Search Techniques for Rational Drug Design

P.W. Finn* L.E. Kavraki+ J.C. Latombe† R. Motwani† S. Venkatasubramanian†

* Pfizer Central Research, Sandwich, Kent CT13 9NJ, UK.
+ Department of Computer Science, Rice University, Houston, TX 77005, USA.
† Department of Computer Science, Stanford University, Stanford, CA 94305, USA.

Abstract

Pharmaceutical drug design is a long and expensive process. Early selection of promising molecules can dramatically improve this process, but this selection is often similar to searching for a needle in a haystack. In most cases, all what a chemist initially has is a small collection of molecules that exhibit enough desired activity to hypothesize that they share a 3-D atomic substructure binding to the same receptor site. A key problem is to identify this substructure, which can then be used as a pattern to screen databases of molecules. This problem is complicated by the fact that a drug molecule is "flexible" and can achieve many low-energy (stable) states. In this paper we present search techniques that we have developed to find these low-energy states and to identify common 3-D substructures that appear in at least one low-energy state of most molecules in a given collection. We also show experimental results obtained with a software system implementing these techniques.

1 Introduction

A basic assumption in rational drug design, which is supported by a number of experimental results [7], is that the activity of a drug molecule – also called a **ligand** – results from the binding of this molecule to a larger molecule – the **receptor** (usually, a protein). In its binding state, a subset of the ligand exhibits geometric and chemical complementarity with a cavity of the receptor called the **binding site** [4]. Many drug molecules are inhibitors, i.e., if a receptor cavity is a favorable environment for a reaction, a ligand that stably fills that cavity can prevent this reaction from happening. Fig. 1 displays a receptor, the protease thermolysin, and (in darker grey) one of its inhibitors, the 1TMN, in its binding state.

In general, a receptor is a fairly rigid molecule. In contrast, a ligand has multiple degrees of freedom, mainly torsional ones around bonds connecting atoms. Any kinematically valid relative placement of the ligand's atoms in 3-D space defines a **conformation**. The ligand's internal energy is a function of its conformation. Only low-energy conformations are stable.

The problem is thus to identify a ligand that admits a low-energy conformation achieving good geometrical and chemical complementarity with the targeted binding site. When a 3-D model of this site is known, docking methods [6] exploit both the geometric and the chemical information available. But the geometric

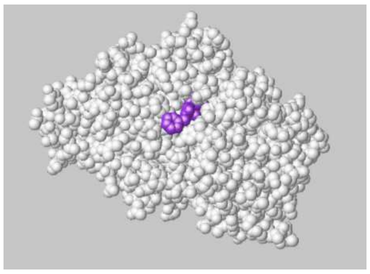

Figure 1: The protease thermolysin with one of its known inhibitors (1TMN)

structures of relatively few receptors have been obtained via X-ray crystallography or NMR techniques. Thus, chemists must often identify drug molecules for receptors whose structure is unknown [5, 20]. The starting point is then a collection of ligands that have been experimentally discovered to exhibit some level of desired activity. While these ligands are not potent enough to be good drug candidates, the hypothesis then made is that they can bind at the (unknown) targeted receptor site. By examining the possible shapes of these ligands, chemists seek to identify a 3-D substructure of features (e.g., atoms or groups of atoms of certain types) that is contained in some low-energy conformation of each (or most) of the ligands. This substructure, called a **pharmacophore**, is then regarded as a partial complement of the receptor's binding site, and is used as a template to find more potent ligands [14]. For example, one may screen large databases of ligands to extract those molecules which have a low-energy conformation that contains the pharmacophore [21]. Measuring the activity of these molecules allows the selection of more potent ligands, which in turn may yield more pertinent pharmacophores. Chemists eventually use the most potent ligands as leads, and modify them in a trial-and-error fashion, by adding/removing groups of atoms and testing the activity level of the modified ligands.

Conformational search, pharmacophore identification, and database screening are three important challenging search problems in rational drug design. The solution to each problem (in the given order) is an input for the next problem. Our research seeks efficient algorithms for solving these problems. In this paper, for limitation of space, we only address the first two

– conformational search and pharmacophore identification. For each, we give a brief state-of-the-art and we present techniques that we have developed. All reported timings are on an SGI Indigo2 with 175MHz MIPS R8000 processor and 384 MB RAM. Code was written in C/C++, and compiled using SGI CC.

2 Conformational Search

2.1 Problem and Background

The **conformational search** problem is:

Given a collection of ligands $M = \{M_1, M_2, \ldots, M_N\}$, the degrees of freedom of each of them, and an energy function E, find, for each M_i, a set of conformations $\{C_{i1}, C_{i2}, \ldots, C_{ik_i}\}$, such that $E(C_{ij}) \leq$ THRESHOLD and $d(C_{ij}, C_{il}) >$ TOLERANCE for $l \neq j$ and $j, l = 1, \ldots, k_i$. THRESHOLD and TOLERANCE are pre-specified values, while $d(\cdot,\cdot)$ is a given distance function.

In theory, the degrees of freedom (DOFs) of a ligand include bond lengths, bond angles (angles between two consecutive bonds), and dihedral or torsional angles (angles formed by the first and third of three consecutive bonds, viewed along the axis of the second bond). In practice, only the torsional DOFs are considered because these are the only ones that achieve large variations [7]. In many drug molecules, the number of torsional DOFs ranges between 3 and 15; e.g., the 1TMN shown in Fig. 1 has 11 DOFs. The energy of a ligand is a function of its conformation. While quantum mechanics allows for a precise modeling of molecular behavior, it yields prohibitively expensive computations. Therefore, the energy function considered is usually an approximation based on van der Waals and Coulomb potentials [19]. Such approximation has produced meaningful results (e.g., see [6, 7]).

Conformational search is a classical problem in computational chemistry. It includes protein folding [10], where one must find with great precision the conformation of a large molecule that achieves the global minimum of the energy function. Instead, smaller, flexible ligands have multiple stable states, each corresponding to a local energy minimum. So, one must find a set of geometrically distinct conformations whose energy is below a threshold [17]. It is not critical to find each such conformation with high precision since, when a ligand binds against a receptor, the energy function is modified by a binding-energy term that one cannot compute in the absence of a model of the binding site; this additional term is usually small compared to the ligand's internal energy, but it is not negligible.

Both systematic and randomized techniques for conformational search have been proposed [17]. Systematic search methods sample each DOF of the ligand at regularly spaced intervals and were among the first to be developed [18]. Using each sampled conformation as a starting point, these methods run some sort of gradient-descent routine to minimize the energy of all generated conformations [18]. To avoid considering a prohibitive number of conformations these methods rely on a very coarse discretization of the torsional values. They also use heuristics to terminate each gradient descent whenever the conformation gets close to a previous considered conformation.

Randomized methods seem faster and more promising than systematic ones [9, 13]. Random sampling techniques have already been successfully applied to other high-dimensional search problems, like motion planning [15], where the shape of the underlying space is irregular. The same observation holds for conformational search: a systematic procedure has a higher chance of missing the irregularly shaped basins of attraction of the energy landscape of the ligand [13]. We describe below one method in more detail, which we have developed and tested [11].

2.2 Randomized Search

Initially, a large number of conformations are picked at random. Unlike previous randomized search methods, which apply random increments to the torsional DOFs starting with a user-specified initial conformation [13] or with a previously discovered low-energy conformation [9], our method picks a random conformation by selecting each DOF from its allowed range according to a user-specified distribution reflecting any a priori information about the preferred values of the DOF [16]. Each generated conformation is retained only if it avoids intersections of non-bonded atoms (classically modelled by spheres of van der Waals radii [7]). An efficient minimizer [8] is then used to continuously deform the conformation until its energy is close to a local minimum.

To eventually obtain a reduced, but yet representative set of low-energy conformations, we partition the optimized conformations into clusters that reflect geometric similarity as captured by a distance measure d. For any two given conformation C_i and C_j, we define $d(C_i, C_j)$ as the square root of the mean of the squared distances of the corresponding atoms of C_i and C_j, after C_i is transformed to C_j. This transformation is computed using a basis of three predefined atoms a_1, a_2, and a_3, by successively aligning the two occurences of a_1, the two occurrences of the vector a_1a_2, and the two occurences of the plane $a_1a_2a_3$. We perform a greedy clustering of the conformations by placing a given conformation in an existing cluster if its distance from the "center" of that cluster is less than the predefined value TOLERANCE. If no such cluster is found, a new cluster is created. The center of a cluster is the conformation with the lowest energy in the cluster. The implementation employs heuristics for ensuring that the cluster center is close to the average cluster conformation and that all cluster centers are sufficiently far apart. We stop producing new conformations when a series of conformations of predefined length yield minimized conformations that all fall into existing clusters.

2.3 Experimental Results

Four inhibitors of protease thermolysin are depicted in Fig. 2 using the so-called "stick" model (only bonds between atoms are shown): 1TLP has 69 atoms and 10 torsional DOFs, 4TMN has 68 atoms and 15 DOFs, 5TMN has 64 atoms and 13 DOFs, and 6TMN has 63 atoms and 12 DOFs. Each of these molecules

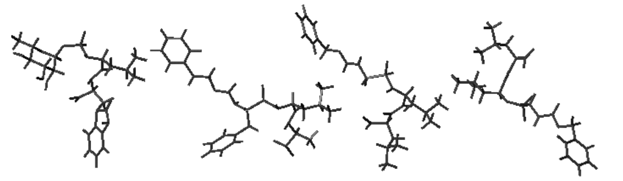

Figure 2: 1TLP, 4TMN, 5TMN, and 6TMN are inhibitors of thermolysin.

was run through our conformational search software. THRESHOLD was set to 20 Kcals/mol and TOLERANCE to 1.23Å. It took 4.3 hours to produce 4,000 valid optimized conformations for each molecule and reduce these to a set of representative clusters. These timings are very reasonable given that 10, 15, 13, and 12 dimensional spaces had to be searched. The numbers of representatives produced were 850, 1192, 1024, and 955 for 1TLP, 4TMN, 5TMN, and 6TMN, respectively. Since the binding conformation of each of these molecules is known, we confirmed that it was close to one of the produced representatives.

3 Pharmacophore Identification

3.1 Problem and Background

In the following we model a ligand as a set of labeled points. Each such point designates an atomic feature, e.g., the center of an individual atom or the center of a benzene ring. While its label identifies the type of the feature, its position (in any particular conformation) defines the feature's position. If needed, a multi-atom feature can be modelled by two or more points to encode orientation.

The **pharmacophore identification** problem is:

Given a collection of ligands $M = \{M_1, M_2, \ldots, M_N\}$*, where each* M_i *has a set of conformations* $\mathcal{C}(M_i) = \{C_{i1}, C_{i2}, \ldots, C_{ik_i}\}$*, determine a set of labeled points* S *in 3-D such that for all* $i \in \{1, \ldots, n\}$*, there exists some* $C_{ij} \in \mathcal{C}(M_i)$ *such that* S *is congruent (w.r.t. rotations and translations) to some subset of* C_{ij}*.*

In our software, the collection of ligands M is input by the user. In practice, the size of M ranges from 5 to 10 ligands. For each molecule M_i, the conformations $\{C_{i1}, C_{i2}, \ldots, C_{ik_i}\}$ are the cluster centers found by the conformational search method presented in Section 2. The size of $\mathcal{C}(M_i)$ ranges from a dozen to several hundred conformations. Each solution S of the above problem is called an **invariant**; it is a candidate pharmacophore.

The above problem formulation assumes that all point positions are known exactly. In fact, low-energy conformations are cluster representatives and are not exactly defined. Thus, we use the convention that two points p_1 and p_2 **match** when $| p_1 - p_2 | \leq \epsilon$, where ϵ is the tolerance on point location. Increasing ϵ would allow conformational search to generate larger clusters; hence, less conformations per ligand would have to be considered for pharmacophore identification. But this could also cause a greater number of spurious invariants to be generated by the pharmacophore identification software.

Searching for an unknown invariant in several ligands each having many possible conformations is obviously a very complex problem. DISCO [20] uses clique searching techniques. But this approach leads to a combinatorial explosion of the number of basic operations performed, and is limited to small sets of conformations per ligand [5]. For that reason, different approaches are being investigated. One of them, which we have proposed in [12], is based on probing conformations at random. We present it below and we give experimental results. We first describe a probing procedure to find invariants in two given point sets (pairwise matching). Next we show how this procedure is used to solve the pharmacophore identification problem (multiple matching).

3.2 Pairwise matching

Let P_1 and P_2 be two point sets in 3-D representing two conformations of different ligands. Assume without loss of generality that both P_1 and P_2 have the same cardinality n. We wish to determine a point set S of size $|S| \geq \alpha n$ that is congruent to some subset of each P_i, $i = 1, 2$, where $\alpha \leq 1$ is a given constant used to characterize the minimal size of the seeked invariants. For the case when $\alpha = 1$, i.e., determining the congruence of two point sets, an algorithm that runs in time $O(n \log n)$ in 2-D and 3-D was presented in [3]. However, for general α, the best known algorithms were obtained by [1]. These algorithms have a worst-case running time of $O(n^{4.6})$ for unknown α, and $O(n^{2.6}/\alpha^2)$ (randomized) when α is known, for 3-D. However, these bounds apply only for the noise-free model. The noisy version was considered in [2] yielding an $O(n^8)$ algorithm in 2-D.

We have developed and experimented with two random-probing procedures for solving pairwise matching with noisy data [12]. One is technically simpler and has lower asymptotic complexity than the other; however, in our experiments the more technical procedure runs consistently faster. Both give solutions of comparable quality. Here, we describe only the simpler procedure. It works as follows:

`pair-probe`(P_1, P_2):
For some constant c, perform $(c \log n)/\alpha^3$ iterations of the following:

1. Pick a triplet of points $< p_1, p_2, p_3 >$ at random from P_1.
2. Determine three points in P_2 congruent to this set and compute the resulting induced transformation.
3. Determine the number of points in P_1 matching points in P_2; if the number exceeds αn, retain this set of points as a solution of the pairwise matching problem.

When the input data is noise-free, one can verify that, if P_1 and P_2 have a common subset S such that $|S| \geq \alpha n$, the probability that the above procedure fails to find it is at most c'/n, for some constant c'.

Moreover, we have also shown [12] that the procedure runs in time $O((n^{2.8}\log n)/\alpha^3)$ using space $O(n^2)$.

To make the above procedure more practical, we have to address two additional issues:

Elimination of redundant solutions: The above procedure could return many solutions that satisfy a containment relationship with respect to each other. To prevent such redundant solutions from propagating through the complete search procedure of Subsection 3.3, we prune them. Given two invariants S_1 and S_2, we check if $S_1 \subseteq S_2$ by invoking the same procedure as above on these two sets with α set to 1.

Inaccuracies in point positions: The core computation in the above procedure is the following: Given points $p_1, p_2, p_3 \in P_1$, and $p'_1, p'_2, p'_3 \in P_2$, compute a transformation T from P_1 to P_2 that maps p_i to p'_i, for $i = 1, 2, 3$. The simplest way to compute T is to successively align p_1 and q_1, the directions p_1p_2 and q_1q_2, and the planes $p_1p_2p_3$ and $q_1q_2q_3$. But this computation over-constrains the transform, since what we really need is a transformation that maps each p_i $(i = 1, 2, 3)$ to a point within distance ϵ of q_i, while maximizing the number of matching points of P_1 and P_2 (again, within ϵ). Unfortunately, there may exist different transformations that yield maximal but incomparable common subsets. To be fast, our implementation proceeds as follows: Determine a seed transformation T using the simple technique indicated above; next, a small number of times, sample a triplet at random from the subset of P_1 that T matches against a subset of P_2, and use this triplet to construct a new transformation.

3.3 Multiple Matching

There are various strategies one could use to process multiple molecules. A simple one is the following:

`multi-probe`$(\{M_1, \ldots, M_N\})$:

1. Extract invariants from two molecules with multiple conformations by calling `pair-probe` on all pairs of conformations.
2. Test each solution S obtained against every remaining molecule M_i, by invoking `pair-probe` on S and each of the point set representing a conformation of M_i.

In Step 2, each invocation of `pair-probe` may return 0, 1, or several solutions (all included in S), since multiple subsets of S of sufficient size may separately match a subset of points in a conformation of M_i.

Another strategy, which is less order-sensitive, consists of running the two-molecule algorithm on distinct pairs of molecules and recursively combining the results using a binary tree.

In some cases, we may wish to find an invariant that does not necessarily exist in all the molecules, but in some fixed number of them. To do this we use a modified merging strategy here to keep track of the number of times an invariant fails to match against a molecule, and only reject those which exceed the maximum allowed number of failures.

Both `pair-probe` and `multi-probe` can be improved using heuristics. For example, if in a given conformation an atom (modeled by a sphere) does not contribute a significant portion of its surface to the total molecular surface, this atom is unlikely to play a binding role, and thus could be discarded to reduce the size of the point sets [11].

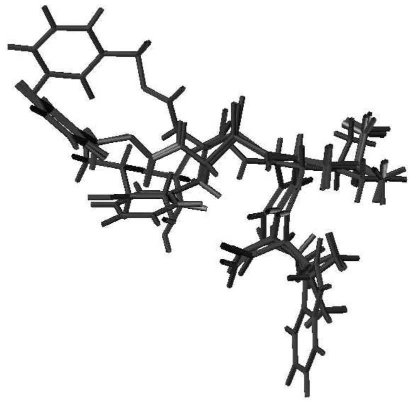

Figure 3: 1TLP, 4TMN, 5TMN, and 6TMN overlapped in their binding conformations.

3.4 Experimental results

Consider again the four inhibitors of thermolysin shown in Fig. 2, for which the pharmacophore was well-known prior to our experiments. Subsets of conformations produced by the software of Section 2 were input to the `multi-probe` algorithm described above. For this test, we considered each non-hydrogen atom to be a separate feature; thus, each molecule was modelled by approximately 30 points and there were 6 distinct point labels (i.e., atom types). The parameters ϵ and α were set to 1.3Å and 0.3, respectively. We set to 5 the number of iterations performed by `pair-probe` at each invocation.

We have run our software with several numbers of conformations per molecule. In all cases, we included the (known) binding conformation in our sets. In every case, the largest invariant was the "correct" solution, i.e., it was contained in the a priori known pharmacophore. This invariant consists of 14 atoms: 7 form the actual pharmacophore, while the other 7 atoms are connecting atoms (or "scaffolding" atoms) with no actual role in the binding. Fig. 3 shows the four molecules in the conformations containing the identified invariant, with the four occurrences of this invariant aligned. For 21 conformations per molecule, the total running time of the software was approximately 6.6 hours.

4 Summary and Conclusions

Rational drug design raises challenging search problems. We discussed two of them: conformational search and pharmacophore identification. Another third important problem, which we currently investigate, is database screening [21], i.e., searching large databases for molecules that admit a low-energy conformation containing a given pharmacophore.

Our current goal is to improve our pharmacophore identification techniques and perform experiments involving 5-20 ligands and a large number of conformations per ligand. Toward that goal we have implemented of a rule-based software module that extracts multi-atom features from ligands. Each such feature (e.g., a benzene ring, an hydrophobic group of atoms) is routinely used by chemists to capture the behavior of a subset of a molecule. We expect that features will reduce the number of labeled points we must consider per conformation, while increasing the number of different labels, hence reducing the total combinatorics of the problem. We also believe that more rigorous analysis of the combinatorial properties of point sets representing conformations should help us reduce the running time of our algorithms.

Concurrently, we are experimenting with our software on new collections of molecules, both at Stanford and at Pfizer. For some collections (e.g., ligands having anesthetic/analgesic action), neither the binding conformations, nor the pharmacophore are known a priori.

Acknowledgments

This research is supported by a grant from Pfizer Central Research, Sandwich, UK. Rajeev Motwani's research is also partially supported by an Alfred P. Sloan Research Fellowship, an IBM Faculty Partnership Award, an ARO MURI Grant DAAH04-96-1-0007, and NSF Young Investigator Award CCR-9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation. Lydia Kavraki is partially supported by startup funds from Rice University. The authors thank Dan Halperin, Piotr Indyk, Steve LaValle, Christian Shelton, and Andy Yao for their multiple contributions to this work.

References

[1] T. Akutsu, H. Tamaki, and T. Tokuyama. Distribution of distances and triangles in a point set and algorithms for computing the largest common point set. *Proc. 13th ACM Symp. on Comp. Geometry*, 1997.

[2] H. Alt, K. Mehlhorn, H. Wagener, E. Welzl. Congruence, similarity and symmetries of geometric objects. *Discrete Comput. Geom.*, 3:237–256, 1988

[3] M. D. Atkinson. An optimal algorithm for geometric congruence. *J. Algorithms*, 8:134–183, 1987.

[4] L. Balbes, S. Mascarella, and D. Boyd. A perspective of modern methods in computer-aided drug design. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Comp. Chem.*, 5:337–370, VCH Pub., 1994.

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

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

[7] D. B. Boyd. Aspects of molecular modeling. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Comp. Chem.*, 1:321–351, VCH Pub., 1990.

[8] R. Brent. *Algorithms for finding zeros and extrema of functions without calculating derivatives.* PhD Thesis, Stanford U, 1971.

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

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

[11] P. Finn, D. Halperin, L.E. Kavraki, J.-C. Latombe, R. Motwani, C. Shelton, and S. Venkatasubramanian. Geometric manipulation of flexible ligands. In M. Lin and D. Manocha, editors, *LNCS Series – 1996 ACM Workshop on Applied Comp. Geometry.* 1148:67–78, Springer-Verlag, 1996.

[12] P. Finn, L.E. Kavraki, J.C. Latombe, R. Motwani, C. Shelton, S. Venkatasubramanian, and A. Yao. RAPID: Randomized Pharmacophore Identification for Drug Design. *Proc. 13th ACM Symp. on Comp. Geometry*, 1997.

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

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

[15] L. Kavraki. *Random Networks in Configuration Space for Fast Path Planning.* PhD Thesis, Stanford U, 1995.

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

[17] A. Leach. A survey of methods for searching the conformational space of small and medium sized molecules. In K. Lipkowitz and D. Boyd, editors, *Reviews in Comp. Chem.*, 2:1–47, VCH Pub., 1991.

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

[19] T. Lybrand. Computer simulation of biomolecular systems using molecular dynamics and free energy perturbation methods. In K. Lipkowitz and D. B. Boyd, editors, *Reviews in Comp. Chem.*, 1:295–320, VCH Pub., 1990.

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

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