PharmID: Pharmacophore Identification Using Gibbs Sampling

Jun Feng, Ashish Sanil, and S. Stanley Young*

National Institute of Statistical Sciences, P.O. Box 14006,
Research Triangle Park, North Carolina 27709-4006

Received September 28, 2005

The binding of a small molecule to a protein is inherently a 3D matching problem. As crystal structures are not available for most drug targets, there is a need to be able to infer from bioassay data the key binding features of small molecules and their disposition in space, the pharmacophore. Fingerprints of 3D features and a modification of Gibbs sampling to align a set of known flexible ligands, where all compounds are active, are used to discern possible pharmacophores. A clique detection method is used to map the features back onto the binding conformations. The complete algorithm is described in detail, and it is shown that the method can find common superimposition for several test data sets. The method reproduces answers very close to the crystal structure and literature pharmacophores in the examples presented. The basic algorithm is relatively fast and can easily deal with up to 100 compounds and tens of thousands of conformations. The algorithm is also able to handle multiple binding mode problems, which means it can superimpose molecules within the same data set according to two different sets of binding features. We demonstrate the successful use of this algorithm for multiple binding modes for a set of D2 and D4 ligands.

INTRODUCTION

The binding of a small molecule into a protein is inherently a 3D event. More often than not, we do not have a crystal structure of the protein to guide drug design efforts, so it is useful to have methods that can work from biological assayed potency data to suggest one or more 3D pharmacophores, the arrangement of key binding features in space. It is important to know the binding conformation and the key features that allow a compound to bind to a protein. A 3D pharmacophore can be used to guide the selection of compounds to screen—virtual screening—and to guide lead optimization efforts.[1−5]

What are the problems? First, small molecules typically have rotatable bonds and often have flexible rings, so they can take any of a vast number of shapes. Each shape will put atoms into different relative positions in space. We need to know the binding conformation for each molecule. Also, if there are multiple binding modes or binding sites, we also need to know which molecules need to be put together as they are binding in the same way. Next, we need to identify the key features that interact with the target protein in the protein-binding pocket. Both problems, determination of the binding conformation and its key features, can have extremely large search spaces. For example, a molecule with five rotatable bonds and three flexible six-member rings might be represented with $3^5 \times 2^3 = 1944$ distinct conformations. Much effort has been expended in selecting features to represent the biologically important characteristics of a molecule.[6]

One method of feature representation is to create a bit string where each bit gives the presence or absence, 1/0, of a feature in the conformation of the molecule. There are a number of features that medicinal chemists have considered important. A molecular feature can include two atoms (or more complex groups of atoms) and the through-space distance between them. Typical molecular groups include the following: hydrogen-bond donor, hydrogen-bond acceptor, positively charged group, negatively charged group, lipophilic group, aromatic ring center, and so forth. The distance is often given as a distance range, for example, 3−5 Å. The default settings of six pharmacophore groups and four distance bins of 2−4.5, 4.5−7, 7−10, and 10−14 give rise to 84 pharmacophore-pair/distance features. When these features are used, each conformation of a molecule can be represented as a bit string of 84 bits. A richer list of features; more distance bins; or extensions to three, four, or five features can vastly increase the number of bits in the string used to represent a conformation. For a flexible molecule, thousands of bit strings can be generated. The 3D superimposition problem can be initiated using a 1D bit string alignment. We define "alignment" as selecting one bit string per molecule and maximizing a number of specific bits in common. We assume that, for a set of conformers (one per ligand) to be superimposed on a common subset of molecular groups, their corresponding bit strings must have common subsets of bits (representing the presence or absence of a molecular feature) turned on. Therefore, if generated bit strings can be aligned very well for a set of conformers from different ligands, their 3D structures would more likely be able to be superimposed together. On the contrary, if their bit strings cannot be aligned optimally, it would be impossible for their 3D structures to be superimposed. Once bit strings are aligned, based on specific bits turned on for most conformers, common molecular features can also be determined. In summary, we need to search over two large search spaces, select one conformation for each ligand, and select a subset of bits from each string.

* Corresponding author phone: (919)782-2759; fax: (919)685-9310; e-mail: young@niss.org.

10.1021/ci050427v CCC: $33.50
Published on Web 02/03/2006

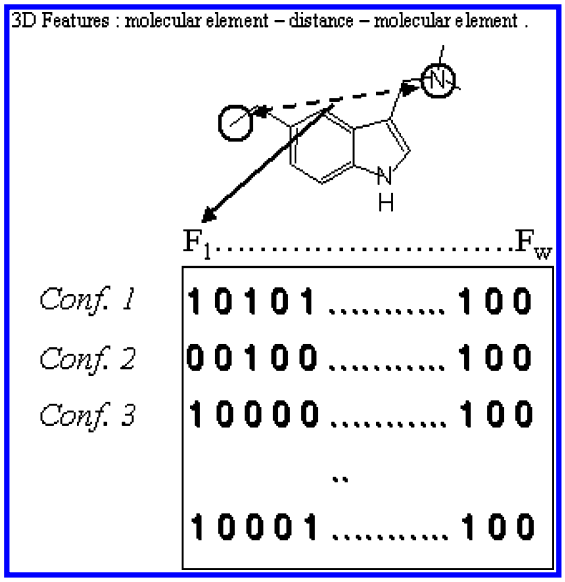

Figure 1. For each pair of molecular elements and binned distance, a bit is set or not depending on the presence or absence of the combination.

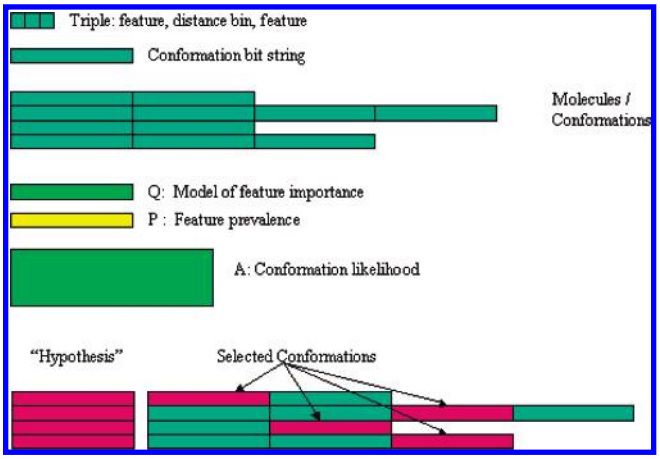

Figure 2. Data structure. Each feature is a triple of two molecular elements and the distance between them. The presence or absence of a feature is coded into 1/0. All the features are written as a bit string for each conformation. Each element of **Q** measures the importance of each feature. Each element of **P** measures the prevalence of each feature in unselected conformations or in a historical data set. **Q** and **P** are used with the molecular conformation to determine the likelihood that a conformation is the binding conformation. The selected conformations (hypothesis) and **P** are used to determine the importance of the features, **Q**.

METHOD

The algorithm includes four steps: feature definition, which generates a bit string for every conformer of every molecule; bit string alignment, which selects one bit string per molecule that has the most features in common between a molecule and the statistical model of the pharmacophore; hypothesis generation, which enumerates combinations of pharmacophore groups within selected conformers; and refinement, which examines all hypotheses over all conformers of all molecules. Our method for bit string alignment is a slightly modified version of the Gibbs sampling algorithm previously published for sequence alignment[7,8] whereby the most likely binding conformations and the key binding features are identified for each molecule. The bit string alignment is a modification of sequence alignment where there are only two types of residues (in our case, they are 0 and 1) and sequences are only allowed to move with the step size of k, where k is the width of each bit string. The resulting pharmacophore hypotheses, consisting of key binding features, are constructed with a clique detection method[9] within those most likely binding conformers in the bit string alignment stage.

Step 1. Feature Definition (See Figure 1). *Step 1a.* Six types of pharmacophore groups are predefined: hydrogen-bond donor, hydrogen-bond acceptor, aromatic ring center, hydrophobic center, negative charge center, and positive charge center. Certain groups can be defined as a vector, which means the angle between the direction or normal vectors must satisfy a predefined threshold (15° by default).

Step 1b. The Daylight SMART language is used to define pharmacophore groups. Users can change the rules on the basis of their understanding of the problem. For certain data sets, like ACE inhibitors, this is crucial for pharmacophore identification since there are some uncommon definitions for pharmacophore features, like zinc binding site.

Step 1c. The distance between each pair of features is computed. The distances are placed into bins. The default bins are 2−4.5, 4.5−7, 7−10, and 10−14. The user can change the definitions of the bins, add more bins, overlap bins, and so forth. The distance ranges for pharmacophore features are not determined by the definition of distance bins. During the pharmacophore construction stage, the selected conformers for all molecules will be superimposed, and the actual distance ranges will be calculated if the superimposition satisfies a user-provided root-mean-square deviation (RMSD) cutoff (0.5 Å, by default). The features and the observed distances between them are used to define the final pharmacophore.

Step 1d. A feature is defined as two pharmacophore groups and the distance (binned) between them. If a conformation has the feature, a certain bit is set to 1. If a conformation does not have the feature, that bit is set to 0. A bit string is used to represent each conformation. W features give rise to a bit string that is W bits wide for each conformation. More than two point features could be considered, for example, a triplet feature.

Step 2. Pharmacophore Fingerprints Alignment (See Figure 2). *Step 2a.* Consider N active molecules. The width of each conformation bit string is W. Each bit of a conformation bit string is "1" or "0", which notes if a specific pharmacophore feature is presented or not.

Step 2b. We place the conformation bit strings end to end to make a composite molecule bit string that has $m_i \times W$ elements where m_i is the number of conformations for molecule i.

Step 2c. For each position k between 1 and W on the bit string, the possibility vector **Q**, with elements $(q_{k,0}, q_{k,1})$, denotes the probabilities of 0 and 1 on position k (As $q_{k,0}$ and $q_{k,1}$ sum to 1, we store only one element). $c_{k,0}$ and $c_{k,1}$ denote the occurrence of 0 and 1 on position k for the currently selected binding conformation. Again, as the number of molecules is fixed at N, we need to keep track of only one **c** for each position. The background frequencies for 1 and 0 are p_1 and p_0, respectively, $p_1 + p_0 = 1$. These background frequencies are computed from the total number of 1's and 0's within the same data set. They may be taken as constants over all the features or computed specifically

for each feature, p_{1k} and p_{0k}. The alignment vectors (A_1, A_2, A_3, ... A_N) denote the likelihood of each conformer to be chosen for each molecule to form the alignment conformation set. Collectively, the A_i components form **A**, the conformation likelihood matrix. The elements of the probability vector **Q** can be calculated by occurrence vector **c** and the number of molecules N:

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B} \quad (1)$$

where $i = 1, 2, \ldots W$; $j = 0, 1$; b_j and B are called pseudocounts, and

$$B = \sum_{i=0}^{1} b_j = \sqrt{N}$$

and

$$b_j = \rho_j B$$

where ρ_j is the frequency of feature j (0 or 1) in the whole or a reference data set. The pseudocounts are quantities that capture baseline "prior information" in the underlying Bayesian statistical model. The above form that we adopt for b_j and B is considered to be a reasonable way to quantify this prior information. See refs 7 and 8 for details. More than capturing additional information, the pseudocounts provide a convenient, effective, and meaningful device for dealing with the awkward cases of $c_{k,j}$ being 0 (which would lead to $q_{k,j} = 0$ without the pseudocount correction). The reference data set could be the data set under consideration in the analysis or it could be some historical data set.

Step 2d. Start with a random alignment matrix **A**, and select one conformer bit string per compound according to the alignment matrix **A**.

Step 2e. Remove the bit string of one compound, *z*, from the alignment.

Step 2f. Use the remaining compounds to update probability vector **Q** using eq 1.

Step 2g. Calculate Q_x ($x = 1, 2, \ldots M_z$), which is the probability of generating every bit string for the removed molecule using the current probability pattern $q_{i,j}$, and P_x, which is the probability of generating every bit string for the removed molecule using background probability pattern p_j. The weight for each conformer bit string $A_x = Q_x/P_x$ is used to sample one conformer bit string for the removed compound. The conformer bit strings with higher weights (A_x) will have a higher chance to be picked.

Step 2h. Repeat the steps from step 2c to step 2g for all compounds over and over again until the scoring function $F = \sum_{i=1}^{W} \sum_{j=0}^{1} c_{i,j} \log(q_{i,j}/p_j)$ is converged. Other scoring functions could be used.

Step 2i. To speed up the whole procedure, in step 2g, we can pick the conformer bit string with the highest weight (A_x) instead of randomly sampling on the basis of likelihood. This may quickly converge to a local minimum, so multiple different starting alignment vectors should be used.

Step 3. Hypothesis Generation. Step 2 selects one conformer for each compound and a series of significant pharmacophore features and the distances between them. However, whether these conformers can be combined to have common three points, four points, or even higher-order pharmacophores needs to be determined.

Step 3a. In this step, all selected conformers are reexamined. A clique detection algorithm[9] is used to find all possible higher-order pharmacophores that contain the features selected by step 1. To be a consistent pharmacophore, the selected points have to be supported by the features and distances selected in step 1. Two points have one supported distance, three points (triangle) have three supported distances, four points (tetrahedral) have six supported distances, and so forth.

Step 3b. All pharmacophore hypotheses are saved into a list. The size of this list is a user-defined parameter. Ideally, this size should be as large as possible to accommodate all possible pharmacophore hypotheses.

Step 3c. There may be some molecules that do not participate in any pharmacophores as a result of picking the wrong conformer in the alignment stage (or the molecule may be binding in an alternative binding mode or in an alternative binding site). In the refinement stage, the program will search all conformers of each molecule that is not mapped to a pharmacophore to attempt to fit it into a pharmacophore hypothesis until it satisfies a hypothesis or all conformers are exhausted.

4. Refinement. *Step 4a.* After the hypothesis generation, a list of pharmacophore hypotheses is created. Every hypothesis is evaluated in the refinement stage.

Step 4b. Each hypothesis is evaluated to determine how many molecules can fit to it. If the current conformer cannot fit to this hypothesis, the program will discard the current conformer and pick another conformer and test it again until it finds one conformer that can fit to this hypothesis or until all conformers are discarded.

Step 4c. The refinement stage may find several hypotheses, each satisfied by different compounds. If several hypotheses are found, that may suggest multiple mechanisms.

RESULTS

Several examples are used to explicate our method. First, a simple example of synthetically constructed bit strings is used to demonstrate the Gibbs sampling step. Next, the algorithm is used to find the binding conformation and some of the key features for a small set of D2 ligands. The third example is a large data set of ACE inhibitors. Two pharmacophores are found that utilize 65 and 57 molecules from the set of 78 ACE inhibitors. Finally, selected D2- and D4-active compounds[10] are used to mimic a situation where there are two distinct binding modes or sites. All conformations were computed using Omega from OpenEye Scientific Software[11] using their default settings.

Example 0. A toy data set of 0's and 1's is used to demonstrate that the modified Gibbs sampling algorithm can successfully identify pharmacophoric features. There are 20 molecules, each with 20 features and 20 conformations. The probability of a feature is set low, ~5%, to mimic the sparse nature of bit string representations of conformations. A conformation for each molecule was selected at random, and bits were set to 1 for positions 1, 5, 10, and 18. So each "molecule" has one "binding" conformation with four features. The algorithm successfully found the binding conformation for each molecule. Figure 3 displays the

Molecule Conformation	Features ....
1_14	10001000010000000100
2_14	10001000010000000100
3_15	10001000010000000100
4_12	10001000010000000000
5_15	10001000010010000100
6_07	10001000010000000000
7_08	10001000010000000000
8_19	10001000010000000100

Figure 3. "Toy" data set constructed with 20 conformations, bit strings, for each of 20 molecules. Four of the 20 features were the pharmacophore, and one conformation for each molecule contained a 1 for those features. PharmID found the correct conformations and features for the toy data set. The selected conformations are shown for eight molecules.

selected conformation for eight of the molecules. All the active conformations were found. The basic algorithm can find conformations that have a consistent set of bits set.

Example 1. A 3D study of D2 and D4 ligands was undertaken by Böstrom et al.[10] They report a 3D pharmacophore. We submitted 21 of the more active D2 compounds to PharmID and reproduced their results, see Figure 4. This data set was chosen as it is considered an easy data set to solve.[12]

Example 2. A total of 78 ACE inhibitors taken from ref 13 with a total of over 46 000 conformations were submitted to PharmID. A total of 65 of the compounds satisfied the pharmacophore shown in Figure 5. A total of 57 of the molecules satisfied the conformation shown in Figure 6.

Example 3. A 3D study[13] of D2 and D4 ligands report 3D pharmacophores for D2 and D4 activity. Six of the compounds that were relatively selective for D2 compounds together with six of the compounds that were relatively selective for D4 were selected. Figure 7 gives a plot of D2 versus D4 activity. The compounds selected for analysis are marked. We submitted both types of compounds *together* to mimic a situation that might be expected to happen in practice where the experimenter is unaware that there are two binding modes or locations. The program reproduced the pharmacophores for both D2 and D4, see Figures 8 and 9. As would be expected, when the compounds were submitted to our program separately, the program individually found the same two pharmacophores. This data set was chosen to demonstrate that the program can find multiple pharmacophores within a single data set. Van Drie[12] points to the problematic nature of training data sets. There are two distinct situations. The data set can be ambiguous. There are multiple mathematical solutions, yet there is thought to be only one binding mode and binding site. Alternatively, there can be two binding modes or binding sites. It is generally believed that, if the data set contains molecules that bind in two different ways, then the problem will be difficult with existing analysis strategies. In fact, Catalyst specifically states that its HipHop algorithm will not work in this situation. We have confirmed that Catalyst fails on this data set. In this case, the program is able to deal with the multiple binding mode situations. This example was chosen to be representative of a potentially difficult problem.

Example 4. Thrombin. The thrombin data set is given in Patel et al.[4] There are five data sets in the original paper, and ligands in this dataset are more flexible than those in the other four data sets. We superimposed the protein structures with Sybyl, and then extracted ligand structures within the active sites of the corresponding receptors. Six pharmacophore groups were mentioned in the original paper, but only four pharmacophore groups (codes: B, H1, H2, and H3) are presented in all ligands. Therefore, we manually defined these pharmacophore groups with the SMART language prior to the pharmacophore searching. The solution based on the active conformations is generated using PharmID with only the crystal structure active conformation from each ligand. Since the receptors are superimposed, the ligands are also superimposed very well. PharmID only slightly adjusted the position for each ligand, forming a better superimposition of the active conformations. The definitions of the pharmacophore groups are shown in Figure 10. The conformations are generated by Omega from OpenEye Scientific using their default parameters. The energy window, which defines the upper bound above the global minimum, was set at 15 kcal/mol. The minimum RMSD between two different conformers is 0.6 Å. The maximum number of conformers for each ligand was set at 1000. For the thrombin data set, 6078 conformers are generated. Six out of seven ligands are very flexible and generated the maximum number of conformers allowed. Only one ligand, 1d4p, is rigid, and only 78 conformers are generated for it. To test the quality

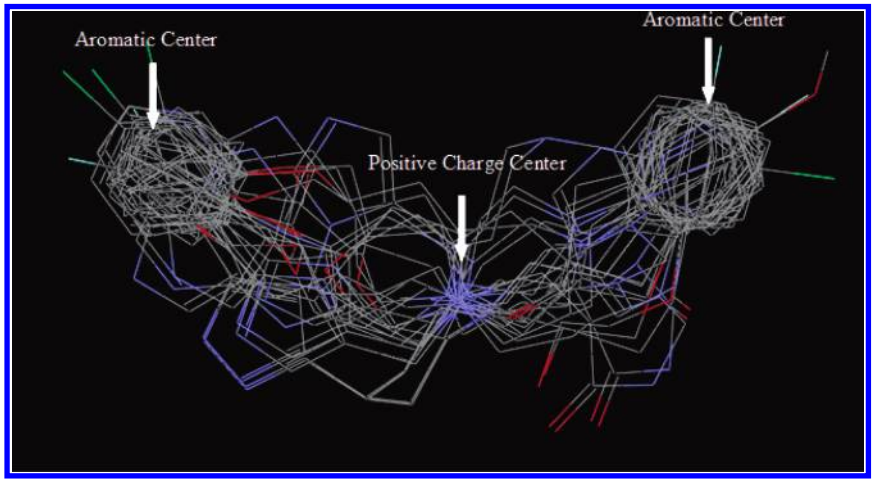

Figure 4. Twenty-one superimposed D4 ligands selected by PharmID. Basic pharmacophore points include two aromatic centers and a positive charge center.

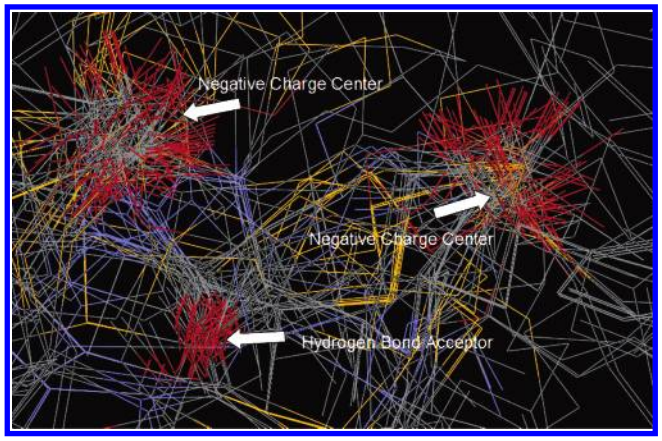

Figure 5. PharmID pharmacophore No.1, with 65 of 78 ACE molecules superimposed. There are three pharmacophore features common to the molecules, and they include two negative charged centers and a hydrogen-bond acceptor. The 65 conformers were selected from the 46 268 conformers; a 0.5 Å RMSD threshold was used for superimposing conformers together.

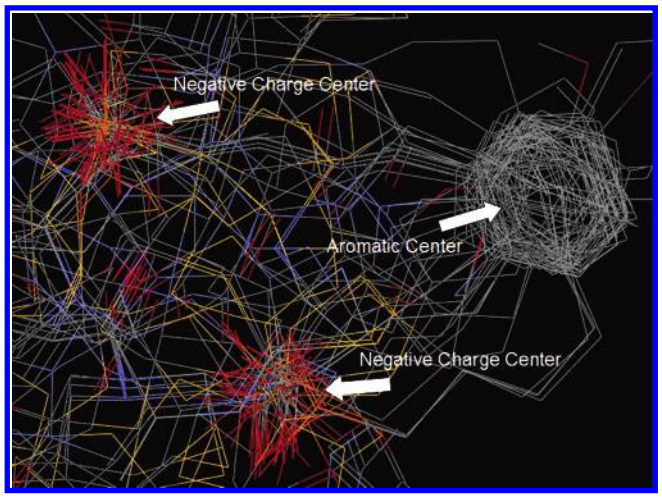

Figure 6. PharmID also determined pharmacophore No.2, with 57 ACE molecules superimposed. Three pharmacophore features are common to these molecules: two negative charged centers (red) and one aromatic center (white/gray).

of the generated conformations, we compared all conformers with their corresponding active conformers. The result is shown in Table 1. For four out of seven ligands, the conformation generation is satisfactory (within a distance of 1.0 Å to the active conformation); for the remaining three compounds, the distance to the binding conformations is within a distance of 2.0 Å. The best hypothesis is shown in Figure 11 (right). Table 2 shows the quantitative comparison between the best hypothesis and the hypothesis generated with only the active conformers. Comparing with the result published by Patel et al.,[4] it is clear that PharmID produced an answer much closer to the answer found by crystallography than any of the commercial methods studied by Patel et al.,[4] Catalyst, DISCO, and GASP, though the pharmacophore features are predefined. Figure 12 shows the IDs for each distance constraint mentioned in Table 2.

DISCUSSION

One way to think about the pharmacophore identification problem is to consider the most potent and rigid molecule and determine if other molecules and their features can be

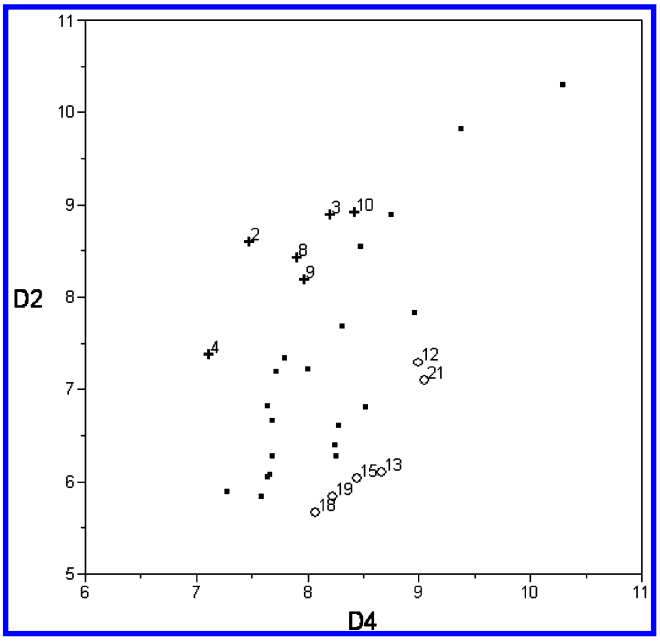

Figure 7. Plot of 2D versus D4 activity. To attempt to find pharmacophores selective for each activity, compounds were selected that deviated from the 45° line. Compounds expected to be more D2-active are marked with a +, and compounds expected to be more D4-active are marked with a ○.

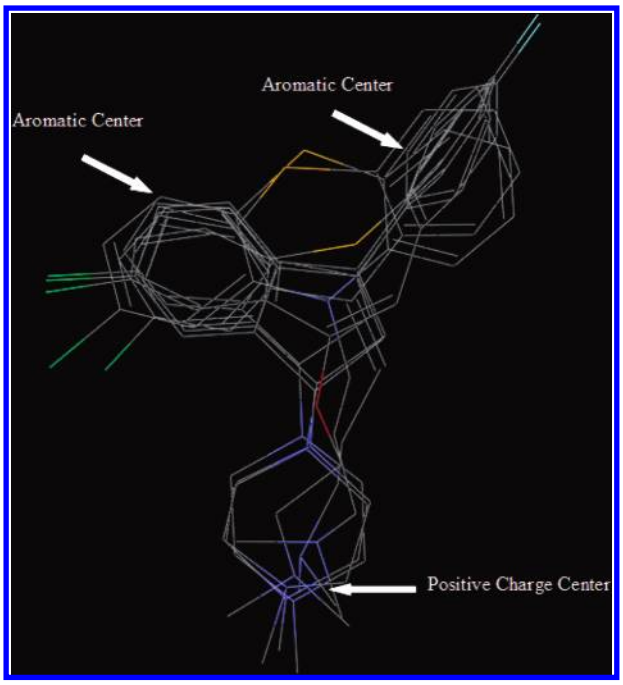

Figure 8. Overlay of compounds cpd08, cpd09, cpd10, cpd02, cpd03, and cpd04 selected to be D2-active. The distance ranges are ACC ~ POC, 5.45 ~ 6.45; ACC ~ POC, 6.69 ~ 8.90; and ACC ~ ACC, 4.94 ~ 5.04.

mapped onto this pharmacophore prototypical molecule. It is a very big intellectual and computational leap to consider all the molecules and allow for multiple binding modes. Even pairwise comparisons among N molecules increase at a rate proportional to N^2. And each molecule will typically have many energetically feasible conformations. The full-scale problem, the comparison of each conformation of each compound to all the conformations of all the other compounds, is essentially computationally hopeless. For example, 500 conformations for each of 100 molecules leads to

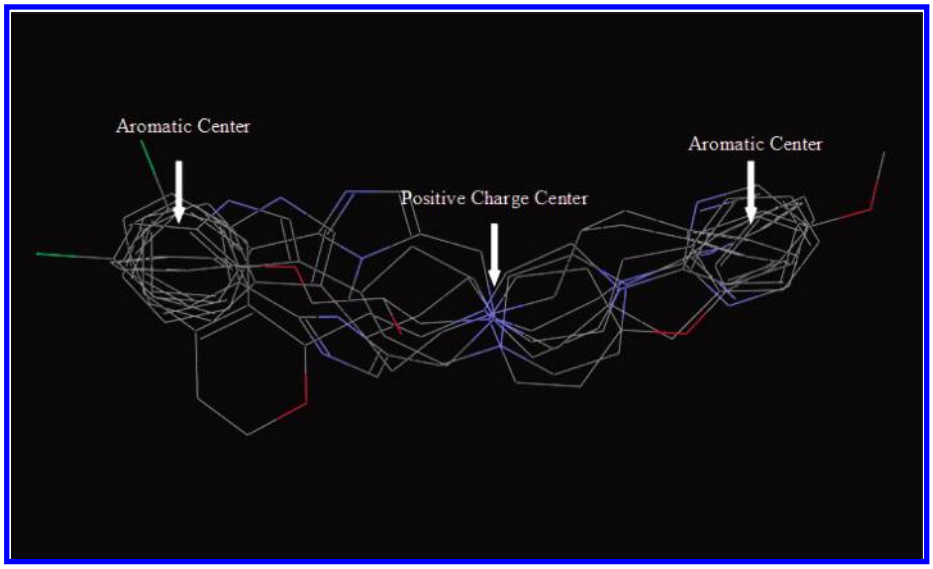

Figure 9. Overlay of compounds cpd12, cpd13, cpd21, cpd15, cpd19, and cpd18. The distance ranges are ACC ~ POC, 5.49 ~ 6.08; ACC ~ POC, 6.29 ~ 7.24; and ACC ~ ACC, 11.78 ~ 12.69.

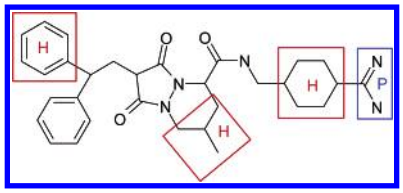

Figure 10. Definition of pharmacophore groups (1c4v, for example).

Table 1. RMSD between the Active Conformation and the Closest OMEGA-Produced Conformation

receptor PDB code	number of conformers	RMSD (Å)
1tom	1000	0.286
1c4v	1000	0.797
1d4p	78	0.553
1d6w	1000	1.846
1d9i	1000	1.606
1dwd	1000	0.624
1fpc	1000	2.009

countless possibilities. The breakthrough is in the creation of a model of the active conformation and its key features, our **Q** vector and **A** matrix. The information about this model is captured in the vector that gives the probability that a feature is important and the matrix that gives the probabilities that a conformation is the binding conformation. Given the model of the possible active conformation and its key features, each conformation for each molecule under consideration can be compared to the model. What would have been a many-to-many comparison now is a many-to-one comparison.

It is somewhat remarkable that the algorithm can start with essentially nothing (neither the active conformation for each molecule nor the set of key features is known) and the process converges to a good answer. In the case of multiple binding modes or multiple targets, the algorithm also needs to sort out which molecules to group together. Statistical theory says the process will converge but is silent on the rate at which the process will converge. The process works on a range of examples, and it is expected to work on data sets where the data is consistent with a solution. A justification on why the process should work goes as follows. A randomly chosen conformation is selected for each active molecule. These conformations are examined for features that are in common. Once the algorithm finds a few features that are in common (and presumed to be important), it is better able to determine which conformations are likely to be the binding conformations. Given better conformations, it is then better able to point to binding features. In practice, in the examples we have studied, the Gibbs sampling process appears to converge rapidly to a solution.

A data set can be trivial. Suppose that the data set contains a very active, rigid molecule. If there is a rigid compound, then our process degenerates to the active analogue method.[14] A data set may allow multiple solutions; that is, the data set is not definitive. Suppose that we are examining close analogues. If all the molecules have essentially the same fragments on opposite sides of a rotatable bond, then the data set is consistent with an infinite number of solutions. Some judgment is necessary in selecting a set of molecules for analysis. Judgment is also necessary in the examination of the solution. Our algorithm is a very rapid computer-aided search for a consistent solution to a very complex 3D search and matching problem. Days or weeks of analysis can be done in minutes to hours.

It is clear that the Gibbs sampling strategy can be used for more complex situations. We mimic the multiple binding mode by applying our program to a data set with two sets of active compounds, D2 and D4, where each set has a different pharmacophore. We also ran several examples by mixing compounds active against two different targets[15] (results not shown). The PharmID program finds elements of the correct pharmacophore when the sets are analyzed separately (it should do this to be an admissible method), and again, it finds the two pharmacophores when the two data sets are analyzed together as one data set. Of course, there is no guarantee that our program will always find multiple binding modes. It appears that we can consider relaxing the usual assumption, "assuming that all the compounds being analyzed bind in the same manner, ..."

One potential weakness in current pharmacophore identification programs is the limitation on the number of conformations that can be used in the search process. Catalyst

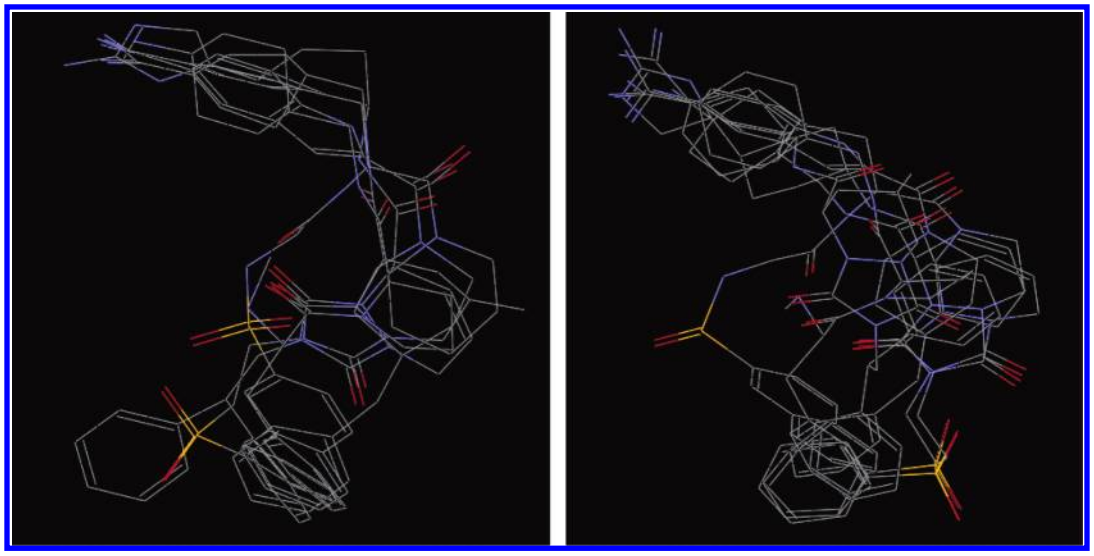

Figure 11. Graphical comparison between the overlaid active conformations (left) and the overlaid PharmID-selected conformations (right).

Table 2. Comparison of the Distance Constraints between the Pharmacophore Hypothesis Generated from Single Active Conformers ("Answer") and that Generated from 6078 Omega-Generated Conformers

ID	crystal structure "answer"	PharmID answer from multiple conformers
No.1	2.79−3.38	2.83−3.34
No.2	10.58−11.16	10.97−12.44
No.3	9.50−10.49	8.62−10.42
No.4	9.21−9.83	9.46−9.88
No.5	7.15−8.02	6.36−7.44
No.6	4.68−5.95	4.86−7.60

The IDs for each distance constraint are shown in Figure 12.

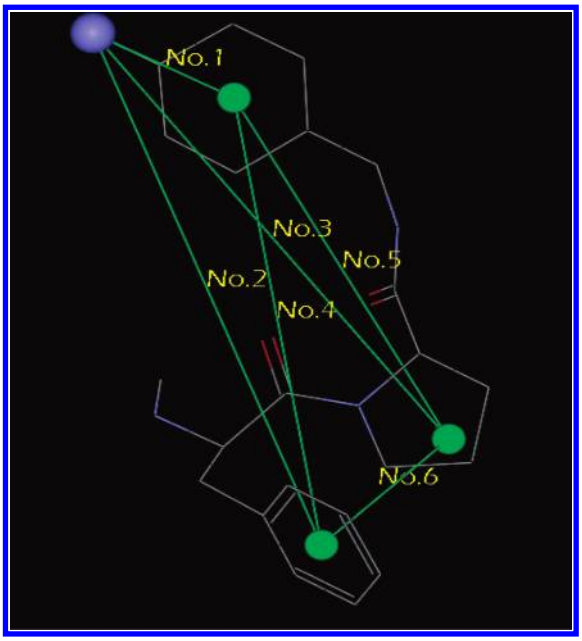

Figure 12. IDs for each distance constraint mentioned in Table 2.

is limited to 255 conformation for each molecule. DISCO appears to be limited to 80 conformations. Our program considers only one conformation for each molecule while it is making a determination of the key features, but once it has tentative key features identified, it can rapidly examine a very large number of conformations for each molecule. It replaces a many-to-many comparison with a small number of many-to-one comparisons. For each of N molecules, compare the "model" active conformation to each of the conformations for each of the compounds under consideration. The calculations are order n rather than $\binom{n}{2}$ (where $n = \sum_{i=1}^{N} M_i$, the total number of conformations). As usual, there is no free lunch; the process is not checking all possibilities, so there is no guarantee that the process will find the optimal solution. And again, behind this process is the assumption that there is a single binding mode (or just a few) and a consistent set of key features for the binding mode (or key features for each of the few binding modes).

Note that PharmID does not formally consider entropy when examining the conformations of a compound. In our examples, we set the upper bound at 15 kcal/mol. A clustering method was not used to select representative conformations of a molecule. As PharmID can use a large number of conformations for each molecule, it is not computationally necessary to be very restrictive. It is clear from example 4, thrombin, that 1000 or more conformations per molecule can be necessary for obtaining a conformation close to the binding conformation. As PharmID obtains literature and crystal structure results for the examples presented, it is possible that being able to examine a large number of conformations makes the careful selection of conformations less important than for other algorithms.

More benchmarking should to be done to understand the range of data sets that can be successfully analyzed with our algorithm. The computational speed is good, so larger numbers of compounds and larger numbers of conformations can be used. For the time being, some of the data set guidelines used for other pharmacophore identification programs can be considered as a guide. There should be multiple, diverse compounds that appear to follow only one or two modes. The compounds and their conformations should be structurally diverse. Having compounds that are fairly rigid should help. The molecular features should capture important information about the activity of the compounds.

For PharmID, the postprocessing looks for and caters to multiple binding modes. An alternative strategy would be, after the execution of PharmID, to use the identified key

features to cluster the compounds. If there are distinct clusters, then the PharmID process can be computed on compounds from each cluster on the reasonable assumption that these compounds from the different clusters might be binding in different ways. Note that, once the key features are identified from the first run of PharmID, the clustering is more likely to correctly identify compounds binding in different ways; clustering algorithms can fail to give good clusters if a large number of unimportant features are included with the important features.

PharmID offers a number of advantages for pharmacophore identification. The key to the speed and size of data sets that can be considered is the fact that searches are many-to-one. In effect, the **Q** vector contains the information on the key features and the features of each selected conformation are compared to it. The **A** matrix holds the information on the likelihood that a conformation is the binding conformation; each conformation within a molecule is compared to **A**. PharmID is never in the situation of trying to compare each conformation of one molecule against all the conformations of another. Relatively large data sets can be used. For the 78 ACE compounds, there were 46 268 conformations; the problem was run in 34 h on a desktop Windows machine (Pentium IV Xeon, 2.4 GHz). A small data set of six compounds and 6000 conformations was finished in 36 min. Our initial efforts have concentrated on getting a working program; future efforts will be directed into optimization of the code and benchmarking on larger and more varied data sets.

Data Sets. The data sets used in this paper are available from the corresponding author: synthetic, D2, ACE, D2/D4, and thrombin. The synthetic data is given as compound, conformation number, and the feature bit strings. The compounds are given as Smiles strings.

REFERENCES AND NOTES

(1) Martin, Y. C. Pharmacophore Mapping. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington D. C, 1998; Chapter 6.
(2) Güner, O. F. *Pharmacophore, Perception, Development, and Use in Drug Design*; International University Line: La Jolla, CA, 1999.
(3) Kurogi, Y.; Güner, O. F. *Curr. Med. Chem.* **2001**, *9*, 1035−1055.
(4) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653−681.
(5) Güner, O.; Clement, O.; Kurogi, Y. *Curr. Med. Chem.* **2004**, *11*, 2991−3005.
(6) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003.
(7) Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F.; Wooton, J. C. *Science* **1993**, *262*, 208−214.
(8) Rouchka, E. C. A Brief Overview of Gibbs Sampling. http://sapiens.wustl.edu/∼ecr/PAPERS/gibbs.pdf (accessed Sep 2005).
(9) Bron, C.; Kerbosch, J. *Commun. ACM* **1973**, *16*, 575−577.
(10) Böstrom, J.; Bohm, M.; Gündertofte, K.; Klebe, G. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1020−1027.
(11) OpenEye Scientific Software: Omega. http://www.eyesopen.com/products/applications/omega.html (accessed Sep 2005).
(12) Van Drie, J. H. *Pharmacophore, Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line: La Jolla, CA, 1999; Chapter 27.
(13) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. *J. Am. Chem. Soc.* **1993**, *115*, 5312−5384
(14) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. The conformational parameter in drug design. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, D. C., 1979; Vol. 122, pp 205−226.
(15) Tripos Inc., 2004.

CI050427V