RAPID: Randomized Pharmacophore Identification for Drug Design*

P. W. FINN[†] L. E. KAVRAKI[‡] J.-C. LATOMBE[§] R. MOTWANI[¶] C. SHELTON[§]
S. VENKATASUBRAMANIAN[§] A. YAO[||]

Abstract

This paper describes a randomized approach for finding invariants in a set of flexible ligands (drug molecules) that underlies an integrated software system called RAPID currently under development. An invariant is a collection of features embedded in $\Re^3$ which is present in one or more of the possible low-energy conformations of each ligand. Such invariants of chemically distinct molecules are useful for computational chemists since they may represent candidate pharmacophores. A pharmacophore contains the parts of the ligand that are primarily responsible for its interaction and binding with a specific receptor. It is regarded as an inverse image of a receptor and is used as a template for building more effective pharmaceutical drugs. The identification of pharmacophores is crucial in drug design since the structure of the targeted receptor is frequently unknown, but a number of molecules that interact with the receptor have been discovered by experiments. It is expected that our techniques and the results produced by our system will prove useful in other applications such as molecular database screening and comparative molecular field analysis.

To appear in the 13th Symposium on Computational Geometry, 1997.

1 Introduction

Computational chemists working in the area of structure-based drug design consider both chemical and geometric properties of the interacting molecules when developing new pharmaceutical drugs [8, 28, 33]. The underlying assumption is that drug activity, or pharmacophoric activity, is obtained through the molecular recognition and binding of one molecule (ligand) to a pocket of another, usually larger, molecule (receptor). This assumption is supported by a number of experimental results showing molecules with geometric and chemical complementarity in their active, or binding, conformations [11].

When the three-dimensional structure of the receptor is known, docking methods [10] exploit both the geometric and the chemical information available. However, the geometric structures of relatively few molecules have been obtained via X-ray crystallography or NMR techniques. As a result, computational chemists often try to develop pharmaceutical drugs for receptors whose structure is unknown [9, 37]. The starting point in this case is a collection of ligands that have been experimentally discovered to interact with the considered receptor. By examining the chemical properties and the possible shapes of these ligands, chemists seek to identify a set of features embedded in $\Re^3$ that is contained in some active conformation of each (or most) of the ligands. This is called the *pharmacophore* and it is considered responsible for the observed drug activity. The features of the pharmacophore interact with features of the receptor, while the rest of the ligand acts as a scaffold. Once a pharmacophore has been isolated, it can be used to further improve the activity of pharmaceutical drugs [23].

This paper considers the following problem: Given a set of ligands that interact with the same receptor, find geometric invariants of these ligands, i.e., a set of features embedded in $\Re^3$ that is present in one or more valid conformations of each of the ligands. We refer to this problem as the pharmacophore identification problem since the pharmacophore is such an invariant. Solving this problem requires dealing efficiently with large amounts of spatial data and shape information. Ligand

[*] This research is supported by a grant from Pfizer Central Research.

[†] Pfizer Central Research, Sandwich, U.K.

[‡] Department of Computer Science, Rice University, Houston, TX 77005. Partially supported by startup funds from Rice University.

[§] Department of Computer Science, Stanford University.

[¶] Department of Computer Science, Stanford University. 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.

[||] Department of Computer Science, Princeton University.

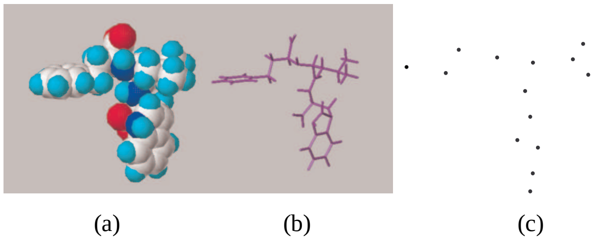

Figure 1: 1TMN: (a) space-filling, (b) stick, and (c) set-of-points models

molecules are very flexible and can assume many distinct potentially valid conformations. A valid conformation is a rigid spatial realization of the atoms of a molecule whose energy is below a predefined threshold [11]. It is not known, a priori, which of the many possible valid conformations of a ligand is the one that contains the desired pharmacophore.

Besides providing starting points for designing effective pharmaceutical drugs, there are several important applications where geometric invariant identification is useful. For example, the geometric invariants identified by our system can help in the formulation of database queries that may retrieve functionally equivalent, but structurally novel, molecules from molecular databases [44]. Geometric invariants can also suggest alignments of molecules for input to CoMFA (Comparative Molecular Field Analysis) and other 3D QSAR (Quantitative Structure-Activity Relationship) methods [9, 28]. These are methods that predict the activity of hypothetical compounds based on the assayed activity of previously synthesized ones.

This paper describes our efforts to prototype an integrated software system, called RAPID (RAndomized Pharmacophore Identification for Drug design), for addressing the pharmacophore identification problem. We present briefly the overall structure of RAPID in Section 2, and outline related work in Section 3. The two main modules of RAPID, conformational search and identification of invariants, are described in Sections 4 and 5. In this paper, we focus on the latter module and describe in some detail a randomized approach for solving an abstract geometric problem lying at the core of pharmacophore identification – the search for common subsets in a collection of point sets embedded in $\Re^3$. Our system attempts to deal with the numerous complications that arise in real examples. We report in Section 6 some preliminary experimental results. Finally, in Section 7, we conclude with a discussion of some open issues that merit consideration in future work.

2 Overview of RAPID

RAPID tries to identify geometric invariants among a collection of small ligand molecules like the molecule shown in Figure 1. This ligand is called 1TMN and it is one of the several inhibitors of the protease thermolysin. Figure 1(a) shows the space filling model of 1TMN. In this model a Van der Walls sphere [11] is considered around each atom center. Figure 1(b) shows the corresponding stick model in which only chemical bonds are drawn. The degrees of freedom of ligands 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 degrees of freedom are considered since these are the ones that exhibit large variations in their values [11]. Figure 1(c) shows the conformation of 1TMN of Figure 1(a) as a set of points in $\Re^3$. These points may represent atom centers or groups of atoms aggregated to one point endowed with a feature common to all these atoms (e.g., a rigid benzene ring) [37]. We assume that once a conformation is given, one can automatically transform it to a unique collection of points. For the purposes of this paper, a conformation C may represent a valid geometric embedding of a molecule where bonds are retained, or a set of points without bond information.

In RAPID, the identification of geometric invariants in a collection of flexible ligands denoted by $M = \{M_1, M_2, \ldots, M_N\}$ is treated as a two-stage process[1] addressing the two following problems:

Problem 1 (Conformational Search) *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 distance function.*

Problem 2 (Invariant Identification) *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 $\Re^3$ with the property that for all $i \in \{1, \ldots, n\}$, there exists some $C_{ij} \in \mathcal{C}(M_i)$ such that S is congruent to some subset of C_{ij}. A solution S, if it exists, is called an* invariant *of M.*

Although at this stage the two modules of RAPID work independently, we plan to support their interaction as the system develops. For example, the invariant identification module will be able to request from the conformational search module conformations that contain certain features of the molecule in specified spatial positions. RAPID also needs to deal with a host of complicating factors. In practice, the input may contain ligands that do not bind firmly to a receptor and do not contain the pharmacophore. This requires us to consider a

[1] A third module of RAPID, currently under development, involves the computation of molecular surfaces (see [20, 27]).

relaxation of Problem 2 above, where a geometric invariant need only be present in conformations of some K of the N molecules. Furthermore, an invariant may be perturbed in its spatial layout in distinct ligands' conformations, or there may be a large number of spurious invariants of differing sizes that do not correspond to a pharmacophore. Each of these contributes to a combinatorial explosion in the search process for the invariants.

3 Related Work

As far as conformational search is concerned, both systematic and randomized techniques are being investigated [10, 32, 33]. Systematic search methods sample each torsional degree of freedom of the ligand at regularly spaced intervals and minimize all conformations produced [34]. These techniques can be prohibitively expensive [32] and several heuristics are employed to quickly prune away conformations that are "close" to previously generated conformations [43]. Randomized methods work as follows: conformations are obtained by applying random increments to torsional degrees of freedom of the molecule starting with a user-specified initial conformation [22], or with a previously discovered low-energy conformation [14]. Recent articles, which attempt to compare different methods, emphasize the superior quality of the results obtained with randomized methods [22]. Other considered methods produce low-energy conformations, which obey distance constraints, using inverse kinematics [36], algebraic methods [24], or distance geometry techniques [17]. Note that the *protein-folding* problem is also a conformational search problem but its large size prohibits the use of the above techniques (see [18]).

Invariant identification is related to the well-studied problem in geometric optimization [1, 3, 2, 30] of finding common point sets. Determining the congruence of two point-sets in $\Re^3$ is tractable [1, 3] in the absence of complications such as noise. However, invariant identification is more closely related to the problem of identifying the *largest common point set* (LCP). Unfortunately, the LCP problem turns out to exceedingly difficult; in fact, even for m collections of n points on the real line, the LCP cannot be approximated to within an n^ϵ factor unless $P = NP$, and only weak positive results are known [2, 30]. Of course the problem is polynomially solvable when $m = 2$ [3] but that is not good enough for our purposes. Also, there are constraints in the molecular structures, but there are also various complications as discussed earlier.

In the computational chemistry literature, the most popular algorithms for invariant detection are based on clique-detection. For instance, DISCO [37] initially considers a pair of conformations c_1 and c_2 belonging to different molecules and constructs a "correspondence graph" G. The nodes of G are all node pairs, and an edge is created if the pairs in each of the connected nodes can be matched simultaneously. A clique detection algorithm [13] is then used to find cliques in G. These correspond to invariants in c_1 and c_2, and thus to candidate pharmacophores. Maximum clique detection is NP-hard [21], but this algorithm seems to work well in practice [37, 44]. The approach has been generalized to n conformations by choosing a reference conformation and comparing it with the other $n-1$ conformations. Unfortunately, if a large number of conformations per molecule are to be considered, there is a tremendous blow-up in the number of primitive operations performed by such algorithms [9], rendering them impractical. Given this situation, several new approaches are under development. For instance, one idea is to start with small invariants (2-3 features) and gradually expand them [9]. Our approach is fairly different and involves heavy use of randomized search techniques [39].

4 Conformational Search

The goal of conformational search as defined in Problem 1 (Section 2) is to produce a number of distinct low-energy conformations of a given ligand. We proceed as follows. Initially a large number of conformations are generated at random. In contrast with previous randomized search methods, we obtain a random conformation by selecting each degree of freedom from its allowed range according to a user-specified distribution. This distribution is frequently the uniform distribution. However, if some a priori information is available about the preferred values of a particular degree of freedom [31], then the corresponding values are selected according to a distribution that reflects the a priori information (i.e. a Gaussian distribution). An efficient minimizer [12, 40] is then used to obtain conformations at local energy minima. Minimization is the most time-consuming step during conformational search, so we have carefully optimized this procedure.

To obtain a representative set of conformations from our sample, we partition it into sets that reflect geometric similarity as captured by the distance measure DRMS. We define DRMS(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 (see Section 5). 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. We omit further details of the clustering, but note that the implementation employs heuristics for improving the quality of output, e.g., it is ensured the cluster center is close to the average cluster conformation and a post-processing step checks that all cluster centers are sufficiently far apart. While we obtain reasonable experimental results, we are investigating optimizations via algorithms that minimize the maximum intercluster distance [25] and incremental clustering techniques [15].

Our experience with randomized techniques for searching high-dimensional spaces has shown that randomized exploration is superior to systematic exploration when the shape of the underlying space is irregular [29]. 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 molecule (see also [22]). This has been our main motivation for the development of the randomized conformational search procedure described above.

5 Identification of Invariants

The set of cluster centers, denoted by $\mathcal{C}(M) = \mathcal{C}(M_1) \cup \ldots \cup \mathcal{C}(M_N)$ is the input for the invariant identification module. Each conformation in $\mathcal{C}(M)$ is now represented as a set of labeled points in $\Re^3$ (see Section 2). We wish to determine a structure S that is congruent to a substructure of some conformation in every molecule. The congruence relation is with respect to 3-D rotations and translations that ensure equality of labels.

Our formulation of the invariant identification problem in Problem 2 (Section 2) assumes noise-free data, specifically that all point positions are known exactly . In practice, atom positions are fuzzy, and it may not be possible to align them exactly. Therefore, we adopt the convention that two points p_1 and p_2 are said to *match* when $\mid p_1 - p_2 \mid \leq \epsilon$, where ϵ is the *point location error*. Similarly, two triangles are said to be congruent if each point in the first triangle is within ϵ of its corresponding point in the second.

The invariant identification problem is a variant of the *largest common point set problem (LCP)* in three dimensions, which is the following. Given s point sets $P_1, P_2, \ldots, P_s$ in $\Re^3$, determine the point set of maximum cardinality congruent to some subset of each point set. For convenience, we assume that each point set P_i has cardinality exactly n. For arbitrary s and dimension d, LCP is hard to approximate within a factor of n^ϵ, for some $\epsilon > 0$ [2]. In the sequel, we consider the following version of LCP, called LCP-α: determine a point set S of size $|S| \geq \alpha n$ congruent to some subset of each $P_i, 1 \leq i \leq s$. The motivation for focusing on this subproblem is that it more accurately captures the primary application, where pharmacophores are desired to have a certain minimum size.

We begin by focusing on the case $s = 2$, and later, in Section 5.2 indicate how to generalize for any value of s.

5.1 Phase 1: Pairwise Matching

In this section, we focus on the invariant identification problem for two point sets, denoted by MATCH. This problem has been studied extensively in the literature [41, 42, 4]. 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 two and three dimensions was presented in [7]. However, for general α, the best known algorithms were obtained by [3]. 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 three dimensions. For two dimensions, the corresponding bounds obtained are $O(n^{3.2})$ and $O(n^{2.2}/\alpha)$. However, these bounds apply only for the noise-free model of point sets. The noisy version of the problem was considered in [5] yielding an $O(n^8)$ algorithm for the two dimensional version.

In this section, we describe two random-sampling schemes for solving LCP-α on noisy data. The first is quite natural, and the second uses more careful random sampling. Note that in these algorithms, we use the notions of congruence and matching that incorporate noise. However, the analysis assumes that the data is exact.

In the sequel, we use the notation $g(n) = \tilde{O}(f(n))$, where f and g are functions, to indicate that $g(n) = O(f(n) \log n)$. Also note that in three dimensions, a unique transformation T (upto reflection) between two point sets P_1 and P_2 is determined by matching three points p, q, and r in P_1 with three points s, t, and u in P_2. Furthermore, it is known that [19]:

Proposition 1 *Given a transformation T from P_1 to P_2, in time $\tilde{O}(n)$ we can determine for each point $p \in P_1$ a corresponding point $p' \in P_2$ such that $|T(p) - p'|$ is minimized.*

The basic random sampling method is as follows.

BASIC-SAMPLE: For some constant c, perform $(c \log n)/\alpha^3$ iterations of the following sampling process: sample a triplet of points $< p_1, p_2, p_3 >$ randomly from P_1; determine three points in P_2 congruent to this set; compute the resulting induced transformation and determine the number of points in P_1 matching corresponding points in P_2; and, if the number exceeds αn, declare SUCCESS.

Theorem 2 *Given a common subset S of size $|S| \geq \alpha n$, the probability that BASIC-SAMPLE fails to declare SUCCESS is at most $1/n$.*

The proof of this theorem is straightforward. Observe that the probability of picking three points that belong to the common substructure is at least α^3. Applying a Chernoff bound [39] yields the desired probability.

Theorem 3 *BASIC-SAMPLE runs in time $\tilde{O}(n^{2.8}/\alpha^3)$ using space $O(n^2)$.*

The running time can be written as $T = n\sum_{i=1}^{1/\alpha^3} \Delta_i$ where Δ_i is the number of candidate transformations generated during iteration i. Let $H_P(t)$ denote the number of triangles congruent to t in point set P (we follow the notation of [3]). Clearly, $\Delta_i = H_{P_2}(t_i)$, where t_i is the triangle induced by the triplet chosen at iteration i. In [3], it is shown that $H_P(t) = O(n^{1.8})$. This yields a bound of $\tilde{O}(n^{2.8}/\alpha^3)$ on the running time for BASIC-SAMPLE.

Run-time profiling revealed that BASIC-SAMPLE examines many spurious triples (i.e., those that do not yield a large invariant). We propose the following modification of the random sampling procedure to fix the problem.

PARTITION-SAMPLE: For some constant c, perform $c \log n$ iterations of the following sampling process: randomly select two subsets A and B of size $1/\alpha$ from P_1; also select a subset C of size $1/\alpha$ from P_2; store all distances $d(p, q)$, for all $p \in C$ and $q \in P_2 - C$, in a hash table; for every triangle (a, b, p) with $a \in A, b \in B$, and $p \in P - (A \cup B)$, probe for $d(p, a)$ and $d(p, b)$ in the hash table to determine all matching triplets (c, p_1, p_2) with $c \in C$ and $p_1, p_2 \in P_2 - C$; finally, as before, if the resulting transformation induces a match of more than αn points, declare SUCCESS.

Theorem 4 *Given a common subset S of size $|S| \geq \alpha n$, the probability that PARTITION-SAMPLE fails to declare SUCCESS is at most $1/n$.*

Proof: Let the transformation yielding S be T. The following statement is clearly true: If $S \cap A \neq \emptyset$ and $S \cap B \neq \emptyset$,then this trial yields a transformation that finds S with probability approximately α.

We try all triplets $(a, b, p), a \in A, b \in B, p \in P_1 - A - B$. Since S is large, at least one of these is a triangle wholly contained in S. We do not find the corresponding triangle in P_2 only if C contains the image of a or b under T (which happens with probability $\leq 1/n$), or if C does not contain any point from S, (which happens with probability $(1-\alpha)^{1/\alpha}$).

The probability f that $A \cap S = \emptyset$ is at most $(1-\alpha)^{1/\alpha} \leq e^{-1}$. The probability that a single random trial fails is the probability that $(A \cap S = \emptyset) \vee (B \cap S = \emptyset) \vee ((P_1 - A - B) \cap S = \emptyset)$. For $S \geq 4$, this probability is bounded by $2f + f^2$ which is constant. Therefore, by repeating $c \log n$ times, we obtain a probability of error $F \leq 1/n$.

Theorem 5 *PARTITION-SAMPLE runs in time $\tilde{O}(n^{3.4}/\alpha^3)$ using space $O(n/\alpha^2)$.*

Proof: The number of candidate triangles $t_1, t_2, \ldots, t_k$ generated by iteration from P_1 is $n \cdot 1/\alpha \cdot 1/\alpha = n/\alpha^2$. Given a triangle t_i, and a point $c \in C$, let $H^c_{P_2}(t_i)$ be the number of triangles in P_2 incident on c and congruent to t_i. The running time of the algorithm can now be given by the following expression:

$$\begin{aligned} T(n) &= O(n/\alpha^2 + \sum_{c \in C} \sum_{i=1}^{k} n H^c_{P_2}(t_i) + \\ &\qquad n/\alpha^2 \sum_{c \in C} n^2). \end{aligned}$$

The first term is the cost of generating the triangles $t_1, \ldots, t_k$. To estimate $H^c_{P_2}(t_i)$, we need to count the number of triangles in P_2 which share the same point c and are congruent to t_i. For any such triangle $\{c, p_1, p_2\}$, let $d_1 = ||c - p_1||, d_2 = ||c - p_2||, d_3 = ||p_1 - p_2||$. Let S_i be a sphere centered at c of radius $d_i, i = 1, 2$. Consider any $\{c, p_1, p_2\}$ congruent to t. Without loss of generality, assume that $p_i \in S_i, i = 1, 2$. Now, once we fix p_1, all such points p_2 lie on the intersection of S_2 and a circle of radius d_3 centered at p_1. Therefore, the number of such points (and also the number of congruent triangles fixed at c and p_1 is the number of incidences between points of P_2 and this fixed–radius circle. Hence, the desired total number of such triangles fixed at c is merely the maximum number of incidences between n circles of fixed radius and n points in $\Re^3$. If we choose an arbitrary direction and project this down to two dimensions, the problem reduces to counting the number of incidences between n ellipses and n points in the plane, which is $O(n^{1.4})$ (from [16]).

Inserting this bound yields:

$$\begin{aligned} T(n) &= O(n^3/\alpha^3 + n \sum_{c \in C} \sum_{i=1}^{k} H^c_{P_2}(t_i)) \\ &= O(n^3/\alpha^3 + n^2/\alpha^3 \cdot n^{1.4}) \\ &= O(n^{3.4}/\alpha^3). \end{aligned}$$

Although the asymptotic running time of PARTITION-SAMPLE is worse than that of BASIC-SAMPLE, experiments (discussed in Section 6) reveal that PARTITION-SAMPLE consistently outperforms BASIC-SAMPLE, generating far fewer spurious triples with an improved degree of success. Intuitively, the modified partitioning favors "large" solutions over small ones. Given a solution S of size K in P_1, the probability that each of A, B, and $P_1 - A - B$ contain at least one point of S increases with K.

Additionally, experimental results suggest that the predicted running times are fairly high. In Section 6 we also discuss possible reasons for this, in terms of actual distance distributions among molecules.

Eliminating Redundant Solutions. An invariant search algorithm 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, we need to *prune* them. Given invariants S_1

and S_2, we can check if $S_1 \subseteq S_2$ by invoking the algorithm on these two sets with α set to 1. Clearly, since α is 1, a solution will be produced only if S_1 is contained in S_2. Essentially, this stage induces many instances of geometric pattern matching, where we wish to determine whether a set of points P is congruent to some subset of a point set Q.

Inaccuracies in Computing the Transformation. The module at the core of the substructure search algorithm 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 $p_i \rightarrow p'_i$, for $i = 1, 2, 3$. Such a transformation can be computed in many different ways. The simplest such way is: **(i)** align p_1 and q_1; **(ii)** align the vectors $\overrightarrow{p_1 p_2}$ and $\overrightarrow{q_1 q_2}$; and, **(iii)** align p_3 and q_3 by a rotation about $\overrightarrow{p_1 p_2}$. This approach over-constrains the fit between the two triangles.

For our purposes we need to compute a transformation that preserves this initial set of correspondences (within distance ϵ), and matches the largest number of points from P to Q. Since atom locations are noisy, such a transformation may not be unique, i.e. there may be different transformations that yield maximal but incomparable common subsets. In general, if we consider the six-dimensional space of all rigid transformations in $\Re^3$, the set of transformations satisfying the given correspondences is actually an enclosed volume over which we must maximize the transformation *score*.

If we knew the correspondences between points in P_1 and P_2, then a "balanced" transformation could be obtained by computing the transformation minimizing DRMS(P'_1, P'_2), P'_1 and P'_2 being the subsets of P_1 and P_2 that are matched to each other [6] (see Section 4 for the definition of DRMS). This suggests the following refinement: Use the three-point procedure above to generate a plausible correspondence list, and then use a DRMS minimizing procedure to compute a refined transformation.

However, this approach is more time-consuming, and it is possible that DRMS minimization merely replaces one form of over-constraining by another. Our implementation employs various heuristics to minimize the effect of this problem. One such heuristic is the following: Determine a *seed* transformation T by any of the above methods, and then sample three random pairs from the set of correspondences that T induces, using these pairs to construct a new transformation. Clearly, in a perfect world, we will obtain T again. However, given the inaccuracies in point location, it turns out that some choices of triplets may yield a larger number of correspondences than before.

5.2 Phase 2: Multiple Matching

The previous phase yields methods for solving LCP-α on two sets. In Phase 2, candidate solutions obtained from the previous phase are tested against the remaining molecules to determine the invariant. Each MATCH call operates on two conformations. Since each molecule is represented by many conformations, we extend MATCH to two molecules by doing all pair-wise matches between the sets of conformations. Note that comparing a candidate solution against a new conformation may result in $0, 1,$ or many solutions, since the solution may decompose into smaller pieces on comparison.

There are various strategies one could use to process multiple molecules. A simple strategy does the obvious linear merging. We take each solution and compare it with the next molecule. We do this for all current solutions, concatenate and prune the results, and repeat with a new molecule. Another approach, which is in some ways less order-sensitive, is a tree-based merge. Here, we run the two-molecule algorithm on distinct pairs of molecules, and recursively combine the results using a binary tree. Both approaches are simple and correct, if point location error is not considered.

In addition to this, we may wish to find an invariant that does not exist in *all* the molecules, but in some fixed number of them. Our current strategy would fail to do this because all candidate solutions are compared against every molecule. 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.

If an invariant S is contained in a point set P, then a call to MATCH(S, P) should return a set of solutions containing S. This can easily be checked, which means that we can verify whether an invariant matches against a conformation. Our marking algorithm associates *marks* with each candidate solution. All original conformations are initialized with zero marks. All invariants that result from a MATCH acquire the sum of the marks of the input structures. Every time a solution fails to match against a new molecule (which means that it fails to match with *any* conformation of that molecule), we add a mark to it and continue to propagate it. We reject any solutions for which the number of marks exceeds the prescribed maximum.

6 Experimental Results

This section reports experimental results for the algorithms described above. All reported timings are on an SGI Indigo2 with a 175 MHz MIPS R8000 processor and 384MB RAM. Code was written in C/C++, and compiled using SGI CC.

Input. In Figure 2, we show four different inhibitors of the protease thermolysin. These molecules fit into the same cavity of thermolysin and by their presence inhibit the activity associated with that cavity. This example was chosen because all the inhibitors have been crystallized with thermolysin and their active conformations are known and recorded in the PDB database (a database of three-dimensional protein structures

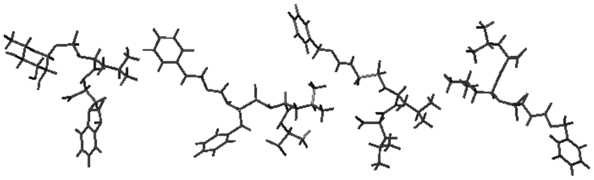

Figure 2: 1TLP, 4TMN, 5TMN, and 6TMN are inhibitors of thermolysin.

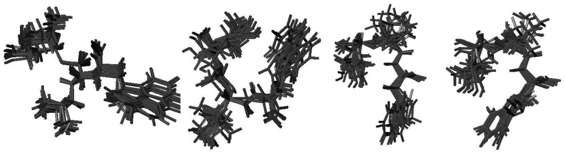

Figure 3: Different clusters of 1TLP.

stored at the Brookhaven National Laboratory). Note that 1TLP has 69 atoms and 10 torsional degrees of freedom, 4TMN has 68 atoms and 15 degrees of freedom, 5TMN has 64 atoms and 13 degrees of freedom, and 6TMN has 63 atoms and 12 degrees of freedom.

Conformational Search. Each of the molecules in Figure 2 was run through our conformational search software. A THRESHOLD value of 20 Kcals/mol was used for the energy of the valid conformations. It took 4.3 hours to to produce 4000 valid conformations for each molecule and reduce these to a set of representatives. The timings reported are very reasonable if one takes into account that 10, 15, 13, and 12 dimensional spaces are searched. The number of representatives produced was 850, 1192, 1024, and 955 for 1TLP, 4TMN, 5TMN, and 6TMN, respectively. The TOLERANCE value was set to 1.23 $\mathring{A}$ for all runs. A few clusters of 1TLP are shown in Figure 3. Since we know the active conformations of all the molecules, we performed a consistency check at the end of conformational search and we confirmed that the active conformations were close to one of the produced representatives.

An important issue in conformational search is to decide the number of valid conformations to produce. At this stage, this number is determined experimentally: we stop producing new conformations when these fall into existing clusters and do not increase significantly the overall number of clusters.

Identification of Invariants. To find the invariants in these four molecules, the conformations produced by the first stage were given as input to the search algorithm described in Section 5. For these tests, we considered each non-hydrogen atom to be a separate feature or point. Thus, each conformation had approximately 30 features drawn from 6 feature classes (the feature classes here being the atom types). We define the "solution" to be the overlapping portions of the molecules when aligned as shown in Figure 4. This is the upper right hand T-shaped portion of this diagram. The entire invariant consists of roughly 7 atoms and an additional 7 atoms of "scaffolding," or connecting atoms with no pharmacophore functionality.

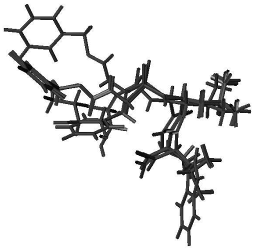

Figure 4: 1TLP, 4TMN, 5TMN, and 6TMN overlapped in their active conformations.

Table 1 details the parameters used for invariant identification. The parameters ϵ, δ, and α are as described in Section 5. We set at 5 the number of times the MATCH algorithm is run per comparison, or the maximum number of invariants found per MATCH. These parameters were chosen by experimentation. For example, δ was selected to be much lower than ϵ. We found that higher values of δ increased the running time without improving the quality of the solution.

For the sake of brevity, we omit detailed experimental analysis of variants of Multiple Search and use a linear search to compute invariants, with no failures in matching invariants permitted.

In Figure 6, we compare the two schemes experimentally. The running time of the two algorithms (measured in terms of the number of transformations produced) is plotted against α for varying α. We see that as α gets smaller, PARTITION-SAMPLE clearly performs far better than BASIC-SAMPLE. At higher values, their behavior is more alike.

We also present in Table 2 a comparison between the two schemes on large data sets, where we vary the number of conformations per molecule. We include the active conformation in our sets. In all cases, the quality of solutions (in terms of the largest solution found) is comparable, and PARTITION-SAMPLE consistently runs faster than BASIC-SAMPLE. When the number of conformations increases, more invariants are produced because some of the added conformations have additional "scaffolding" which also could be matched. Yet, in every case, the invariant of the largest size is the "correct solution." Many non-trivial solutions of size $8 - 14$ are also produced.

Parameter	Search Value	Prune Value
ϵ	1.3	3.5
δ	0.5	1.0
α	0.3	1.0

Table 1: Parameters for invariant search trials.

Although we prove worst-case running times for invariant search that are quite high, the algorithms run very quickly in practice. One source of discrepancy is in the analysis of the algorithms, where we use the best-known combinatorial upper bounds on the multiplicity of a fixed triangles induced by a set of points in $\Re^3$. However, these bounds are not tight, and actual molecules tend not to exhibit worst case behavior. In Figure 5, we plot the distribution of inter-atom distances in four of the molecules with which we experimented. For each molecule, the upper-most graph indicated the complete inter-point distance distribution. The lower graphs plot the distributions induced by the points chosen during random sampling (for various values of α. Not surprisingly, the distribution becomes flatter as α increases. What is also interesting is that the figures indicate that all these molecules have a bounded diameter, and that the distribution is quite similar to a Poisson distribution. These facts indicate that a more careful analysis incorporating this information might lead to a better running time prediction for the algorithms, and possibly better algorithms.

7 Discussion

Our goal is to optimize the modules of the system to perform experiments that involve 5-20 ligands and a large number of conformations per ligand. As far as conformational search is concerned, we are investigating dynamic clustering techniques. By looking at the relative changes in the number of clusters found, we hope that we will be able to automatically end the conformational search stage. For the invariant identification stage, we will try to reduce the dependence of the implementation on the ordering of the input, a phenomenon observed in the current system. Our short-term plans also include the implementation of an algorithm that will extract features from conformations in the way that the computational chemists understand them. Currently, each atom gives rise to a feature that bears as a label the corresponding atom type. Computational chemists, however, have rules to group atoms into features that describe the chemical behavior of parts of the molecule, e.g., hydrophobic parts and positively charged parts. Features will reduce the number of points we consider per conformation and will hopefully help us deal with larger examples.

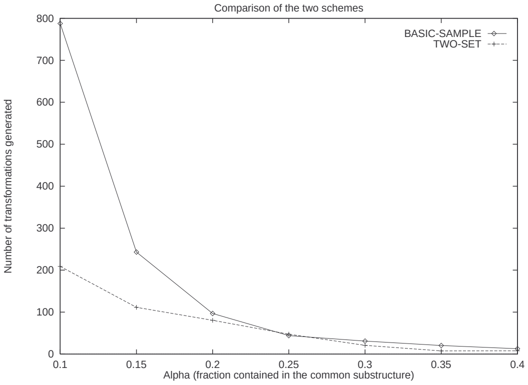

Figure 6: Relative performance of BASIC-SAMPLE and PARTITION-SAMPLE.

In Section 5 we mentioned certain combinatorial properties of point sets in two and three dimensions. Most of these properties are well understood for exact models i.e models in which all point locations are known perfectly. It would be interesting to determine analogous properties for noisy data sets. These would contribute directly to improving the worst case running time of the algorithms we have presented. In addition, it would be instructive to examine the average case behavior of these algorithms in the light of the remarks made previously about distance distributions in molecules.

8 Acknowledgements

The authors would like to thank Piotr Indyk and Dan Halperin for many helpful discussions.

References

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

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

[3] T. Akutsu, H. Tamaki, and T. Tokuyama. Distribution of distances and triangles in a point set and algorithms for computing the largest common point set. *These proceedings*, 1997.

[4] H. Alt and L. Guibas. Discrete Geometric Shapes; Matching, Interpolation and Approximation. A survey. Manuscript 1996.

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

[6] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares Fitting of Two 3-D Point Sets. *IEEE Trans. Pattern. Anal. Machine Intell.*, volume PAMI-9, No. 5, September 1987.

[7] M. D. Atkinson. An optimal algorithm for geometric congruence. *J. Algorithms* 8 (1987). pages 134–183.

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

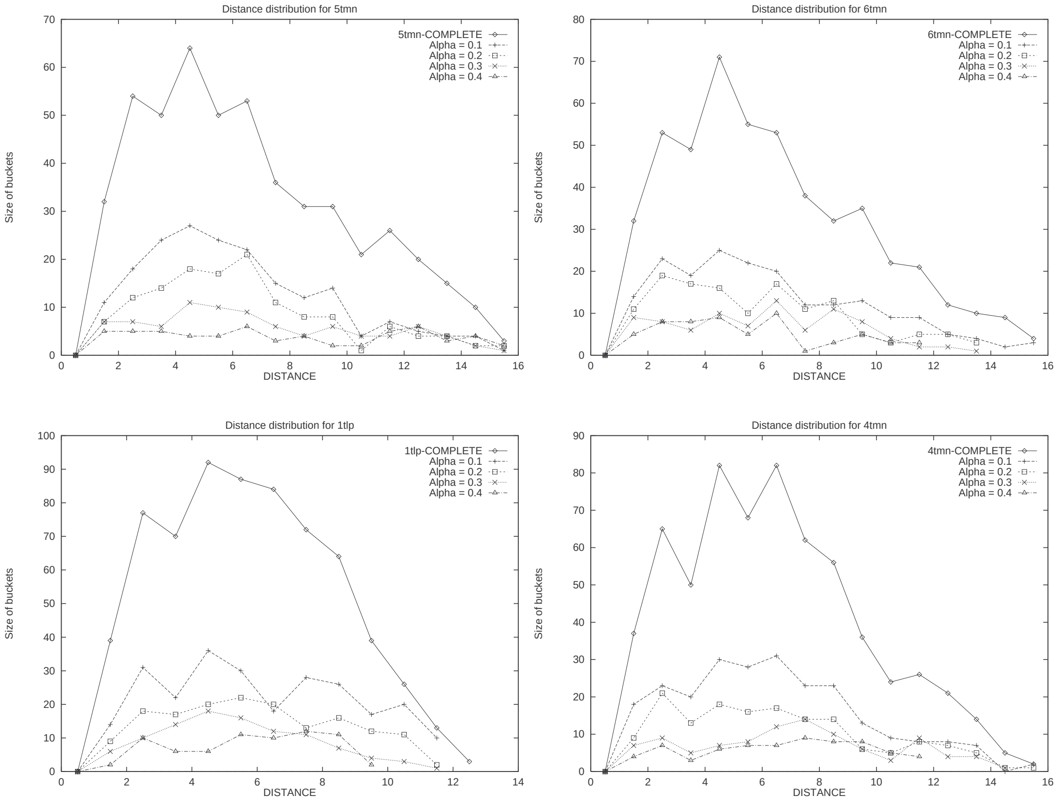

Figure 5: Distance distribution plots for 1TLP, 4TMN, 5TMN, and 6TMN

Number of	Time to	Histogram of Solution Sizes											
Conformations	Completion (in s)	4	5	6	7	8	9	10	11	12	13	14	15	total
BASIC-SAMPLE														
1	4.62		2	1								1		
11	2127.66	90	34	20	5	6	4	1	1	1				
21	24175.42	289	110	46	35	42	18	3	6	2	2	1		
PARTITION-SAMPLE														
1	2.85			1		1					1			
11	656.47	20	9	8	11	11	3							
21	8060.00	73	55	53	30	21	25	7	8	1	1	4	1	

Table 2: Comparison of BASIC-SAMPLE and PARTITION-SAMPLE

.

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

[10] 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.

[11] 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.

[12] R. Brent. *Algorithms for finding zeros and extrema of functions without calculating derivatives*. PhD thesis, Stanford University, 1971.

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

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

[15] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental Clustering. To appear in *Proc. 29th Annual ACM Symposium on Theory of Computing*, May 1997.

[16] K. Clarkson, H. Edelsbrunner, L. Guibas, M. Sharir, and E. Welzl. Combinatorial Complexity Bounds for Arrangements of Curves and Spheres. *Discrete Comput. Geom.* 5:99–160 (1990)

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

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

[19] D. Dobkin and R. Lipton. Multidimensional search problems. *SIAM J. Computing* 5:181–186 (1976).

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

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

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

[23] 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 Chemistry*, 38:3566–3580, 1995.

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

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

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

[27] D. Halperin, C. Shelton. A perturbation scheme for spherical arrangements with application to molecular modeling. *This proceedings, 1997*

[28] A. Itai, N. Tomioka, M. Yamada, A. Inoue, and Y. Kato. Molecular superposition for rational drug design. In Kubingi, editor, *3D QSAR in Drug Design: Theory, Methods, and Applications*. ESCOM, 1995.

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

[30] S. Khanna, R. Motwani, and F. F. Yao. Approximation algorithms for the largest common subtree problem. Technical Report STAN-CS-95-1545, Department of Computer Science, Stanford University, 1995.

[31] 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.

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

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

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

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

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

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

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

[39] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

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

[41] P. Raghavan and S. Irani. Combinatorial and experimental results for randomized point matching algorithms. *Proc. 12th. ACM. SCG*, pp. 68-77, 1996.

[42] P. J. Rezende and D. T. Lee. Point set pattern matching in d-dimensions. *Algorithmica*, 13, 387-404, 1995.

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

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