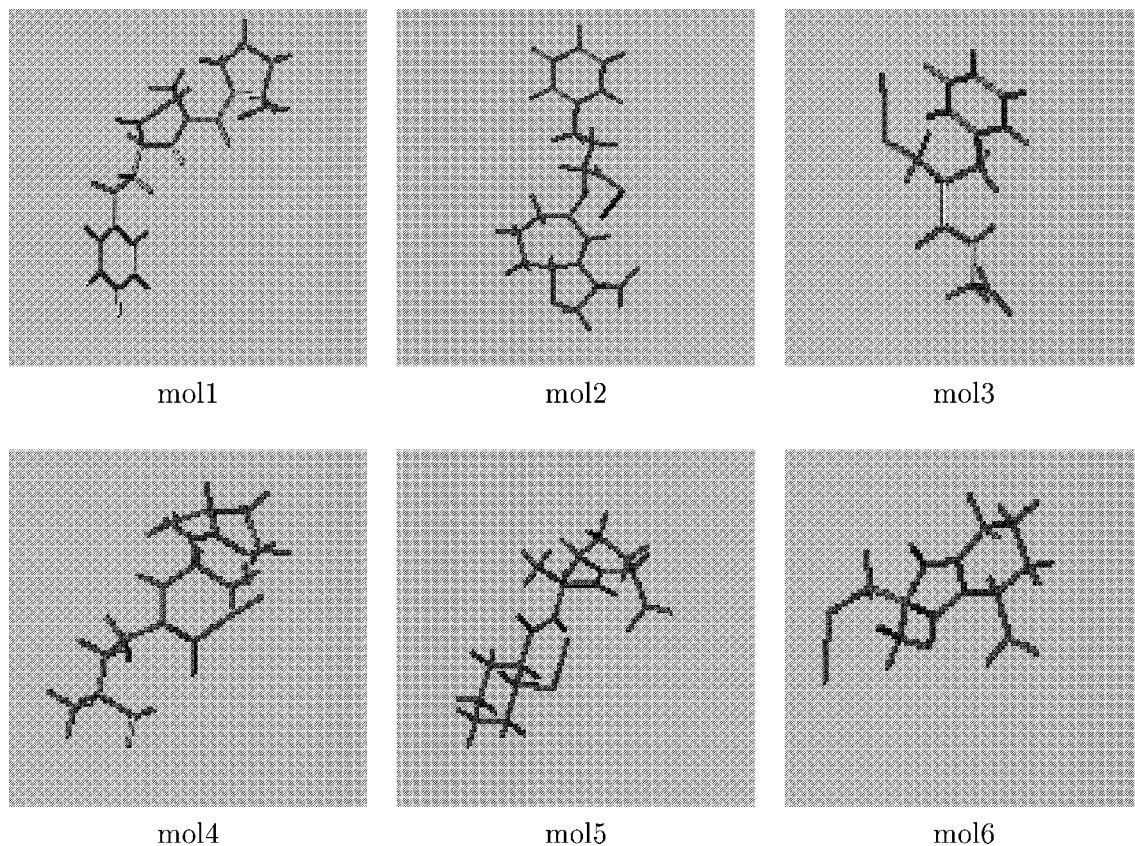

FIGURE 11. The 3D models of the inhibitors of ACE. Each molecule is at a random conformation.

TABLE II.
The Results of a Four-Feature Thermolysin Search.

Thermolysin Inhibitors	Number of Clusters	Minimum 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

TABLE III.
The Results of a Three-Feature ACE Search.

Ace Inhibitors	Number of Clusters	Minimum 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

Our implementation allows us to answer the two questions posed earlier. Question 1 is answered 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 the second section, the position of nonfeature atoms may help the pharmaceutical chemist select among
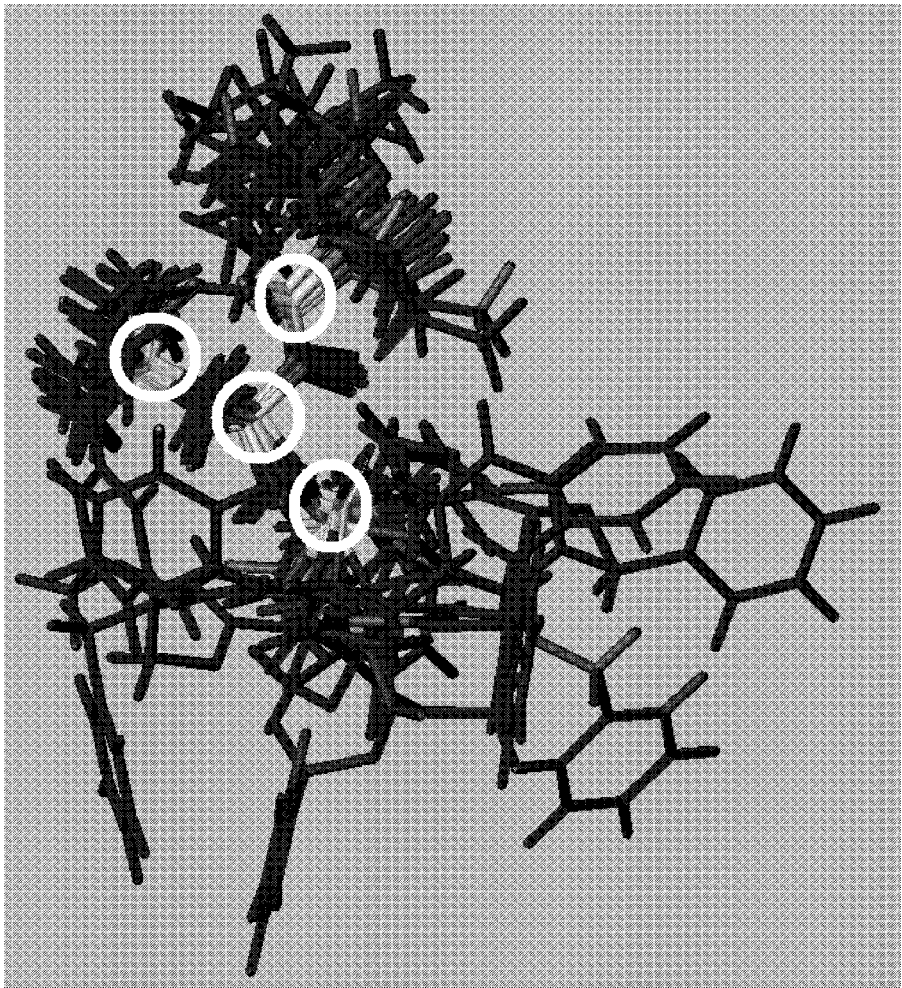

FIGURE 12. Twelve clusters were obtained for mol5 of the thermolysin inhibitors.
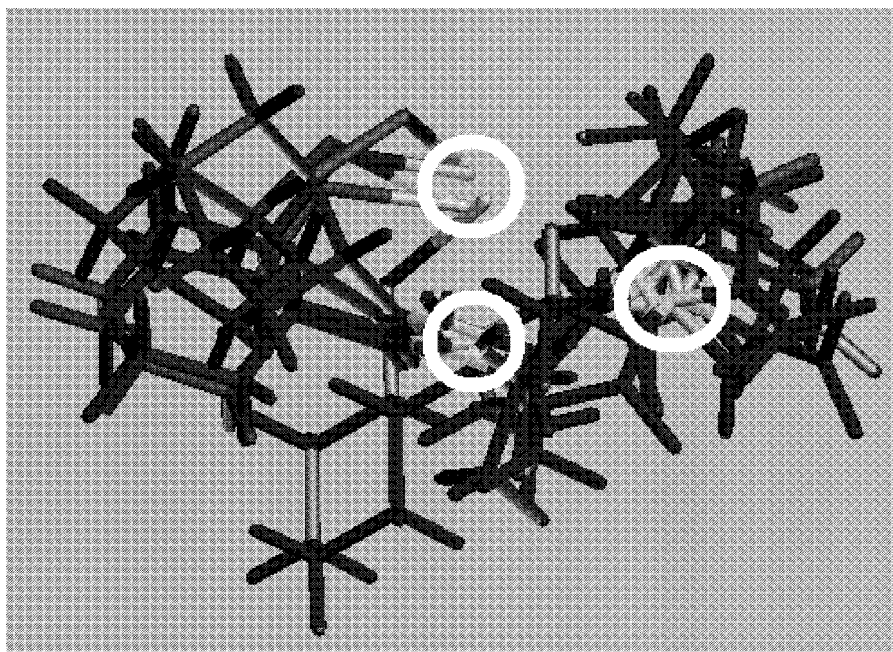
FIGURE 13. Five clusters were obtained for mol5 of the ACE inhibitors.

possible ligands or modify existing ligands in search for a better drug.

Our previous work with randomized techniques has shown[4, 8] 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 five thermolysin inhibitors, we obtained conformations whose RMS distances from the corresponding docked conformation were 0.50, 2.96, 0.59, 0.81, 2.40, and 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, and 1.98 Å. Notice that the more features we use, the closer we can get to the docked conformation (four features for the thermolysin inhibitors vs. three features for the ACE inhibitors). The ACE inhibitors have less flexibility, which helps to reduce the fitted RMS distance. In certain cases, we get very close to the docked conformation (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.

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 of the approach should enable straightforward extensions to other classes of molecules, such as those with flexible rings.

We point out, however, that a variety of candidate evaluation models are used by other researchers. Our approach focuses on one particular model (a rigid pharmacophore), and evaluates the correctness of our computational approach. More extensive experimentation would be needed to evaluate our approach in comparison to other evaluation models proposed in pharmaceutical drug design and computational biology. One possible constraint that could be added to our models is exclusion spheres. These would serve as additional kinematic constraints that could be incorporated directly into the search.

A difficulty with the current approach is the selection of the step size for the random neighbor-

hoods 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.

Another limitation of our approach is the local minimum problem, which could cause false negatives. By generating multiple conformations for each molecule, and by appropiately focusing the database screening in an "any time" manner, we have avoided focusing matching efforts on a single, local optimization. This avoids many local minimum problems. Over time, we expect the search to gradually find a global minimum and sufficiently cover the conformation space of each molecule, but it is difficult to guarantee that this will be done efficiently.

Acknowledgments

References

1. Craig, J. J. Introduction to Robotics; Addison–Wesley, Reading, MA, 1989.

2. Hartenburg, R. S.; Denavit, J. J Appl Mech 1955, 77, 215.

3. Canny, J. F. The Complexity of Robot Motion Planning; MIT Press, Cambridge, MA, 1988.

4. Kavraki, L. E. PhD thesis, Stanford University, 1994.

5. Latombe, J.-C. Robot Motion Planning; Kluwer Academic Publishers, Boston, MA, 1991.

6. Sepetov, N. F.; Krchnak, V.; Stankova, M.; Wade, S.; Lam, K. S.; Lebl, M. Proc Natl Acad Sci USA 1995, 92, 5426.

7. Oprea, T. I.; Waller, C. L. In Reviews in Computational Chemistry; Lipkowitz, K. B.; Boyd, D. B., Eds.; John Wiley and Sons, New York, 1997, p. 127.

8. Finn, P. W.; Halperin, D.; Kavraki, L. E.; Latombe, J.-C.; Motwani, R.; Shelton, C.; Venkatasubramanian, S. In Applied Computational Geometry (Lecture Notes in Computer Science, 1148); Springer-Verlag, Berlin, 1996, p. 67.

9. Ghose, A. K.; Logan, M. E.; Treasurywala, A. M.; Wang, H.; Wahl, R. C.; Tomczuk, B. E.; Gowravaram, M. R.; Jaeger, E. P.; Wendoloski, J. J J Am Chem Soc 1995, 117, 4671.

10. Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J. In Trans in QSAR and Molecular Modeling; Wermuth, C.-G., Ed.; ESCOM, Leiden, 1993, p. 20.

11. Henry, D. R.; Özkabak, A. G. In Encyclopedia of Computational Chemistry; Schleyer, P. v. R., Ed.; Wiley, New York, 1998.

12. Gund, P.; Wipke, W. T.; Langridge, R. Comput Chem Res Ed Technol 1974, 3.

13. Murrall, N. W.; Davies, E. K. J Chem Inf Comput Sci 1990, 30, 312.

14. Lemmen, C.; Lengauer, T.; Klebe, G. J Med Chem 1998, 41, 4502.

15. Willett, P. J Mol Recogn 1995, 8, 290.

16. Güner, O. F.; Henry, D. R.; Pearlman, R. S. J Chem Inf Comput Sci 1992, 32, 101.

17. Clark, D.; Jones, G.; Willett, P.; Kenny, P.; Glen, R. J Chem Informat Comput Sci 1994, 34, 197.

18. Lipton, M.; Still, W. J Comput Chem 1988, 9, 343.

19. Blaney, J.; Crippen, G.; Dearing, A.; Dixon, J. Dgeom: Distance geometry, Quantum Chemistry Program Exchange, 590, Dept. of Chemistry, Indiana Univ., IN.

20. Crippen, G. M.; Havel, T. F. In Chemometrics Research Studies Series; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH Piblisher, New York, 1988.

21. Chang, G.; Guida, W.; Still, W. J Am Chem Soc 1989, 111, 4379.

22. Saunders, M.; Houk, K. N.; Wu, Y.-D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. J Am Chem Soc 1990, 112, 1419.

23. Smellie, A.; Kahn, S. D.; Teig, S. L. J Chem Inf Comput Sci 1995, 35, 285.

24. Smellie, A.; Kahn, S. D.; Teig, S. L. J Chem Inf Comput Sci 1995, 35, 295.

25. Smellie, A.; Teig, S. L.; Towbin, P. J Comp Chem 1995, 16, 171.

26. Fontain, E. J Chem Informat Comput Sci 1992, 32, 748.

27. Jones, G.; Willett, P.; Glen, R. C. J Comput Aided Mol Design 1995, 9, 532.

28. Hurst, T. J Chem Informat Comput Sci 1994, 34, 190.

29. Moock, T. E.; Henry, D. R.; Ozkabak, A. G.; Alamgir, M. J Chem Inf Comput Sci 1994, 34, 184.

30. UNITY; Tripos, St. Louis, MO.

31. Emiris, I. Z.; Mourrain, B. Technical report, INRIA, Sophia-Antipolis, France, 1997.

32. Wang, C. J Comput Chem 1992, 13, 730.

33. Boyd, D. B. In Reviews in Computational Chemistry; Boyd, D. B.; Lipkowitz, K., Eds.; VCH Piblisher, Weinheim, 1990, Vol. 1, p. 321.

34. Byrne, D.; Li, J.; Platt, E.; Robson, B.; Weiner, P. J Comput Aided Mol Design 1994, 8, 67.

35. Fadrna, E.; Koca, J. J Phys Chem B 1997, 101, 7863.

36. Ghose, A.; Kowalczyk, J.; Peterson, M.; Treasurywala, A. J Comput Chem 1993, 14, 1050.

37. Perkins, T. D.; Dean, P. M. J Comput Aided Mol Design 1993, 7, 155.

38. Cox, D.; Little, J.; O'Shea, D. Ideals, Varieties, and Algorithms; Springer-Verlag, Berlin, 1997.

39. Manocha, D.; Canny, J. In IEEE Int Conf Robot Autom, Nice, May 1992, p. 383.

40. Leach, A. R.; Kuntz, I. D. J Comput Chem 1992, 13, 730.

41. Bohacek, R.; McMartin, C. J Med Chem 1992, 35, 1671.

42. Klebe, G.; Abraham, U. J Med Chem 1993, 36, 70.

43. Bohacek, R.; De Lombaert, S.; McMartin, C. J Am Chem Soc 1996, 118, 8331.

44. Fournie-Zaluski, M.-C. J Med Chem 1996, 39, 2594.