Parameterization and Conformational Sampling Effects in Pharmacophore Multiplet Searching

Peter C. Fox, Philippa R. N. Wolohan, Edmond Abrahamian,[†] and Robert D. Clark*

Tripos International, 1699 South Hanley Road, St. Louis, Missouri 63144

Received July 13, 2008

Pharmacophore patterns in ligands can be effectively characterized in terms of their constituent pharmacophore multiplets. Bitsets (fingerprints) encoding which particular multiplets are found in a given ligand have been and continue to be used as molecular descriptors in a range of molecular modeling applications, from ligand alignment and diversity analysis to pharmacophore-based flexible searching. Being able to create, store, and manipulate multiplets in compressed form - as *bitmaps* - has made it possible to integrate them into high-throughput technologies. A number of key parameters affect how well multiplets perform, including the granularity of edge length binning; how different multiplets are weighted in creating hypotheses from multiple ligands; and the number of bits that should be included in a pharmacophore hypothesis. The similarity metric employed for bitmap comparisons also affects search performance, as does the conformational sampling regime used for characterizing flexible molecules. In this report we explore the effect of parameter variation on within- and between-class similarity across seven different pharmacological classes and introduce a new measure of molecular similarity - the asymmetric stochastic cosine - uniquely suited to searching a database for matches to query hypotheses deduced from multiple ligands. Surprisingly, it turns out that the most discriminating bitmaps are obtained using relatively few conformers. The extreme discrimination power seen for single conformers, however, seems to reflect consistent effects of 2D connectivity on the 3D structure obtained. Conformational sampling by systematic search reinforces such circumstantial discrimination and should be avoided. The potential for systematic bias becomes clear when the behavior of otherwise similar conformational ensembles created by local energy minimization or by random sampling is considered. Consolidating information from multiple known actives or establishing single “bioactive” conformations *a priori* are safer ways to improve discrimination in pharmacophoric multiplet searching.

INTRODUCTION

Modeling protein−ligand interactions is a key step in modern drug design. If the structure of a binding site is known, docking studies can provide valuable insight into such interactions and, in favorable cases, make it possible to identify high-affinity ligands by virtual high-throughput screening (vHTS). For many proteins, however, no structure is available or the target protein is so flexible that accurate docking studies are difficult or impossible to carry out. Ligand-based analysis is very attractive in such cases, with pharmacophore-based 3D searching having proven itself particularly useful.[1] Pharmacophore distance multiplets - wherein the overall pharmacophore is characterized in terms of subgroups of two, three, or four features and the spatial relationship between them - provide a particularly fast and flexible tool for carrying out such searches. They may be based on a single conformer of one or more ligand molecules[2] or extended to encompass a number of conformers that the ligand molecule can adopt.[3,4] The latter approach allows multiple potential binding modes for a ligand molecule to be encoded into a single fingerprint descriptor. These descriptors were originally developed for diversity analysis and library design, [5−7] but they have more recently been used to create queries (“multiplet hypotheses”) for screening 3D databases.[2−4] The exact character of bitmaps for candidate ligand molecules and for the query hypotheses used to search them is determined by several parameters, each of which affects qualitative and quantitative performance to a greater or lesser degree.

The importance of some of these parameters has been debated over the years, including the appropriate granularity to use in mapping the continuous distances between features into the discontinuous space represented by binary fingerprints; [8] the appropriate similarity measure to use in comparing pharmacophore fingerprints; [8,9] and how the accessible conformational space is best sampled.[10] These critical questions can all be addressed using the fast pharmacophore multiplet technology embodied in the Tuplets module of SYBYL.[11]

How best to combine fingerprints from multiple actives has been much less thoroughly examined than the questions cited above, in part because the diversity analysis for which multiplet methodologies were originally developed focused primarily on 1:1 comparisons.[8,12−14] We introduced the concept of *Tuplet hypotheses*, which are pharmacophore multiplet bitmaps comprised of the most discriminating bits shared by a set of known actives.[3] Search results can be affected substantially by how the structures in the training set and in the database of candidates to be searched are

* Corresponding author phone: (314) 660-4499; e-mail: bclark@bcmetrics.com. Current address: Biochemical Infometrics, 827 Renee Lane, St. Louis, MO 63141.

[†] Current address: Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110.

10.1021/ci800234q CCC: $40.75
Published on Web 11/19/2008

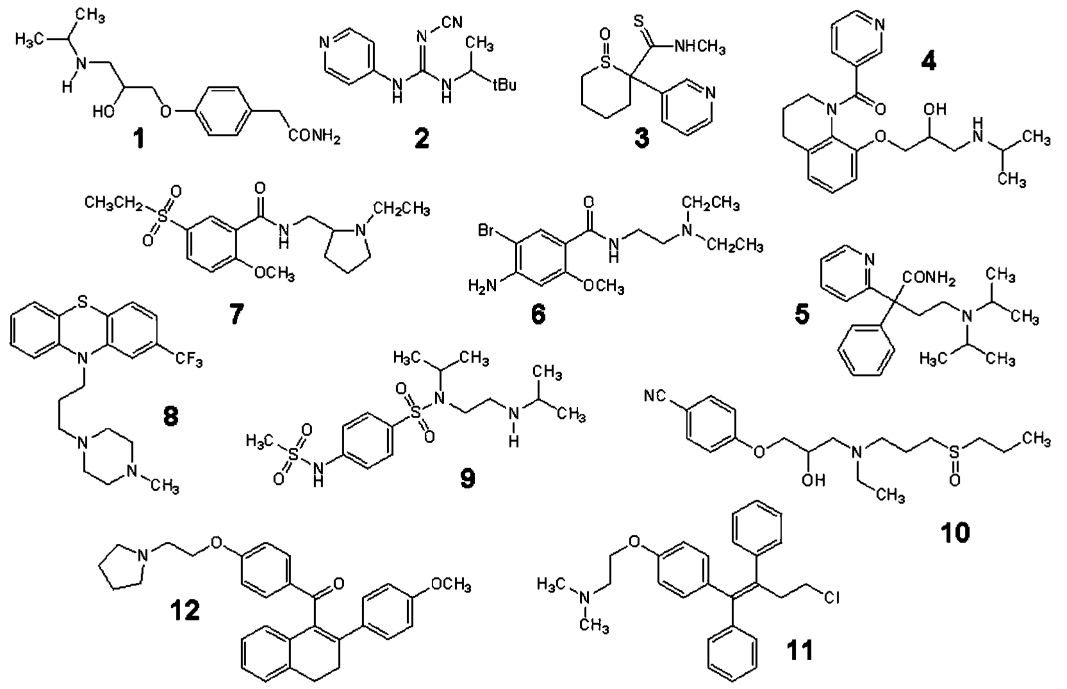

Figure 1. Representative structures from the seven pharmacological classes included in the data set include β-blocker 1 (atenolol); potassium channel openers 2 (pinacidil) and 3 (aprikalim); type I antiarrhythmics 4 (nicainoprol) and 5 (disopyride); benzamides 6 (bromopride) and 7 (sultopride); phenothiazine 8 (trifluoperazine); type III antiarrhytmics 9 (risotolide) and 10 (almakolant); and estrogen antagonists 11 and 12.

manipulated as well as by how stringent conditions are when selecting which particular multiplets are ultimately represented in the hypothesis. Relevant variables include the granularity of distance binning; the number of bits set (multiplets included) in the hypothesis; how conformers are generated; and the number of conformers considered for each compound.

Here, we surveyed the distribution of interfeature distances across classes to identify a set of distance bin definitions suitable for most applications. The similarity of the compounds in each class was then compared to the similarity to compounds in all other classes using three established similarity coefficients: the Tanimoto, the cosine, and the stochastic cosine.[3] We also examined the behavior of a new similarity coefficient that is better able to compare Tuplet hypotheses to bitmaps for individual target compounds - the *asymmetric stochastic cosine*. Finally, we examined the effect of the number of conformers used for bitmap construction on discrimination for various methods of conformer generation and found that the best discrimination is found at surprisingly low levels of conformational sampling, in part because underlying consistencies in how the initial 3D structures are generated account for much of the total discrimination observed and in part because critical contingent information is lost when too many bitmaps are OR'ed together.

The work described here was originally presented, in part, at the third Joint Sheffield Conference on Cheminformatics.[15]

METHODS

Data Set. The data set used by Mannhold et al.[16] to characterize molecular log P predictions is made up of compounds that fall into six broad pharmacological classes: type I and type III antiarrhythmics, β-blockers, potassium channel openers, and two classes of neuroleptic agents: phenothiazines and benzamides. The set was augmented by addition of twenty estrogen antagonists taken from the literature[17,18] to yield a total of 88 compounds. This set of molecules spans a wide range of pharmacophoric types, size, and flexibility. Moreover, though all of the compounds in it are drugs and therefore "druglike", the data set was compiled with an eye toward structural diversity and a spread of physical properties rather than on the basis of molecular targets. As a result, it exhibits less intrinsic analog bias than larger sets drawn from crystallized complexes are prone to.[19] Classification was based on pharmacology rather than on biochemistry, however, so the various classes differ considerably in their pharmacophoric consistency. They also vary somewhat in their structural homogeneity: UNITY[20] 2D fingerprint similarity within classes averaged 0.611 (0.418 to 0.807; SD ± 0.173) and similarity between classes averaged 0.313 (0.281 to 0.332; SD ± 0.242).

Figure 1 shows the structures of one or two compounds from each class, the particular examples being chosen so as to be representative of their respective class in terms of structure and size.

Conformational Sampling. Concord[21] provided the 3D structures used as starting points in all cases. Discrete conformational ensembles were pregenerated by augmenting these with systematic search[22] or with Confort.[23]

Confort is an extension of Concord that surveys torsional space to efficiently identify diverse sets of conformations from low-lying local energy minima. This is done by analytically solving for inflection points in torsional energy profiles for cyclic and alicyclic bonds and by carrying out subsequent diversity selection in distance space. Here the program was set to sample up to 20 rotors for each molecule and to generate up to 500 nonredundant conformers. The 200 most diverse conformations in this set were selected, and their geometries optimized using the Tripos force field[24] as implemented within Confort. Coordinates from the 100 optimized conformers lowest in energy were stored in a UNITY multiconformer database for those molecules that had sufficient conformational flexibility, along with all conformers for less flexible structures.

For systematic search, pharmacophorically relevant rotatable bonds were marked on each compound and sampled at 60° increments over a 360° range. All conformers within 30 kcal/mol of the minimum were preserved. For many of the compounds in the training set, the default general van der Waals scaling factor of 0.95 was so stringent that few if any conformers passed the energy filter. In those cases, a bump check was performed, and the compensating van der Waals factor indicated by the program was used for subsequent work. The conformers obtained from these computations, too, were loaded into a UNITY database. For twenty of the compounds in the data set, the number of conformers generated by systematic search exceeded 1000, with six compounds affording more than 10000 conformers. For these compounds, the 1000 lowest energy conformers were separated out, and another database was created for them.

When continuous ("random") conformational sampling was desired, it was carried out on the fly by serial randomization of the torsions about all rotatable bonds, after which the directed tweak algorithm[25] was applied to relieve steric clashes.[3,26]

The number of conformers obtained for the various ligands using the different sampling methods is provided as Supporting Information.

Bitmap and Hypothesis Generation. Pharmacophore doublet and triplet bitmaps were generated and analyzed using the Tuplets module in SYBYL 7.0 and five feature types: donor atom, acceptor atom, hydrophobic center, positive nitrogen, and negative center. This entails identifying the features in each molecule and then extracting the coordinates for the corresponding atoms and centroids in each conformation of that molecule. These are passed on to the bitmap generator via an XML stream. Identifying the features and extracting the coordinates are rate-limiting steps in this process, so it is much more efficient to redirect the XML stream for each molecule for storage in a database and retrieve it as needed rather than to send each directly to a bitmap creation or comparison routine. The stored XML files can subsequently be used to create different types of multiplets - e.g., pairs, triplets, or quartets. This was done here to speed up the many related comparisons made among compounds in the data set.

In order to support stochastic similarity measurements (see below), each bitmap consists of four parts: a bitmap union taken across all conformations, a bitmap intersection across all conformations, and two component bitmaps, each representing the union across one-half of the conformations, with conformations assigned to one or the other subpopulation at random. The four parts of a bitmap constructed from a single conformer are identical by convention.

The first step in creating a Tuplet hypotheses is to construct a vector whose elements are determined by the number of compounds (ν_i) in the training set in which the corresponding multiplet i appears and weighting factors based on feature type and distance bin - φ_k and δ_j, respectively. The contributing weights are summed across the three feature types and three pairwise distances that define a pharmacophore triplet or the four features and six distances that define a pharmacophore quartet. Values of the elements s_i in the full similarity vector are then calculated according to eq 1.

$$s_i = \nu_i \times \sum_{j=1}^{n_d} \delta_j \times \sum_{k=1}^{n_f} \phi_k \qquad (1)$$

Equation 1 addresses the fact that all multiplets are not created equal in terms of discrimination power. Those which are found in most or all actives are obviously important, but the spacing between features is important as well.[27] In particular, closely spaced features tend to represent common substructures like primary and secondary amides, each of which includes an acceptor atom (the carbonyl oxygen) as well as a donor atom (the amide nitrogen). The triplet from a primary benzamide is another, more specific example. It is comprised of three closely spaced features - a hydrophobic center as well as the donor and acceptor atoms from the pendant amide group. More widely spaced features are less common, so larger multiplets are given additional weight. Experience has shown that differentially weighting feature types is not particularly useful, so each type was given the same weight for the experiments described here.

A hypothesis bitmap of size k is one constructed by setting the bits corresponding to those multiplets having the k highest scores s_i to 1 and the bits corresponding to other multiplets to 0.

Similarity Measures. The similarity between two candidate ligands can be assessed in several ways, of which the Tanimoto and cosine coefficients (denoted herein as T and Cos, respectively) have been used most extensively. Both are dependent only on bits set (i.e., multiplets found) in one or the other bitmap; bits set in neither do not contribute to similarity. Moreover, each is scaled with respect to the overall cardinality of the pair. For binary fingerprints and bitmaps

$$T(a,b) = \frac{|a \cap b|}{|a \cup b|} = \frac{|a \cap b|}{|a| + |b| - |a \cap b|} \qquad (2)$$

where a and b are the bitsets or bitmaps corresponding to the fingerprints being compared and the vertical bars indicate cardinality.[28] Similarly, the Cosine coefficient is given by

$$Cos(a,b) = \frac{|a \cap b|}{\sqrt{|a| \times |b|}} \qquad (3)$$

A potential problem with applying either of these measures to multiplet bitmaps is that they can be dependent on exactly how the accessible conformational space is sampled. In particular, they may be inappropriate for stochastic sampling methods.[10] The stochastic cosine coefficient (Sto)[29] is defined by

$$Sto(a,b) = \frac{E(|a \cap b|)}{\sqrt{E(|a \cap a'|) \times E(|b \cap b'|)}} \qquad (4)$$

where $E()$ represents an expectation, and a and a' are replicate conformational samples, as are b and b'. This amounts to scaling by the self-similarity to compensate for random variations in conformational sampling, which makes the expected value of $Sto(a,a')$ unity, by construction. It is an efficient alternative to forcing convergence to an arbitrary

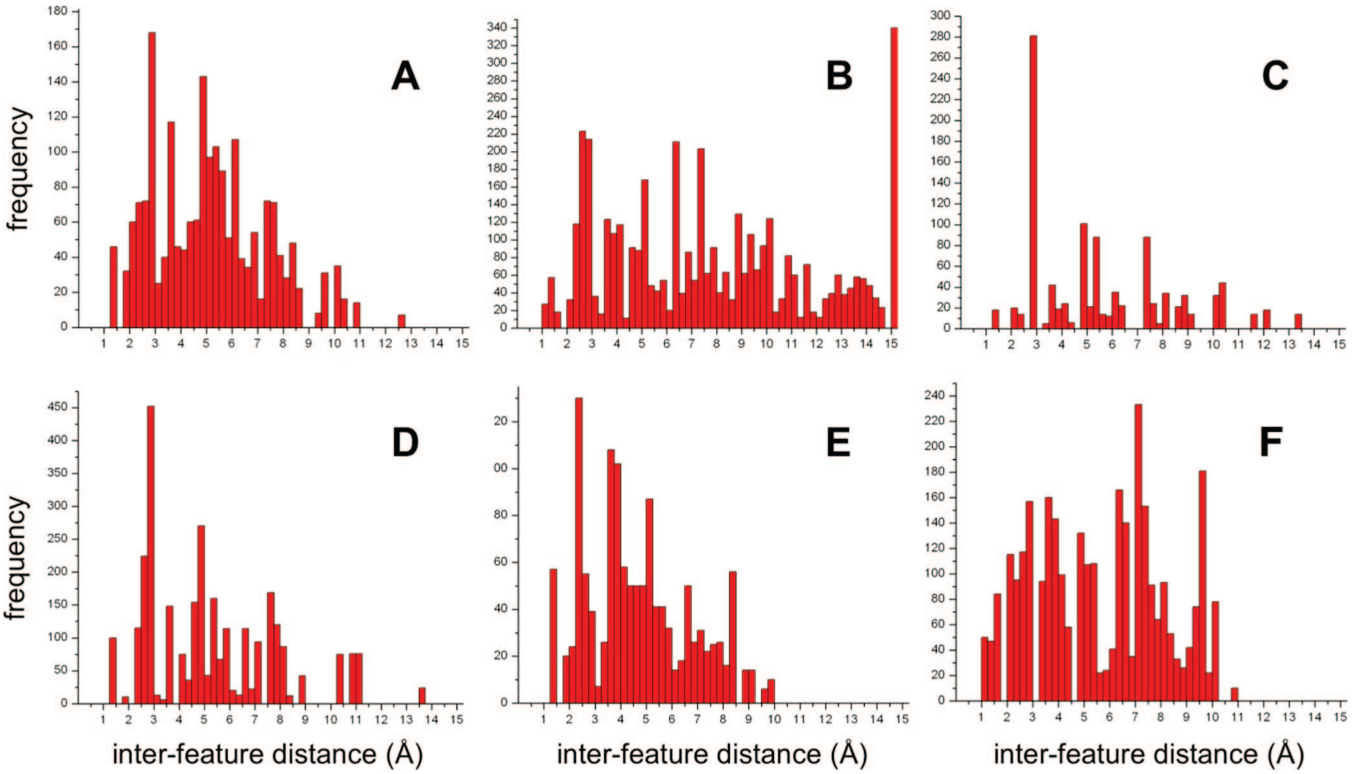

Figure 2. Frequency of occurrence for interfeature distances within activity classes for Concord structures from the data set compiled by Mannhold et al:[16] (**A**) type I antiarrhythmics, (**B**) type III antiarrhythmic, (**C**) β-blockers, (**D**) phenothazines. (**E**) potassium channel openers, and (**F**) benzamides.

ensemble of conformations, especially when analogous ensembles are not uniformly accessible to all ligands.

Many more bits are usually set in the bitmap for a flexible ligand than the k set in the corresponding Tuplet hypothesis, which can distort the meaning of symmetrical similarity measures such as those described above. In particular, a Tuplet hypothesis may show no appreciable similarity to any compound in a target database when too few bits are set in the hypothesis and a symmetric similarity measure is used (see below). This situation can be addressed by using the Asymmetric Stochastic Cosine (*ASC*) to evaluate multiplet similarity

$$ASC(q,b) = \frac{E(|q \cap b|)}{E(|q \cap q'|)} \tag{5}$$

where q is the query hypothesis, and b is a candidate bitmap from the target database. Note that *ASC*(*q*,*b*) is not in general equal to *ACS*(*b,q*) and that q may be the full bitmap of a single molecule rather than a consensus hypothesis based on several.

The partition of a standard bitmap into four parts - a Boolean AND across all conformers, a Boolean OR across all conformers, and bitmaps for two randomly selected subpopulations - is done in order to support the ASC and stochastic cosine coefficients. The expectation values are calculated from the two subpopulation bitmaps. For bitmaps based on a single conformation, the expected self-similarity is 1 by definition.

Discrimination between classes was evaluated by comparing the average similarity between test compounds and structures *within* the same pharmacological class to the average similarity to all of the structures in other classes.

RESULTS

Characteristic Pharmacophore Distance Distributions. In order to have each pharmacological class represented by a similar number of compounds, eight compounds were taken at random from each class, this being the number of compounds in the smallest class (the potassium channel blockers). To examine the distribution of interfeature distances, compressed count vectors were created using a binning scheme in which distances were binned in 0.25 Å increments between 1.0 Å and 15 Å. The total of 58 bins produced includes an "extra" pair of bins, one for all interfeature distances equal to or greater than 15 Å and another for those less than 1.0 Å. Perfectly coincident feature pairs such as the donor and acceptor atoms represented by the hydroxyl group in an alcohol were suppressed. Count vectors were then summed within each pharmacological class. Plots of the frequency distributions for single conformations produced by Concord are shown as Figure 2 for each individual class.

The class profiles clearly differ from each other, even when these very simple Tuplets - doublets with differences in feature type ignored - are considered. Nonetheless, some striking patterns emerge when the plots are overlaid, especially when frequency distributions based on multiple conformations are used (Figure 3). Some smaller interfeature distances occur surprisingly few times: bins from 1.75 to 2 Å, from 3 to 3.25 Å, and from 4.25 to 4.5 Å all have much lower frequencies than do the bins surrounding them.

This might have been expected for the single conformations because Concord typically produces extended conformations, but many of the distinctive peaks and troughs are also evident for the multiconformer distributions (Figure 3).

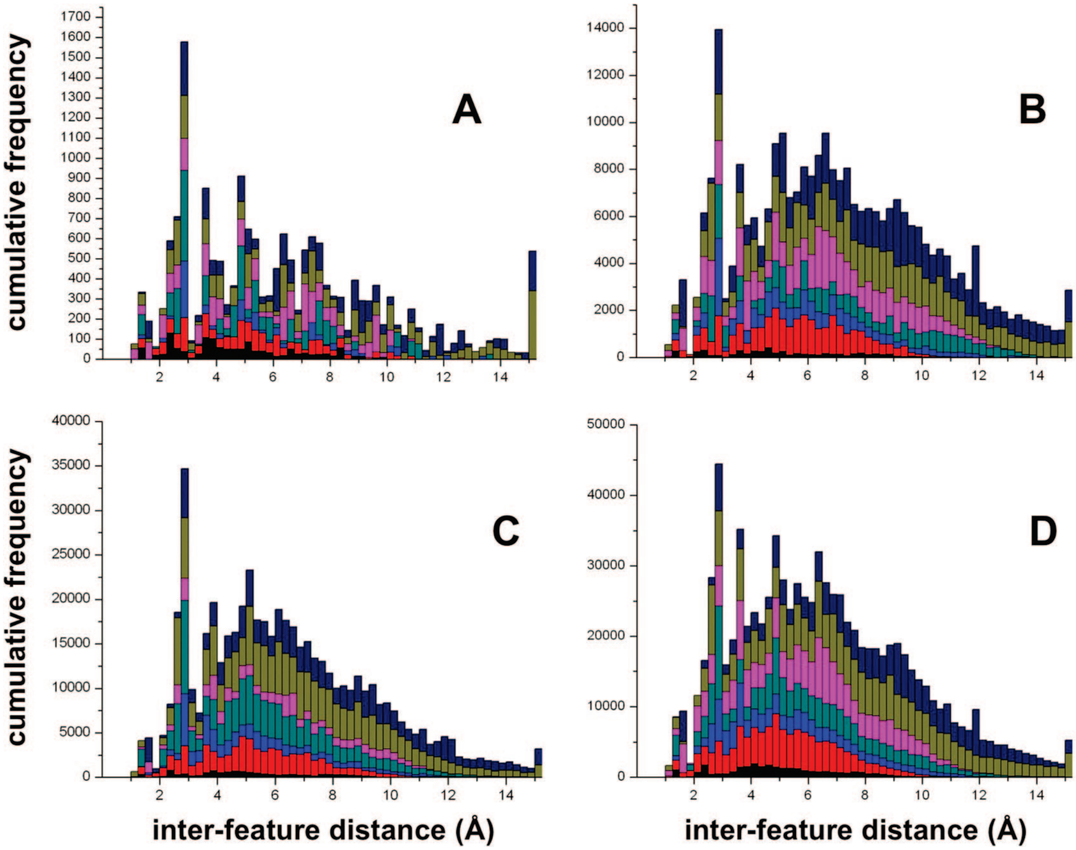

Figure 3. Cumulative frequency distribution for interfeature distances across activity classes. Stacked classes are potassium channel openers (bottom bars), type I antiarrhythmics, β-blockers, phenothiazines, benzamides, type III antiarrhythmic, and estrogen antagonists (top bars). Conformer sets consisted of single Concord structures (**A**) or conformational ensembles generated using systematic search (**B**), Confort (**C**) or random torsional sampling with steric clashes relieved using directed tweak (**D**).

Overall, the patterns of pharmacophore distance distribution are very similar whether systematic search, Confort, or random sampling is used to generate conformers. Note that fewer bits are set for the ensembles created using systematic search than for those generated by Confort, whereas random sampling sets more bits than does Confort. This is a direct result of the relative thoroughness of conformational sampling under the different sampling regimes. The average number of conformations generated by the systematic search was 1914, but setting the maximum number of conformers used for each molecule to 100 drops the average to 61, as this method generates fewer than 100 conformers for many of the compounds in the data set. This lower value was used in the analysis. The average number of conformers used by the directed tweak algorithm was 98, whereas Confort generated an average of 86 conformers per compound. It is this difference in sampling that accounts for the differences in the cumulative frequencies of the different methods.

Examination of the patterns of both the single and multiconformer distributions suggests the edge length binning scheme in Table 1 is a reasonably general ("natural") one for pharmacophore multiplets. The short-distance bins nicely bracket the distinctive peaks in this range, and bin sizes of 1 Å or less fit the profiles well up to at least 10 Å. The natural breaks between bins become less clear at larger distances, especially for the multiconformer profiles, but the data contain nothing to suggest that creating larger bins is necessary or warranted.

Table 1. Generalized Edge Length (Interfeature Distance) Bin Definitions (Å) and Weights

lower bound	upper bound	bin weight
0.0	1.75	1
1.75	3.00	1
3.00	4.00	2
4.00	5.00	3
5.00	6.25	4
6.25	7.5	5
7.5	8.75	6
8.75	9.75	7
9.75	10.75	8
10.75	11.75	9
11.75	13.00	10
13.00	15.00	12
15.00	none	12

Conformer Counts and Similarity Measures. Understanding the influence of conformational sampling on the pharmacophoric multiplet comparisons is crucial to using these descriptors effectively. Bitmaps were created using the binning scheme shown in Table 1 for 1, 10, 50, 100, 300, or 1000 conformers, and the average pairwise similarity of each of the compounds to others in their class was compared to the average pairwise similarity of that same compound to all compounds *outside* its class using each of the three symmetrical similarity coefficients: the Tanimoto, the cosine, and the stochastic cosine. Plots of the average within- and between-class similarities as a function of conformer count are shown as Figure 4 for three classes: estrogen antagonists, benzamides, and potassium channel openers. Conformers were generated by random sampling in these cases.

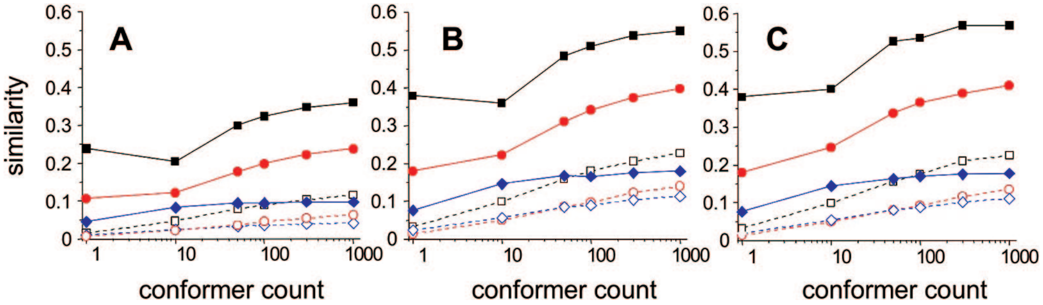

Figure 4. Dependence of average similarity measures within (filled symbols and lines) and between (open symbols and dashed lines) classes on the number of random conformations considered for different measures of triplet bitmap similarity. Results are shown for estrogen antagonists (□,■), potassium channel openers (●,○), and benzamides (◇,◆). Tuplet similarity was assessed using the Tanimoto (**A**), cosine (**B**), or stochastic cosine coefficient (**C**).

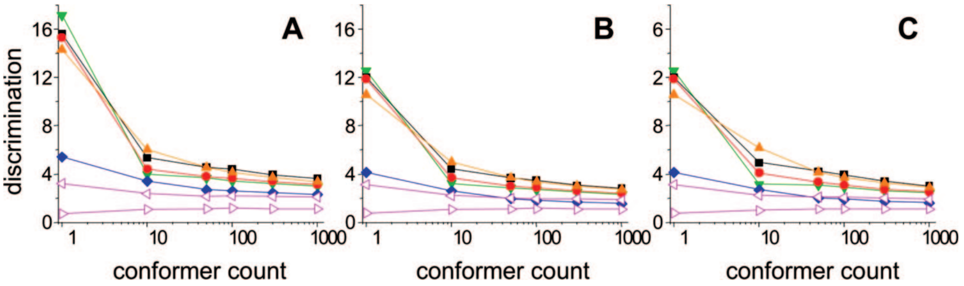

Figure 5. Dependence of discrimination on conformer count for β-blockers (▼), estrogen antagonists (■), potassium channel openers (●), phenothiazines (▲), benzamides (◆), and antiarrhythmics type I and type III (triangle pointing to the right and triangle pointing to the left, respectively). Tuplet similarity was assessed using the Tanimoto (**A**), cosine (**B**), or stochastic cosine coefficient (**C**).

Including more conformations in the ensemble bitmaps being compared increases similarity by all three measures, provided at least 10 conformers are being taken into consideration. Unfortunately, this is true for the between-class similarities (filled symbols) as well as for the within-class similarities (open symbols), the net result being that the ratio between the two - the *discrimination* - falls off as more and more conformations are included in the calculations (Figure 5). At the extreme, the level of discrimination falls to levels just above the levels observed for the corresponding 2D fingerprints. It should be noted, however, that the hits retrieved by the two methods are complementary with respect to individual actives.[4]

The triplet bitmaps are so sparse (<0.25% of the bits are set for 1000 conformers of the largest and most flexible molecules in the data set) that this result cannot be ascribed to random “collisions”, an effect encountered when fingerprints become saturated.[30] It is more likely due to the loss of contingent information - i.e., which multiplets occur *together* - that results from merging (an OR operation) a series of individual bitmaps, each of which represent an AND operation for a single conformation (see below).

Tanimoto similarities (Figure 5A) are consistently higher than the corresponding simple cosine similarities (Figure 5B). Some of that similarity can be “recovered” by using the stochastic cosine (Figure 5c), where the self-similarities included in the denominator during its calculation (eq 4) are less than unity for multiple conformers of flexible compounds.

The Role of Differences in Conformational Sampling Method. The discrimination for single-conformer bitmaps generated from naïve Concord structures for individual ligands deviate appreciably from the trend for all ligand classes and all similarity measures. The possibility that this is an artifact of the random conformational sampling ordinarily used in Tuplets analysis was explored by examining the corresponding profiles for conformational ensembles derived from systematic search or Confort as well (Figure 6). Here, the average discrimination is shown for the top four pharmacological classes (β-blockers, estrogen antagonists, potassium channel openers, and phenothiazines). Confort tracks random conformational sampling up to its 100 conformer limit, whereas bitmaps based on conformations derived from systematic search are significantly more discriminating than are bitmaps derived from either of the other two methods.

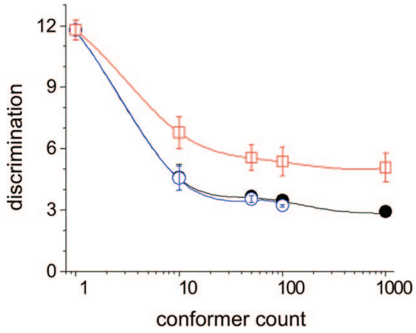

Figure 6. Within- vs between-class discrimination averaged across β-blockers, estrogen antagonists, potassium channel openers, and phenothiazines as a function of conformer count. Conformational ensembles were generated by torsional randomization (●), systematic search (□), or Confort (○). Single conformers were taken directly from Concord in all cases, and similarity was assessed in term of stochastic cosine similarity. Error bars are ± SEM, and points are connected with simple splines.

This result is striking, given that Confort is specifically designed to produce a diverse representative set of low-energy conformers, whereas systematic search and random sampling are both relatively energy-blind, simply ignoring conformations that exhibit steric clashes or tweaking torsions to relieve such clashes. The simplest rationalization is that the high discriminating power of the initial Concord con-

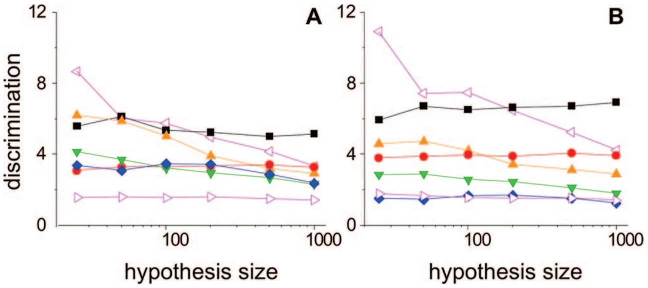

Figure 7. Dependence of discrimination on hypothesis size for β-blockers (▼), estrogen antagonists (■), potassium channel openers (●), phenothiazines (▲), benzamides (◆), and antiarrhythmics type I and type III (triangle pointing to the right and triangle pointing to the left, respectively). Tuplet discrimination for 100 conformer bitmaps was assessed based on symmetric cosine (**A**) or asymmetric stochastic cosine similarity (**B**).

formation is propagated through the elaborated ensemble by systematic search, so that if the starting torsions in the seed conformations differ by anything other than some multiple of the corresponding systematic search increment, all subsequent torsions will differ as well. Conversely, if the starting torsions in the seed conformations are very similar or differ by some multiple of the corresponding systematic search increment, the incremented torsional profiles will also be similar. Concord employs an incremental construction approach utilizing substructure based rules, reliably producing 3D structures that are consistent and energetically reasonable. The resulting consistencies within lead classes are useful in the context of flexible searching and topomeric analysis,[31] but they can distort multiplet similarities. This effect can be positive when members of a class share a relatively unusual type of substructure (e.g., phenothiazines) but can be negative when they share a substructure common to multiple classes, especially when that substructure is pharmacophorically rich (e.g., the ethoxyethanolamine moiety in **1**, **4**, and **10**). Note that the "cross-docking" type of validation approach used here, where drug molecules are used as decoys rather than some database of more generic structures, can only serve to exaggerate this effect.

Hypothesis Size and Asymmetric Searching. The analyses presented to this point have involved only pairwise comparisons between individual molecules. The most powerful use of fast pharmacophore multiplets, however, is in database screening applications in which hypothesis bitmaps created from *several* active compounds are used to screen large databases for pharmacophorically similar compounds.[2,3] The discriminating power of such searches is relatively insensitive to conformer count but is affected by the similarity measure used to compare the hypothesis to the bitmaps of compounds in the database and by the size of the hypothesis (i.e., number of bits included in it). The latter criterion is a reflection of the degree to which multiplets are shared among actives and how distinctive they are likely to be.

The behavior of similarity metrics was characterized by creating bitmap hypotheses for each class of compound from 100 conformer bitmaps and retaining 25 to 1000 multiplets in the hypothesis. The discriminative power for each hypothesis was then determined using symmetric and asymmetric stochastic cosine coefficients, with the results shown in Figure 7. Hypothesis bitmaps focus in on informative multiplets that are shared among multiple ligands, substantially increasing the discrimination for pharmacological classes that exhibit relatively large pharmacophores (type III antiarrhythmics, estrogen antagonists, and phenothiazines). Discrimination is enhanced further under ASC in most cases (Figure 7B), but those classes characterized by compact, rigid cores (e.g., phenothiazines) fare less well under the asymmetric measure. This is especially so when that core is a relatively common one, as in the case of the benzamides.

Hypotheses specifying as few as 25 multiplets are nearly as discriminating as single Concord conformations. The fact that small hypotheses perform well indicates that the weighting scheme used to rank multiplets for inclusion in the hypothesis (eq 1) is a reasonable reflection of reality. The fact that adding less informative multiplets to form larger hypotheses can degrade performance in some cases is broadly consistent with the effect of including more conformations in target and query bitmaps. It, too, can be rationalized as reflecting a loss of contingent information. Discrimination was not particularly dependent on training set size (varying from 8 to 20 here), an observation that is in accord with our previous experience.[4]

DISCUSSION

The bin sizes used to map the continuous distance space into the discrete distance bins of pharmacophore multiplet space is a critical consideration when encoding information from the molecules, as is how many bins are used. Previous efforts have used either fixed increment bins or a proportional binning scheme[8,32,33] where the size of the bin was dependent on the middle point distance of the bin. Alternatively, they have used encoded perimeter and area information to provide a potential descriptor.[12] As a practical matter, the number of bins used has often been limited to keep fingerprint sizes manageable in terms of either memory demand or stored file size, neither of which limits Tuplets analysis. An analysis of populations of hits from rigid 3D database searches of single conformer databases[32] has indicated that there are more pharmacophoric "hits" at shorter distances, and so the authors opted for finer binning at smaller distances, a conclusion that is in good qualitative agreement with our results. Our scheme is based on detailed empirical data and is, in that sense, "natural". Whether or not it performs better than others in prospective studies remains to be seen; it is likely to depend upon the target and the context in which it is applied.

Despite clear quantitative differences in the similarity values produced, the various symmetric similarity measures considered here yield very similar discrimination profiles, indicating that those differences are similar within and between classes and therefore largely cancel out. The only exception to this is that stochastic cosine similarity gives better discrimination at intermediate conformer counts, where it serves to offset sampling differences for the random sampling used here. Such rescaling by the expected value of the self-similarity is eminently reasonable for any random conformational sampling approach. If this is not done, the pharmacophoric similarity between a flexible ligand and another snapshot of itself will be less than unity, a behavior which is clearly inconvenient even though it reflects the physical reality of having a moving molecular target. It has been suggested that full local minimization is necessary for robust multiplet characterization, [10] but the results presented here comparing random conformational sampling with Confort conformers (Figure 6) indicate that stochastic normalization is an effective alternative.

The consistent falloff in discrimination with increasing conformer count (Figures 5 and 6) is unexpected and somewhat unsettling. Nonetheless, the effect is strong and evident across all seven pharmacological classes and across all three conformational sampling regimes, so it is likely to be very general. The falloff has not been noted earlier, probably because others have relied on systematic search as a way to ensure "exhaustive" conformational sampling, where increasing sampling was a matter of geometrically decreasing the torsional increment (60°, 30°, 15°,...) applied to each bond.[8] Given such a bias in sampling, it is not surprising that the diminishing returns seen here with stochastic conformational sampling was not noticed, though some earlier studies hinted at the fact that surprisingly few conformations are required to capture most relevant multiplets.[10]

Another reason for the effect of increasing conformer count is that most discriminating pharmacophores involve more than three features. Such a pharmacophore manifests itself in the bitmap of a single conformation as a *combination* of several triplets. If symmetries are ignored, four features generate four triplets, five features generate twelve triplets, six features generate 72 triplets, and so forth; all must be set together to fully match all features in the molecule. When several such bitmaps are OR'ed together, information about which multiplets belonged together is lost. Actually, this is desirable to some degree: inclusion of some features is often favored but not required, as is the case for "partial match" pharmacophores.[34] As bitmaps from more and more conformers are combined, however, the ANDs implicit in the individual bitmaps will be diluted out and discrimination suffers. This issue does not arise in diversity applications, where identifying a new multiplet - i.e., setting a new bit - is an unalloyed benefit. A similar effect accounts for cases where discrimination falls as the number of bits set in the hypothesis increases (Figure 7).

This confounding effect does not inadequately explain the falloff in discrimination between bitmaps constructed using one and ten conformers, however. The dependence on how conformers are generated (Figure 6) indicates that this effect is due to similar starting conformations being generated for structurally similar molecules. This also accounts for the elevated class-to-class variation in discrimination (i.e., larger SEMs in Figure 6) for systematic search and suggests that the results obtained using "exhaustive" conformational sampling approaches[8] akin to systematic search are biased and probably overly optimistic. Fortunately, random conformational sampling divorces query bitmap generation from the bias introduced by the dependence of systematic search on initial conformation. Such sampling, coupled with appropriate stochastic similarity measures, will help separate pharmacophoric similarity from substructural similarity, thereby encouraging more effective "lead-hopping".

The risk of bias might be worth the artificial increase in discrimination it produces were there no other way to achieve such an increase. Screening an ensemble of query bitmaps wherein each is based on a relatively small number of randomly chosen conformations is one alternative approach, albeit a computationally expensive one. It is probably better to look to increase discrimination by constraining the conformations going into the query rather than by reducing conformational sampling. The identification of shared discriminating multiplets through hypothesis generation seems to work well in many cases, as does generation of queries using conformations from crystal structures of ligand complexes or flexibly aligned structures.[2]

CONCLUSION

We performed several tests intended to characterize the behavior of pharmacophoric bitmaps and obtained several useful insights into that behavior. Examination of the pairwise interfeature distance distributions in our diverse data set of drugs revealed a "natural" distance binning scheme that captures most of the information content in such structures regardless of the conformational sampling method used. The pattern seen indicates that these typical distances are for the most part conserved across the full range of energetically reasonable conformers. Plots of discrimination (within- vs between-class similarity) show that only a modest amount of conformational sampling is required for maximal discrimination: 100 or fewer conformations proved generally adequate to sample accessible conformational space, at least for the moderately flexible compounds considered here. This is so because beyond this point, bitmap similarity between pharmacological classes increases at least as rapidly as does similarity within classes, the net result of which is a loss in discrimination. Coupling random conformational sampling with similarity measures that incorporate stochastic rescaling is a fast and effective alternative to full local minimization that avoids problems with potential bias due to shared substructures, though most practical applications of this technology will require consolidation of information of several actives into a unified query hypothesis; the preliminary identification of "active" ligand conformations; storage of several bitmaps for each molecule, each based on a small, independently drawn sample of conformations; or some combination thereof.

ACKNOWLEDGMENT

Anonymous reviewers provided several helpful suggestions, which the authors appreciate.

Supporting Information Available: Number of conformers initially obtained for each ligand in the data set tabulated for each different conformational sampling method. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *Pharmacophore Perception, Development, and Use in Drug Design;* Güner, O., Ed.; International University Line: La Jolla, CA, 2000.

(2) Shepphird, J. K.; Clark, R. D. A marriage made in torsional space: Using GALAHAD models to drive pharmacophore multiplet Searches. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 763–771.

(3) Abrahamian, E.; Fox, P. C.; Nærum, L.; Christensen, I. T.; Thøgersen, H.; Clark, R. D. Efficient generation, storage and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 458–468.

(4) Clark, R. D.; Fox, P. C.; Abrahamian, E. Using pharmacophore multiplet fingerprints for virtual HTS. In *Virtual Screening in Drug*

Discovery; Alvarez, J., Shoichet, B., Eds.; CRC Press, Taylor & Francis: Boca Raton, 2005; pp 207−225.

(5) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263–272.

(6) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB - Strategies for the Design and Comparison of Combinatorial Libraries using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144–150.

(7) Good, A. C.; Mason, J. S.; Green, D. V. S.; Leach, A. R. Pharmacophore-Based Approaches to Combinatorial Library Design. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker, Inc.: New York, NY, 2001; pp 399−428.

(8) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(9) Good, A. C.; Cho, S. J.; Mason, J. S. Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 523–527.

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

(11) *Tuplets in SYBYL 7.0*; Tripos International: St. Louis, MO 63144.

(12) Good, A. C.; Kuntz, I. D. Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373–379.

(13) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.

(14) Cato, S. J. Exploring Phrarmacophores with Chem-X. In *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O., Ed.; International University Line: La Jolla, CA, 2000; pp 107−125.

(15) Clark, R. D.; Fox, P. C.; Abrahamian, E. Characterization of Pharmacophore Multiplet Fingerprints as Molecular Descriptors. Presented at the Sheffield Conference on Cheminformatics. [Online], Sheffield, U.K., April 21−23, 2004. Abstracts. http://cisrg.shef.ac.uk/shef2004/abstracts.htm#cla (accessed January 1, 2008).

(16) Mannhold, R.; Rekker, R. F.; Sonntag, C.; ter Laak, A. M.; Dross, K.; Polymerpoulos, E. E. Comparative Evaluation of the Predictive Power of Calculation Procedures for Molecular Lipophilicity. *J. Pharma. Sci.* **1995**, *84*, 1410–1419.

(17) Waszkowycz, B.; Perkins, T. D. J.; Sykes, R. A.; Li, J. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM Syst. J.* **2001**, *40*, 360–376.

(18) Waszkowycz, B.; Perkins, T.; Baxter, C.; Sykes, R.; Li, J.; Harrison, M. Receptor-Based Virtual Screening of Very Large Chemical Datasets. In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, 2001; pp 372−381.

(19) Good, A. C.; Oprea, T. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.

(20) *UNITY 7.0*; Tripos International: St. Louis, MO 63144.

(21) Peralman, R. S. *Concord 7.0*; Tripos International: St. Louis, MO 63144.

(22) Beusen, D. D.; Shands, E. F. B.; Karasek, S. F.; Marshall, G. R.; Dammkoehlerb, R. A. Systematic search in conformational analysis. *J. Mol. Struct. THEOCHEM* **1996**, *370*, 151–171.

(23) Pearlman, R. S.; Balducci, R. *Confort 7.8*; Tripos International: St. Louis, MO 63144.

(24) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(25) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190–196.

(26) Dorfman, R. J.; Smith, K. M.; Masek, B. B.; Clark, R. D. A knowledge-based approach to generating diverse but energetically representative ensembles of ligand conformers. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 681–691.

(27) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 2. Applications of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295–304.

(28) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094–1102.

(29) Designated C* in ref 3.

(30) Flower, D. On the properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.

(31) Cramer, R. D. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J. Med. Chem.* **2003**, *46*, 374–388.

(32) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.

(33) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.

(34) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 567–587.

CI800234Q