Recursive Distance Partitioning Algorithm for Common Pharmacophore Identification

Fangqiang Zhu* and Dimitris K. Agrafiotis

Johnson & Johnson Pharmaceutical Research & Development, L.L.C. 665 Stockton Drive,
Exton, Pennsylvania 19341

An improved method for exhaustively identifying common pharmacophores from a given list of 3D conformers is proposed. The method partitions feature lists into multidimensional boxes according to the distances between the pharmacophore centers. Unlike some existing techniques, each feature list is mapped into multiple boxes to ensure that good matches will never be missed due to the partitioning. To circumvent the computational complexity of the problem, a recursive distance partitioning (RDP) algorithm is introduced, in which the partitioning and the elimination of unqualified feature lists are carried out at multiple levels. The method is demonstrated to be both accurate and efficient.

INTRODUCTION

Identifying or designing active ligands against a specific biological target is one of the most important tasks in drug discovery. While structure-based techniques have greatly advanced our ability to design bioactive molecules, the approach is only applicable to a limited number of targets whose 3D structures have been experimentally determined through X-ray crystallography or NMR. Many important target families, such as GPCRs, still remain challenging to structural means, and in such cases alternative rational methods can be employed in order to increase the likelihood of success. Pharmacophore modeling is one such method.[1] A pharmacophore is the spatial arrangement of steric and electronic features that are necessary to confer the optimal interaction with a particular biomolecular target and to induce or inhibit its biological response. In pharmacophore modeling, a set of high-affinity ligands is analyzed, and common pharmacophore (CP) hypotheses on the relative 3D arrangement of key molecular recognition sites are inferred. These recognition sites, which are often referred to as pharmacophore centers, are generalized interaction points that are usually defined in the form of chemical patterns or substructures, mapped onto the target molecule. The resulting pharmacophore hypotheses can then be used to search 3D databases to identify ligands with similar 3D features, which may have a greater-than-average probability of being active against the target, and could therefore be subject to further examination. When combined with quantitative activity data, the pharmacophore hypotheses can also be used to build quantitative structure−activity relationship (QSAR) models to predict the activity of any given compound.

A variety of computer programs and algorithms for pharmacophore modeling have been reported in the literature.[1] For example, DISCO is one of the earlier software products that enable automated pharmacophore matching.[2] The Catalyst/HipHop program constructs CPs by gradually extending initially small sets of features.[3] The software package Chem-X employs special bit strings called pharmacophore keys to address pharmacophore diversity and facilitate database searching.[4] A genetic algorithm has been adopted in the GASP program for the alignment of flexible molecules.[5] Gibbs sampling has also found application in the CP identification problem.[6] Phase, a versatile and flexible software package for pharmacophore modeling, is a recent development in this area.[7,8] Unlike previous methods, Phase identifies CPs by enumerating and partitioning all *k*-point pharmacophores from the active ligands, thereby achieving an improved exhaustiveness. Phase also offers the functionality of 3D QSAR and database searching.

Identifying CPs from active ligands is a key step in pharmacophore modeling. A CP represents a common 3D configuration of the chemical features shared by the active ligands and corresponds to a hypothetical pattern of interaction between the ligands and the target. In this article, we propose a new recursive distance partitioning (RDP) algorithm for efficiently finding CPs. Unlike some existing methods, our algorithm does not suffer the drawback of possibly missing good CPs and therefore improves the accuracy and reliability of identifying the best hypotheses from a given data set.

METHOD

In this section, we describe the key steps in our method for identifying common pharmacophores (CPs). First, we provide a brief description of the preprocessing steps, including conformation generation and pharmacophore site assignment. We then formally define CP and the problem of CP identification. Following a discussion of distance partitioning techniques, we introduce the RDP algorithm. Next, the scoring scheme for CPs and the exact method for finding the highest-scored CP are discussed. Finally, we describe in detail the implementation of our method.

1. Preprocessing Steps. Given a set of active ligands, the first step for pharmacophore modeling is the generation of their 3D conformations. Conformational sampling of organic molecules has been the subject of many studies and has been reported extensively in the literature, such as our recent

* Corresponding author phone: (610)458-5264 x6620; fax: (610)458-8249; e-mail: fzhu2@prdus.jnj.com.

work[9−12] and the references therein. In this article, we will not cover this topic and will instead assume that the 3D conformers for each active ligand have already been generated and given as input.

Pharmacophore sites (or features) for each conformer can be created according to certain feature definition rules. Typical feature types include H-bond acceptors (A), H-bond donors (D), negatively charged groups (N), positively charged groups (P), and aromatic centers (R). Additional features, such as hydrophobic groups and lone-pairs, may also be defined in a similar way. Each pharmocophore site is associated with one or more atoms in the ligand, and its 3D coordinates are defined as the centroid of these atoms. For the sake of simplicity, in this work the directionality of the features is ignored, and therefore each site is represented by a point in 3D space. The strategies for dealing with directionality will be discussed at the end of this article.

2. Problem Definition. First, we define a *feature list* (FL) as a list of individual pharmacophore sites in a conformer. FL is sometimes referred to as a *k*-point pharmacophore in the literature.[8] FLs can be classified into different *variants*, a term used in Phase[8] to describe the number of sites for each feature type. For example, a variant "AANRRR" specifies that the FL consists of two H-bond acceptors, one negatively charged group, and three aromatic centers.

Assuming there are *m* ligands, a *common pharmacophore* (CP) is composed by *m* FLs of the same variant (one from each ligand). Note that the order of the sites in the FLs is important since it determines the site correspondence between the FLs—the *i*th sites in each FL correspond to each other and are represented by the *i*th site in the CP. Obviously, the corresponding sites must share the same feature type. Therefore, one can demand that the sites in each FL be sorted according to their feature type. Furthermore, a shuffle of the sites with a same feature type in an FL will alter its site correspondence to other FLs and will give rise to a different CP.

In addition to the component FLs, each CP also has its own pharmacophore sites, which represent the common positions for the corresponding sites in the FLs. The CP sites may simply adopt the sites in one of the FLs (which is called the reference FL) or may be assigned different coordinates than any of the FLs, as will be discussed later. The quality of a CP can be characterized by a score. The exact definition of the score is deferred to a later section, but intuitively, in a highly scored CP, the component FLs must align closely with each other and with the CP sites. The problem of CP identification is then to find CPs with the best scores from the given conformers.

Since a CP of a given variant must be composed by FLs of the same variant, in practice different variants are usually processed separately. When the total number of sites (N_{site}) in the CPs is specified by the user, all possible variants can be enumerated based on the numbers of sites the ligands have for each feature type, and CPs of each variant will then be identified independently.

3. Distance Partitioning. In order to exhaustively identify good common pharmacophores (CPs) of a given variant, all feature lists (FLs) of that variant must be enumerated. Each ligand usually gives rise to a large number of FLs, due to the use of multiple conformers and the many different permutations of pharmacophore site selections (for example,

given a variant "AAADDR", each acceptor site in the conformer can be at the first, second, or third position in the FL). Every combination of these FLs, one from each ligand, corresponds to a unique CP. Consequently, the total number of the CPs is huge, and exhaustively enumerating and scoring all of them is computationally prohibitive.

To reduce the required amount of computation, FLs with similar geometries, which are more likely to align closely with each other and contribute good CPs, can be grouped together. The geometry of an FL can be represented by its intersite distances, i.e., the distances between pairs of pharmacophore sites in the FL. For an FL with N_{site} sites, there are a total of $N_{pair} = N_{site}(N_{site}-1)/2$ feature pairs or intersite distances. If the entire distance range is divided into a number of bins, an intersite distance can be mapped into one of the bins. With N_{pair} intersite distances, each FL can then be mapped into an N_{pair}-dimensional box, with each dimension representing the distance of one feature pair. The length of the box in each dimension is equal to the width of the bin. When all of the FLs are partitioned in this way, the ones found in the same box should share a similar 3D geometry. Since the number of FLs in the box would be much smaller, it becomes feasible to apply exact methods to find the best CP within the box.

However, even if two intersite distances are very close to each other, it is still possible that they are partitioned into separate, adjacent bins, if they happen to be on either side of a bin boundary. In such cases, the two corresponding FLs will be found in different boxes and will not be incorporated in the same CP. Consequently, good CPs may be missed due to the partitioning. To address this problem, when retrieving a CP from a box, one needs to examine the FLs in its neighboring boxes as well. An approximate solution would be to only consider the FLs in the direct, one-dimensional neighbors of the box. However, although this strategy reduces the possibility of missing good matches, it does not entirely eliminate it. If two FLs have two or more pairs of corresponding intersite distances falling into different bins, they will be mapped into two boxes that are not direct neighbors of each other and therefore cannot appear together in any CP, even if they actually align perfectly well with each other. To be completely exhaustive, one needs to examine the FLs in all the N_{pair}-dimensional neighbors of a given box as well as the box itself, which amounts to a total of $3^{N_{pair}}$ boxes.

An alternative exhaustive strategy is to map each FL into multiple boxes. Specifically, each intersite distance can be mapped into the two nearest bins. For example, a distance of 2.3 would be mapped into bins [1,2) and [2,3), and a distance of 4.9 to bins [4,5) and [5,6). A similar strategy was previously adopted in the OSPPREYS method, where the bits for the near-neighbor bins were set in the pharmacophore fingerprint.[13] With N_{pair} intersite distances, an FL will be mapped into a total of $2^{N_{pair}}$ boxes. When retrieving a CP from a box, one no longer needs to consider the FLs in its neighbors, since those FLs are already included in the box. Consider a pair of distances under this scheme. If their difference is smaller than the bin width L, the two can always be found in at least one of the bins; if the difference is larger than $2L$, none of the bins contains both of them; if their difference lies between L and $2L$, the two may or may not be found in the same bin, depending on the relative positions

ALGORITHM FOR COMMON PHARMACOPHORE IDENTIFICATION

J. Chem. Inf. Model., Vol. 47, No. 4, 2007 **1621**

of the bin boundaries. Consequently, for a group of FLs, if no difference between any corresponding intersite distances exceeds L, they are guaranteed to appear together in at least one box. Therefore, with a properly chosen bin width, good CPs will not be missed due to the partitioning.

A naïve implementation of the distance partitioning would be to identify the box that each FL belongs to by checking all of its intersite distances, insert the FL into the box, and finally examine all of the boxes after the partitioning. However, this approach is very inefficient for the exhaustive strategies, because one would need to explicitly process a huge number of boxes either during or after the partitioning. For example, for 7-point pharmacophores (N_{site}=7), if each intersite distance is mapped into two bins as described above, every FL needs to be mapped into $2^{N_{pair}} \approx 10^6$ boxes. Such an amount of computation is prohibitive in practice. In order to make the exhaustive strategy affordable, a new algorithm is introduced in the following section.

4. Recursive Distance Partitioning (RDP) Algorithm. As discussed above, if each intersite distance is mapped into two bins, the feature lists (FLs) would be mapped into a very large number of boxes. However, a box "survives" only if it contains at least one FL from each molecule. Only the surviving boxes could yield common pharmacophores (CPs) and need to be actually processed. Furthermore, if a group of boxes is known not to contain any FL from at least one of the molecules, none of these boxes would survive. Accordingly, our RDP algorithm adopts a top-down approach to eliminate unqualified FLs as early as possible, as described in Chart 1.

The second parameter (i) of the function in Chart 1 specifies the level of the partitioning or the dimension being partitioned. Initially, the function should be invoked as *partition(FLlist, i=1)*, where *FLlist* is the list of all enumerated FLs of the given variant. In the first level (i=1), the FLs are mapped into the distance bins according to the first intersite distance. Then the distance bins are examined one by one. If a bin does not contain at least one FL from each molecule, it can be simply ignored. If the bin does contain FLs from all of the molecules, the FLs in this bin should be further partitioned according to the next intersite distance, by making a recursive call of *partition()* with an increment of the level i (step 4). If this recursion already reaches the last level (i=N_{pair}), an N_{pair}-dimensional box that contains those FLs has been identified, and exact methods can be applied to retrieve a CP from them (step 5), as described in the next section. The recursive partitioning is illustrated in Figure 1, where each intersite distance is assigned to only one bin for the sake of simplicity. However, the procedure remains essentially the same when mapping each distance into two bins. Although in principle each FL is still mapped into $2^{N_{pair}}$ boxes according to this scheme, only the surviving boxes are examined, while others are effectively eliminated at earlier levels without even being fully identified. This significantly reduces the required computation for the partitioning, making it feasible to apply the exhaustive strategy for CP identification.

A related tree-based partitioning algorithm is adopted in Phase, in which the FLs are filtered through a binary decision tree.[8] The algorithm achieves good efficiency by attempting to eliminate disqualified subtrees at each level. Since our RDP procedure can also be perceived as a multibranched

Chart 1

Input: *FLlist – list of FLs to be partitioned;*

    i – indicates the feature pair (or the dimension) to be examined in this recursion.

partition(FLlist, i)

{

    1. Associate an empty list with each of the bins.

    For each *FL in FLlist* **do**

    {

        2. Check the i-th intersite distance in FL, identify the bin(s) it should be mapped

            into, and insert FL into the associated list(s) for the bin(s).

    }

    For each *bin* **do**

    {

        3. Examine the list FLlist associated with this bin*

        if *(each molecule is represented by at least one FL in FLlist*) then*

        {

            if *(i < N_{pair})* **then**

            {

                4. Call partition(FLlist, i+1) to further partition the FLs in this*

                   bin according to the next intersite distance.

            }

            else

            {

                5. A surviving box has been found. Exhaustively examine the FLs

                   in FLlist to retrieve a CP.*

            }

        }

    }

}

tree (see Figure 1), the two algorithms bear certain resemblance, but they also exhibit distinct differences. For example, in a binary tree, several levels of partitioning are needed to assign an intersite distance to its terminal node, which requires more computation than directly mapping the distance into the corresponding bins, as in our algorithm. Also, the binary tree works most conveniently when the number of the terminal nodes is a power of 2, whereas our binning method works equally well for any number of bins. More importantly, the tree-based technique needs to pool the neighbors of a surviving box when retrieving CPs.[8] Probably because the N_{pair}-dimensional neighbors of the box are too many to enumerate, Phase only considers the direct neighbors of the box, and thus there is always the possibility of missing good matches, as discussed in the previous section. In contrast, by mapping each distance into two bins, our method is both exhaustive and efficient. Furthermore, during the essentially depth-first search in our algorithm, no more than one bin is expanded at each level at any point in time, which helps to keep memory requirements under control.

5. Retrieving Common Pharmacophores. In a surviving box, each molecule is represented by one or more feature lists (FLs). Since a common pharmacophore (CP) must consist of exactly one FL from each molecule, one needs to
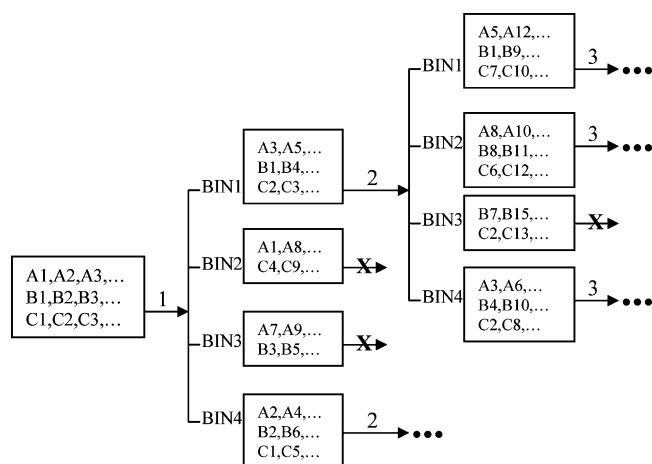
Figure 1. An illustration of the recursive distance partitioning (RDP) algorithm. Feature lists (FLs) from the three molecules (A, B, C) are labeled by their molecule name and index, such as A1, B3, etc. In the first level of the partitioning, the FLs are placed into the 4 distance bins according to their first intersite distance. A bin will not be further processed if it does not contain any FL from one of the molecules. For example, in the first level, bin 2 is ignored because it does not contain FL from molecule B. On the other hand, if a bin (such as bin 1 and bin 4 in the first level) contains FLs from all of the molecules, then these FLs should be further partitioned at the next level and placed into the distance bins according to the next intersite distance. In the diagram, the numbers above the arrows indicate the level of the partitioning, or the intersite distance to be examined at this level.

select the best combination of FLs to construct the CP for that particular box. This can be done through the method adopted in Phase,[8] in which each of the FLs in the box is temporarily treated as a reference. Given a reference FL from a given molecule, the FLs from each of the other molecules are superimposed against it. Among all FLs from a particular molecule, the one with the best overlap (measured by RMSD) to the reference is selected as its partner. The average RMSD over the partners across all molecules is taken as the score for that reference. Each FL in the box is treated as a reference, and the one with the best score is identified. The CP for this box thus contains the best reference FL and its partner FLs from the other molecules. The 3D superimposition mentioned above also serves to eliminate FLs with inconsistent chiralities, since if two FLs are the mirror images of each other, they share the same intersite distances and would be mapped into the same box.

In Phase,[8] the pharmacophore sites of the CP simply assume the sites of the reference FL. Here we propose an iterative procedure to further refine the CP sites. After the component FLs are aligned against the reference, a template pharmacophore is constructed in which the 3D coordinates of each site take the average position of the corresponding sites in the FLs. Then the component FLs are aligned against the template, and the average of their new coordinates is used to update the position of the template. This procedure is repeated until the position of the template converges, which is then taken as the coordinates of the CP sites. Our experiments showed that a couple of iterations are usually sufficient for the template to achieve convergence within 0.01 Å. The CP sites obtained in this way lie at the centers of the aligned FLs and therefore better represent the common geometry of the component FLs and are insensitive to the specific choice of the reference FL among them.

6. Implementation. The recursive distance partitioning (RDP) algorithm was implemented in the C++ programming language. After reading in the conformers, the program creates the pharmacophore sites according to predefined feature definition rules. The user may specify the desired total number of sites for the common pharmacophores (defaulted to $N_{site}=5$) and the minimum and maximum numbers for each feature type (defaulted to 0 and the largest possible number, respectively). Based on these numbers, the program determines all valid variants and then processes each of them one by one.

To find common pharmacophores (CPs) for a given variant, the first step is to enumerate all possible feature lists (FLs) from each conformer. Specifically, for each feature type, given the available sites in the conformer and the required number of sites specified by the variant, all possible selections of the sites are enumerated. As mentioned before, the order of the sites in an FL is important since it determines their corresponding sites in other FLs. To ensure the uniqueness of the CPs, if a conformer is from the first molecule, the sites in its FLs are ordered and not permuted. For conformers from all other molecules, all permutations of the sites with the given feature type are enumerated, each resulting in a unique sublist of sites for that feature type. The enumerated site sublists for different feature types are combinatorially joined, with every combination giving rise to a unique FL. In addition, an FL is rejected if the distance between any pair of the sites is below a user-defined minimum distance, defaulted to 2 Å as in Phase,[7] in order to avoid including the same chemical group more than once (e.g., the user might not want to have an OH group associated with both a donor site and an acceptor site in the same FL). Each FL can be directly accessed by a unique index assigned to it.

All valid FLs are then passed to the recursive distance partitioning routine. The entire range of the intersite distances is divided into bins with a width of 1 Å. At each level of the recursion, each FL is placed into two distance bins according to the corresponding intersite distance, as described in sections 3 and 4. Whenever a group of FLs are found to survive all the N_{pair} levels of partitioning and therefore reside in a same N_{pair}-dimensional box, they are immediately processed according to the methods described in section 5, before the partitioning proceeds to search for other surviving boxes.

Retrieving a CP from a surviving box involves the computation of pairwise RMSDs between the FLs in the box. In our scheme, some FLs may be found in multiple surviving boxes. Accordingly, after the RMSD between two FLs is computed, we store the value in a hash table, searchable by FL index pair. Whenever an RMSD is needed, it is first looked up in this table and is directly retrieved if it exists, thus avoiding repeated computations. As in Phase, our program uses a cutoff (defaulted to 1.2 Å) to reject a CP if the RMSD between the reference and any of the component FLs exceeds this value.[8]

The retrieved CPs are stored in a queue sorted by the alignment score (i.e., the average of the RMSD values). When a CP is retrieved from a surviving box, it is first checked for uniqueness, as the FLs were mapped into multiple boxes and an identical CP may have already been identified from other boxes. If it is indeed unique, then the
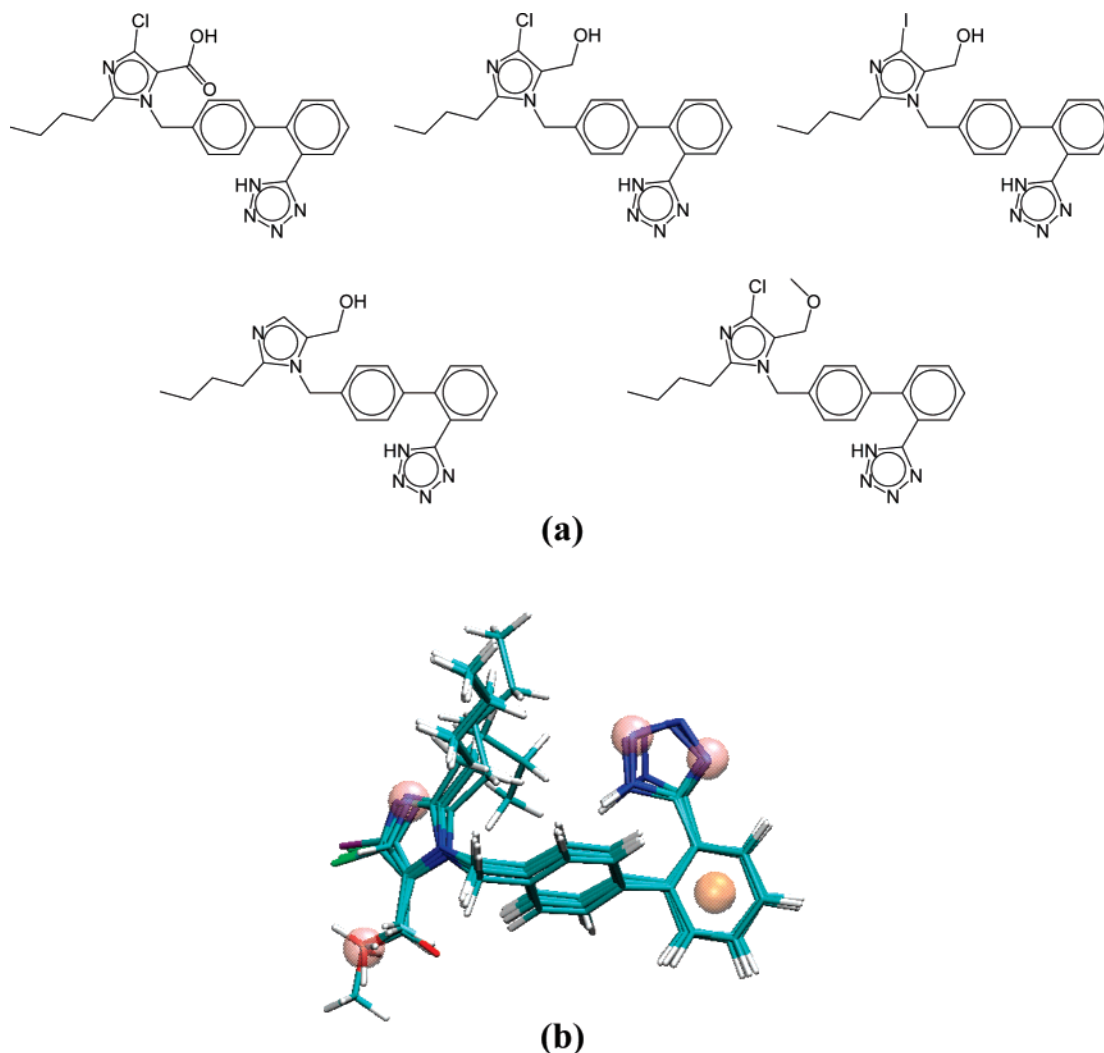
Figure 2. (a) The chemical structures of the five molecules in the test suite. (b) The common pharmacophore with the best score for variant AAAAR, along with the associated conformers from each molecule aligned to it. The pink and orange spheres represent the pharmacophore sites for acceptor (A) and aromatic center (R), respectively. The image was rendered using VMD.[16]

new CP is inserted into the queue. The user may specify the maximum number of CPs to be reported, which determines the capacity of the queue. Whenever an insertion causes an overflow, the CP with the worst score is removed from the queue. When the RDP routine completes, the CPs stored in the queue are exported as the results for that particular variant.

RESULTS AND DISCUSSION

In order to validate our method, we compared the results from our program and Phase,[8] which uses a tree-based partitioning technique described earlier. Our test suite contains 5 molecules taken from a previous study[14] (Figure 2a), which are also used in a tutorial from the Phase User's Guide.[7] The torsion search method in MacroModel[15] was employed to generate a total of 196 conformers for these molecules, which were taken as the input for both programs. Due to some differences in the feature definition rules used in our program and Phase, in this validation study we only focus on two feature types, namely, H-bond acceptor (A) and aromatic center (R). Accordingly, the maximum feature frequencies for all other types (D, H, N, P) were manually set to zero in Phase,[7] so that they would not appear in the

variants and common pharmacophores (CPs). Since the individual variants are processed separately in both Phase and our program, such treatment should not affect the fairness of the comparison. For these particular molecules, the A and R pharmacophore sites were identically assigned in the two programs, allowing the two methods to be compared on the same footing.

A valid comparison must also be based on the same scoring function. In Phase, the final score of a CP is given by the weighted sum of several terms (such as the site score and the vector score) reflecting various aspects of the match.[8] In our program, however, the score is only based on the RMSD values, which correspond to the Phase site score. Accordingly, the weighting factors for all other terms were manually set to zero in Phase,[7] leaving the site score as the only constituent term of the final score. In addition, the vector score threshold was set to the minimum (-1) in Phase,[7] so that the program was not allowed to discard any CPs due to poor vector scores. The Phase site score s_i for each nonreference feature list (FL) is determined by its RMSD with respect to the reference FL: $s_i = 1 - \text{RMSD}_i/\text{cutoff}$, where cutoff (defaulted to 1.2 Å) is the maximally allowed RMSD value,[8] as described in the previous section. The

Table 1. Scores (Defined in the Text) of the Best 5-Point Common Pharmacophores for Each of the Four Variants[a]

	AAAAR	AAARR	AARRR	ARRRR
phase	0.740	0.858	0.836	0.837
RDP	0.803	0.887	0.851	0.838

[a] Identified by phase and by our recursive distance partitioning (RDP) algorithm for the test suite.

overall site score then takes the average of s_i over the nonreference FLs in the CP. For the purpose of this validation, we temporarily modified our program to adopt the above definition for the site score and turned off the iterative refinement introduced earlier. With these measures, the two programs have an identical scoring function and thus can be directly compared. For the sake of simplicity, both programs were requested to return only the best-scored CP for each variant (in Phase,[7] this was done by setting both the upper and lower limits for the number of hypotheses to 1). All other parameters assumed their default values in both programs.

When running the test suite to search for 5-point CPs (N_{site}=5), both Phase and our program found valid CPs in 4 variants, namely, AAAAR, AAARR, AARRR, and ARRRR. The scores of the best CPs identified by Phase and by our program are given in Table 1 for each of the variant, and the best CP from our program for variant AAAAR is illustrated in Figure 2b, as an example. Table 1 indicates that for all of the 4 variants, our RDP algorithm obtained better or similar scores compared to Phase. We note that the sole purpose of this comparison is to validate the algorithmic rigidity of our method, instead of addressing the biological implications of the specific results. In this particular case, the best CPs identified by the two programs actually share similar patterns, although the conformers picked by our program seem to align slightly better. However, it is more critical that good results, if they exist, should always be found by the program. In our method, as discussed before, a CP can never be missed as long as the distance deviations between its component FLs do not exceed the bin width. Therefore, the method eliminates the possibility of missing good results due to the artificial partitioning. Indeed, when trying a broader search by using a larger bin width of 2 Å (instead of the default value of 1 Å), we obtained the same best CPs, which suggests that they might actually be the best answer among all possibilities. The ability to identify the best CPs in this case demonstrates the exhaustiveness of our method, which should result in an improved reliability and may help identify multiple binding modes when applied to other cases as well.

In Phase, the CP identification procedure consists of the "Find Common Pharmacophores" and "Score Hypotheses" steps.[7] For this test suite, the total time taken for the two steps was found to be ~160 s on a Linux workstation equipped with a 3.4 GHz Intel Xeon processor. In contrast, it took our program merely ~12 s to complete the job on the same computer. We note that these numbers should not be directly compared due to the following reasons. Although the weighting factors for vector score and other terms were set to zero in Phase, it is unclear whether these properties were still calculated during the run, thus requiring additional

computing time. Moreover, Phase was run from the graphical user interface, which would consume more time than running in command-line mode. Nevertheless, the performance of the RDP algorithm appears to be quite satisfactory in this test, suggesting that the exhaustive method proposed here can be affordably applied in practice.

The workflow in our method is similar to that in Phase.[8] After the user sets the total number of sites and the ranges for each feature type, the program lists all possible variants that satisfy the requirement. The user may then specify which variants are of interest and worth pursuing, and the program searches for CPs for each of these variants separately. The basic task in both methods is to identify CPs of an individual variant. This is different from some earlier algorithms, where CPs are constructed by gradually extending the feature set. In Phase and our method, all features in a given variant must be matched, and partial matches are not directly supported.[8] In practice, one may sometimes want to match only a subset of a group of optional features. In such cases, all variants satisfying that requirement can still be enumerated and processed separately, although they may differ in length. We note that if good CPs are found for a feature set, they will match any subset of the features too, and it is probably not necessary to examine the subsets in that case. Therefore, one may process the variants of larger feature sets first and only examine their subsets if they do not yield satisfactory CPs. Since the partitioning usually terminates very quickly for unsuccessful variants, such an approach should be efficient in practice. Furthermore, following the philosophy adopted in Phase,[8] pharmacophore modeling is by nature an empirical method that should be combined with the knowledge of the user. Very often the user has some idea of the desired model and may be interested in only a fraction of the possible variants. With improved performance, our algorithm is able to quickly and reliably find CPs for any given variant, allowing the user to explore different hypotheses in an interactive manner.

In our current implementation, a relatively simplistic representation of the pharmacophore sites is adopted. More sophisticated pharmacophore models would incorporate the directionality of the sites. In addition, other measures such as volume overlap between the conformers could also be included to evaluate CPs. In Phase,[8] the directionality and those measures are ignored during the partitioning and are only taken into consideration when scoring the CPs. Since the partitioning is still based on the intersite distances as described in this study, our RDP technique can be similarly applied in such cases when combined with new scoring functions. In principle, the directionality could also be represented in the partitioning phase if each pharmacophore site is split into two points on the line defining its direction. Then the relative directions between different sites can be reflected by the distances between these points. This proposal might be worth pursuing in a future study. In Phase, the requirement for a CP can be relaxed so that its component FLs only need to represent at least a certain number of molecules, and not necessarily all of them.[8] Our algorithm can also be easily adapted to incorporate this feature. Therefore, adopting this algorithm should not limit the existing functionality in the software. In conclusion, the RDP algorithm presented in this study is shown to achieve both

ALGORITHM FOR COMMON PHARMACOPHORE IDENTIFICATION

J. Chem. Inf. Model., Vol. 47, No. 4, 2007 **1625**

exhaustiveness and efficiency and should prove valuable for the task of pharmacophore model building.

REFERENCES AND NOTES

(1) Guner, O. F. *Pharmacophore perception, development, and use in drug design*; International University Line: La Jolla, CA, 2000.

(2) Martin Y. C. Distance comparisons (DISCO): A new strategy for examining 3D structure-activity relationship. In *Classical and 3D QSAR in agrochemistry*; Hansch C., Fujita T., Eds.; American Chemical Society: Washington, DC, 1995; pp 318−329.

(3) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(4) Murrall, N. W.; Davies, E. K. Conformational freedom in 3-D databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.

(5) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(6) Feng, J.; Sanil, A.; Young, S. S. PharmID: pharmacophore identification using Gibbs sampling. *J. Chem. Inf. Model.* **2006**, *46*, 1352−1359.

(7) *Phase, version 2.0*; Schrödinger, LLC: New York, 2005.

(8) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D, E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided. Mol. Des.* **2006**, *20*, 647−671.

(9) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1186−1191.

(10) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J. Comput. Chem.* **2006**, *27*, 1962−1969.

(11) Zhu, F.; Agrafiotis, D. K. Self-organizing superimposition algorithm for conformational sampling. *J. Comput. Chem.* **2007**, *28*, 1234−1239.

(12) Agrafiotis, D. K.; Gibbs, A.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, in press.

(13) Martin, E. J.; Hoeffel, T. J. Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS): A substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation. *J. Mol. Graphics Modell.* **2000**, *18*, 383−403.

(14) Krovat, E. M.; Langer, T. Non-peptide angiotensin II receptor antagonists: chemical feature based pharmacophore identification. *J. Med. Chem.* **2003**, *46*, 716−726.

(15) *MacroModel 9.1*; Schrodinger, LLC: New York, 2006.

(16) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics Modell.* **1996**, *14*, 33−38.