Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching

Edmond Abrahamian,[†] Peter C. Fox,[†] Lars Nærum,[‡,§] Inge Thøger Christensen,[‡] Henning Thøgersen,[‡] and Robert D. Clark*[,†]

Tripos, Inc., 1699 South Hanley Road, St. Louis, Missouri 63144, and Novo Nordisk A/S, Medicinal Chemistry Research, Novo Nordisk Park, 2760 Måløv, Denmark

Received September 1, 2002

Pharmacophore triplets and quartets have been used by many groups in recent years, primarily as a tool for molecular diversity analysis. In most cases, slow processing speeds and the very large size of the bitsets generated have forced researchers to compromise in terms of how such multiplets were stored, manipulated, and compared, e.g., by using simple unions to represent multiplets for sets of molecules. Here we report using bit*maps* in place of bit*sets* to reduce storage demands and to improve processing speed. Here, a bit*set* is taken to mean a fully enumerated string of zeros and ones, from which a compressed bit*map* is obtained by replacing uniform blocks ("runs") of digits in the bitset with a pair of values identifying the content and length of the block (run-length encoding compression). High-resolution multiplets involving four features are enabled by using 64 bit executables to create and manipulate bitmaps, which "connect" to the 32 bit executables used for database access and feature identification via an extensible mark-up language (XML) data stream. The encoding system used supports simple pairs, triplets, and quartets; multiplets in which a privileged substructure is used as an anchor point; and augmented multiplets in which an additional vertex is added to represent a contingent feature such as a hydrogen bond extension point linked to a complementary feature (e.g., a donor or an acceptor atom) in a base pair or triplet. It can readily be extended to larger, more complex multiplets as well. Database searching is one particular potential application for this technology. Consensus bitmaps built up from active ligands identified in preliminary screening can be used to generate hypothesis bitmaps, a process which includes allowance for differential weighting to allow greater emphasis to be placed on bits arising from multiplets expected to be particularly discriminating. Such hypothesis bitmaps are shown to be useful queries for database searching, successfully retrieving active compounds across a range of structural classes from a corporate database. The current implementation allows multiconformer bitmaps to be obtained from pregenerated conformations or by random perturbation on-the-fly. The latter application involves random sampling of the full range of conformations not precluded by steric clashes, which limits the usefulness of classical fingerprint similarity measures. A new measure of similarity, The Stochastic Cosine, is introduced here to address this need. This new similarity measure uses the average number of bits common to independently drawn conformer sets to normalize the cosine coefficient. Its use frees the user from having to ensure strict comparability of starting conformations and having to use fixed torsional increments, thereby allowing fully flexible characterization of pharmacophoric patterns.

INTRODUCTION

Generalized formulations of the key interaction points in ligand binding to a specific protein—*i.e.*, pharmacophore hypotheses—play a key role in 3D database searching, making it possible to identify lead compounds which can interact in the same way yet fall outside existing lead series or patent estates. Fully flexible 3D searching has proven particularly effective in this regard. Given this success, it seems very reasonable to characterize ligands of pharmacological interest in terms of all possible pharmacophores they might present to a potential binding site. Unfortunately, even relatively small and rigid ligands can present remarkably complex pharmacophoric patterns, so it is generally necessary to decompose them into component feature multiplets of manageable size. The most common multiplets encountered are pharmacophore triplets and quartets. Triplets are defined in terms of the three features at the vertices and the three pairwise distances separating them, whereas quartets involve four features and six distances plus an indication of chirality, where necessary (Figure 1).

Hopes for this approach were buoyed by the successful application of 2D substructural fingerprints, based on small constituent fragments, in diversity analysis and library design.[1−3] Thereafter several groups investigated the use of pharmacophore distance triplets—trios of feature types and the three corresponding pairwise distances separating them[4]—in library design and diversity analysis.[5−12] Most of these groups used tools available in the ChemX software[13] (or in modified versions thereof[12]), in the PD Triplets module in SYBYL and UNITY,[14] or in analogous software suites

* Corresponding author phone: (314)647-8837 x3365; fax: (314)647-9241; e-mail: bclark@tripos.com.
[†] Tripos, Inc.
[‡] Novo Nordisk A/S.
[§] Current address: Combio A/S, Gamle Carlsberg Vej 10, DK-2500 Valby, Copenhagen, Denmark.

10.1021/ci025595r CCC: $25.00
Published on Web 01/28/2003

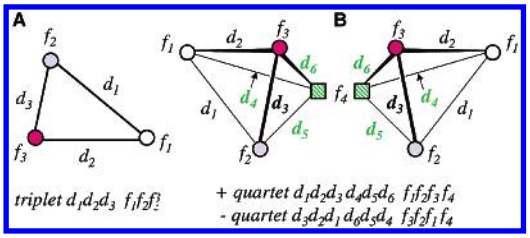

Figure 1. Schematic illustration of the multiplet encoding scheme used, where f_i denote feature types and d_i indicate pairwise distances (i.e., edge lengths). (A) Triplet encoding, where $d_1 \geq d_2 \geq d_3$. (B) Quartet encoding, where d_1 is the longest edge and $d_1 \geq d_2 \geq d_3$; f_4 is a privileged substructure, an extension point connected to f_2 or f_3 (augmented triplets), or a generalized feature type (quartets).

developed internally by pharmaceutical or academic research organizations.

The extensive literature available in this area makes it clear that, at least for existing implementations, simple triplets of features generally fail to capture pharmacophoric information at an adequately high level of complexity.[5,8−10] This is perhaps not surprising, given that a single pharmacophore query consisting of only three features is usually not specific enough for 3D database searching. Pharmacophore quartets are expected to be more discerning,[12] but their behavior has proven somewhat difficult to characterize adequately. This is in part due to their very large size: the length of four point pharmacophore (quartet) fingerprints increases as f^4d^6, where f is the number of feature types and d is the number of distance bins considered, with at least one extra bit needed to account for chirality (Figure 1B). Most analyses involve seven feature types and so are limited to nine or fewer distance bins if the resulting bitsets are to be kept below 32 bits in length.

Other difficulties in working with multiplets lie in uncertainty as to how pharmacophores presented by different conformations should be consolidated and in how pharmacophore fingerprints from different molecules—or sets of molecules —are to be meaningfully compared. Hydrogen bond donor and acceptor extension (site) points, which augment the information contained in the corresponding atomic features by indicating the points in space where complementary binding site features are likely to be located, are particularly problematic. That information is not captured in "classical" pharmacophore quartet fingerprints restricted to atom-centered features. Unfortunately, simply including extension points as independent feature types typically adds more noise than signal to the system.

Here we describe and characterize a particular method of encoding, compressing and manipulating multiplet bitmaps that supports fast generation, storage, manipulation, and analysis of very large pharmacophore triplet and quartet fingerprints as bitmaps without unduly straining memory resources, even for large databases. Data compression has been used by others in regard to multiplet storage,[15,17] but to the best of our knowledge, this is the first detailed report in which all manipulations are themselves carried out in a compressed format (i.e., as bitmaps) using an encoding scheme specifically designed to make such manipulations computationally efficient. The method is designed to operate fast enough to allow on-the-fly characterization of candidate compounds in terms of suitability for follow-up of hits generated by high-throughput screening (HTS), yet is flexible enough to incorporate privileged substructures[5] as well as augmented triplets, in which complementary extension points are linked to the corresponding hydrogen bond donor and acceptor atoms. Taken together, these tools provide a powerful infrastructure for pharmacophore multiplet similarity analysis in several contexts, including the 3D database searching application described here. Given its speed and flexibility, the method is well suited for use in combinatorial library design, just as earlier multiplet tools have been.[12,16,17]

METHODS

Conformer Generation. When only a single starting conformer was considered for a molecule, the conformer generated by CONCORD[18] was used. Where multiple, mutually diverse low-energy conformers were desired they were generated using CONFORT[19] and stored in a UNITY[14] multiconformer database for subsequent processing. When more conformers were called for in the configuration file than were provided in the specified database, the extra conformers needed were generated by randomizing the torsions about all rotatable bonds and using the directed tweak method[20] to relieve steric clashes in the perturbed conformations obtained.

Most pharmacophore multiplet analysis systems described in the literature to date have opted for torsional sampling at fixed increments so as to ensure "complete" conformational coverage. The configurations actually found in bound ligands often deviate significantly from the energetic minima found or calculated for isolated ligands, however. Moreover, the ensemble of conformations obtained is critically dependent upon the initial conformation to which the torsional increments are applied. This problem has been addressed by using a standardized, rule based system like CONCORD to generate starting conformations.[16] Unfortunately, there are still many cases (e.g., diaryl ethers) where near-equivalence of constituent substructures and having a multiplicity of low-energy conformations inevitably results in pharmacophorically similar molecules having quite different presentations of their constituent features.

Our aim was to create a system that would allow—but not require—complete conformational flexibility in the analyzed molecules. Except for very rigid structures, this implies that each bitmap will represent a finite sample of a very large number of possible conformations. Hence the conformations for each molecule were partitioned at random to form two subsamples to yield "replicate" pharmacophore multiplets, thereby providing a convenient mechanism for capturing the variability between different samplings of each conformational manifold. Four bitmaps were generated and stored for each molecule: one representing the union bitmap taken across all conformations; one union bitmap for each of two conformational subsamples; and, finally, a bitmap for the intersection across all conformations. This intersection bitmap can be used to quantitatively assess molecular flexibility, but its use will not be further described here.

Multiplet Generation. The first step in encoding multiplets is to sweep through the molecules of interest and identify pharmacophoric features specified in the multiplet definition file, *BinBounds.def*. These may be centroids of substructures or extension points, defined in-line using SYBYL line notation (SLN[21]) or by reference to the (user-

```
<STREAM>
   <META>
      <DATE>Feb 28 2002 13:06:53</DATE>
      <RANDOM_SEED>1015216305</RANDOM_SEED>
   </META>
   <STRUCT>
      <SLN_ID>3</SLN_ID>
      <CONF>
         <FEATS>
            <FEAT>
               <FEAT_TYPE>HY</FEAT_TYPE>
               <PROPS>
                  <XYZ>2.474 0.627 0.076</XYZ>
               </PROPS>
            </FEAT>
            <FEAT>
               <FEAT_TYPE>AA</FEAT_TYPE>
               <PROPS>
                  <XYZ>0.567 3.720 -0.460</XYZ>
               </PROPS>
            </FEAT>
            <FEAT>
               <FEAT_TYPE>DS</FEAT_TYPE>
               <PROPS>
                  <XYZ>-0.527 6.397 -0.247</XYZ>
                  <XYZ>-0.354 1.939 -2.555</XYZ>
                  <XYZ>-1.444 4.618 -2.346</XYZ>
               </PROPS>
            </FEAT>
         </FEATS>
      </CONF>
   </STRUCT>
</STREAM>
```

Figure 2. Example of XML format used for data transfer between 32 and 64 bit executables.

editable) default macro definition file, *sln3d_macros.def*. The features found are compiled into an ASCII file using the extensible markup language (XML) format illustrated in Figure 2, which is then passed on to a separate executable for generation of the desired bitmaps. The intermediate XML files constitute a "feature stream" which allow database access and feature identification programs to communicate with programs responsible for bitmap generation per se, so the latter function can be embodied in a 64 bit (or higher) executable even though the former functions are carried out by existing 32 bit programs. Being able to store feature streams also makes it possible to efficiently generate different "flavors" of multiplets in parallel, since feature identification is a relatively slow step in the overall process but is independent of the type of multiplet being generated.

Five centroid feature types are used by default: hydrogen bond donor and acceptor atoms; hydrophobic centers; positive nitrogens; and negative centers. The definitions for these and for the complementary hydrogen bond extension points[5] used in augmented triplets were identical to those distributed with SYBYL 6.8.1 and UNITY 4.3; a copy of the pharmacophore definition file is provided as Supporting Information for this article. These definitions take into account the likely protonation state(s) of common drug substructures at physiological pH as well as potential tautomerization of groups such as imidazole. The aromatic nitrogen in pyridine, for example, is marked as *both* a donor and an acceptor atom to accommodate the ambiguity of its likely protonation state; this reflects the ability of pyridines as a class to take on either role in protein binding, depending on the substitution pattern around the ring and the "local" pH of the binding site being considered.

Negative centers include oxy acids of sulfur and phosphorus, carboxylic acids, tetrazoles, sulfonimides, and acidic sulfonamides but do not include weak acid classes such as phenols. Similarly, positive nitrogens include substructures that are *always* positively charged: amidines and aliphatic amines, for example, but not weak bases such as anilines and pyridines. Some pharmacologically important special cases are also accounted for: *N*-acyl amidines, for example, are not basic and so are not marked as positive nitrogen centers. Features for complex acids and bases were usually specified as centroids rather than as specific atoms. Hence a carboxylic acid defines a single negative center positioned between the constituent oxygen atoms, and an amidine defines a single positive center positioned between the two nitrogen atoms.

No attempt was made to distinguish between aromatic and other hydrophobic features in the work reported here. Such a distinction can be introduced by editing *sln3d_macros.def*; the difficulty is in developing a self-consistent and broadly applicable definition for non-hydrophobic substructures (e.g., unsaturated lactams) capable of participating in aromatic interactions.

By default, edge lengths were binned starting with 0−2 Å (one bond length) and running up to 15 Å in 1 Å increments. All distances greater than 15 Å were consolidated into a final "omega" bin. This gave a total of 15 distance bins. Others have found that fewer, somewhat wider distance bins are better for quartets based on complete coverage at fixed torsional increments.[4,17] It is not obvious a priori, however, that the tradeoff between getting similar multiplets into the same bin and distinguishing between truly different ones would be the same in that case as for fully flexible fingerprints, especially when stochastic sampling effects are taken into account (see below).

Multiplet Encoding. If a pharmacophore multiplet analysis system is to be robust, an unambiguous, two-way, one-to-one mapping must exist between every possible pharmacophore multiplet and some position in the corresponding bitset. For triplets, the distances d_i are sorted in descending order and the vertex falling between the longest and shortest edges is assigned to the central position (f_2) in the feature index. The first vertex (f_1) is then set to match the feature connected to f_2 by the longest edge (d_1), whereas the third feature index (f_3) is determined by the type of the feature connected to f_2 by the shortest edge (d_3; see Figure 1A). In the event that different edges fall in the same distance bin, the tie is broken lexicographically, where feature priority is taken as the order in which features are specified in *BinBounds.def*.

The first step in encoding quartets involves identifying a base triangle, which is encoded as described above for triplets. The "extra" edge indices d_4, d_5, and d_6 then indicate the distances between the fourth feature (f_4) and f_1, f_2, and f_3, respectively.

Specifying two edges is sufficient to specify a base triangle. In parallel to the logic used for triplets, the base triangle is chosen so as to include the longest edge in the tetrahedron and the shortest edge that it shares a vertex with (which may not be the shortest edge of the tetrahedron). If there are two or more longest edges (in terms of the distance bin they fall into), the one bearing the shortest adjacent edge is given priority. Other ties are broken lexicographically, first by reference to d_2 (i.e., the base triangle is the one with the third edge falling into the largest distance bin) and then by feature types, just as described above for triplets.

Six distances and four feature types are only sufficient to completely specify quartets that contain symmetry elements. In the majority of cases chirality must be specified as well (Figure 1B). In the present scheme, chirality is assigned by taking the cross-product between the $f_2 \rightarrow f_1$ vector (along the

longest edge in the base triangle) and $f_2 \rightarrow f_3$ (along the shortest edge) to get a normal to the base triangle. The sign of the dot product between that normal and the $f_2 \rightarrow f_4$ vector then determines the nominal chirality of the quartet. Quartets of positive chirality are indexed by direct extension of the scheme for triplets, whereas the indexes of quartets with negative chirality are reversed with respect to the base triangle—i.e., from $d_1d_2d_3\ d_4d_5d_6\ f_1f_2f_3\ f_4$ to $d_3d_2d_1\ d_6d_5d_4\ f_3f_2f_1\ f_4$. Since the second edge index cannot be greater than the first under the encoding used for + quartets, this is by definition "free space" in the bitset. Note that all quartets for which the two orderings are completely identical are symmetrical and so are not chiral. Chirality may be a function of either the geometry of the multiplet (unsymmetrical edge lengths) or an unsymmetrical distribution of feature types or some combination thereof.

Multiplets involving privileged substructures are handled somewhat differently,[5] as are multiplets augmented with extension points. In this case, the privileged substructure or extension point is always assigned the "last" feature slot—f_4 in quartets involving privileged substructures and in augmented triplets (which are actually special cases of quartets). Although privileged substructures and extension points can be directly entered as feature types, specifying them separately is generally more powerful. When either is specified in the latter case, only multiplets containing the "extra" features in question contribute to the bitmap obtained.[22]

Bitmap Compression. There is no need to ever generate or manipulate pharmacophore multiplet bitsets as such. Rather, they can be generated, stored, and manipulated as compressed bitmaps, which goes a long way in accounting for the efficiencies in both processing time and memory usage for the method described here. Under the run-length encoding scheme used for compression, a bitmap is specified as a series of pairs of values, the first of which specifies a run content and the second of which specifies a run length (hence the name). For example, the bit*map*

0, 46; 1, 3; 0, 289; 1, 1; 0, 12; 1, 2; 0, 12234; 1, 5; 0, 2311; ...

corresponds to a bit*set* in which the first 46 elements are 0, the next three are 1, the next 289 are 0, the next one is 1, and so forth. [The punctuation used here is purely for clarity of presentation; the actual bitmap is simply a list of numbers in which run values alternate with run lengths.] Such compression is intrinsically very effective for pharmacophore multiplets because they are characteristically sparse, and become more so as their dimensionality increases. Note that "impossible" multiplets—whether they are such because they are geometrically impossible in Euclidean space (e.g., having 7, 2, and 2 Å edges) or because of the particular encoding scheme applied—form long runs of zeros when edge indices are given priority and sorted in order of length. The computational cost of creating, storing, and manipulating bitmaps scales with the number of runs, not their length, which is the reason the "edge-first" logic described here is preferred over "classical" feature-first encoding. Utilization of this scheme allows for the manipulation of the compressed bitmaps directly, without expansion of the full bitset at any time. Boolean and vector operations can be carried out with a significant reduction in computer resources as a result.

Hypothesis Generation. This scoring approach used here extends the minimum triplet area and frequency of occurrence filters employed by others[15] in a generalized way that better supports full conformational flexibility. Hypotheses were built up by sweeping through bitmaps for the training set to build a compressed count vector, each element of which indicates the number of bitmaps in which the corresponding "bit" is set, i.e., its frequency ν_i. A score S_i was generated for each element i based on that frequency, edge weights δ_j summed across the n_d distance bins (three for triplets and six for quartets) in i, and feature weights φ_k summed across its n_f feature types (three for triplets and four for quartets):

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

This formula is generalized from the discrimination power found for a series of searches using tetrahedral and triangular queries of various feature type mixes and sizes run against the Dictionary of Pharmacological Agents[23] using the UNITY 3D flex search engine. The form of the equation agrees with results of similar analyses carried out previously by other groups.[24] By default, the weight for the two smallest distance bins was set to 1 in the configuration file, with the weight for each subsequent bin in the series incremented by 1, yielding a sequence of weights of 1, 1, 2, 3, 4, 5 etc. This gives a summation over δ_j for the edges of a 6−5−3 triplet of $5 + 4 + 2 = 11$, for example. In the experiments described here, the feature weights φ_k were set to 1 for all feature types, so that the second summation simply equaled the number of features in the multiplet. This is a reasonably good first approximation so long as extension points are not included except as augmenting features, but it is unlikely to be optimal. Now that the requisite infrastructure is in hand, further research is underway to determine exactly what feature weights work best in general.

Nonzero scores were then sorted in decreasing order, and hypothesis bitmaps were created from the specified number of highest scoring bits.

Similarity Measures. Several similarity measures were explored for use in comparing multiplet bitmaps. The measure most commonly used[25,26] to assess similarity between two substructural bitsets a and b is the Tanimoto coefficient T

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

where the bracketing bars indicate cardinality. This is convenient to calculate and is readily obtained for unexpanded bitmaps by using a set of rolling counters to track positional indices in each of the two input bitmaps and within the current run in each bitmap. It is generally not optimal for pharmacophore multiplet applications, however, and several more or less complicated variants have been proposed as alternatives for torsionally constrained fingerprints.[12]

Allowing for full flexibility in the input conformations is particularly problematic. Tanimoto similarities between even modestly flexible molecules tend to be very low, unless the distance binning used is very coarse. In addition, such similarities are biased because similarities involving larger

and more flexible molecules are underestimated with respect to those involving rigid ones. A philosophical drawback is that for any conformational sampling scheme, the Tanimoto similarity of a molecule "to itself" will generally be less than unity. It must be appreciated that the real similarity being calculated is between *collections of conformations*, not between individual conformations. Because each conformation will likely display a distinct pattern of multiplets and similar conformations are unlikely to be produced by chance, different samplings of conformational space are likely to differ significantly, especially for complex multiplets. This problem can be overcome by normalizing the similarity by the extent to which replicate conformational samples reproduce each other. Doing so yields the Stochastic Cosine similarity, C^*

$$C^*(a,b) = \frac{E(|a \cap b|)}{\sqrt{E(|a \cap a'|) \times E(|b \cap b'|)}} \quad (3)$$

where E represents the expectation for the bracketed function obtained by averaging across multiple samples (here, of conformation), and primes indicate bitmaps obtained from distinct conformational subsamples. When there is no variation between subsamples (e.g., for rigid molecules and for complete coverage at fixed increments and a common starting point), this reduces to the more familiar[25] cosine coefficient:

$$Cos(a,b) = \frac{|a \cap b|}{\sqrt{|a| \times |b|}} \quad (4)$$

RESULTS

Storage Efficiency and Speed. "Raw", fully expanded triplet bitsets would occupy 53 Kbytes for the default parameters used here. In contrast, the average file size for bitmaps created from each of 1000 compounds drawn from among 123 000 compounds tested in the National Cancer Institute (NCI) anti-cancer screen[27] was 2, 3, and 6 Kbytes for 1, 10, and 100 conformers, respectively, when stored in 32-bit format. The corresponding quartet bitmaps[28] averaged 5, 12, and 40 Kbytes in size in 32-bit format and 9, 33, and 80 Kbytes in size in 64 bit format, a considerable savings over the 230 *Mbyte* required for the corresponding bitsets. As noted in the Methods section, each bitmap file actually includes four bitmaps—union and intersection bitmaps across all conformations plus union bitmaps evaluated for each of two conformer subsets.

Generation of the intermediate XML required an average of 0.08 CPU seconds for one or 10 pregenerated conformers on a Silicon Graphics, Inc., computer equipped with an R10K processor. Identification of features is rate-limiting, so 100 pregenerated conformers took only a little longer (0.10 s). Generating conformers on-the-fly using directed tweak was significantly slower, requiring 0.51 s on average for the NCI data set. Subsequent bitmap generation, on the other hand, required an average of only 0.01 s for one or 10 conformers and 0.09 s for 100 conformers. Loading bitmaps into memory from disk typically required 0.2 CPU s or less. Applying a Boolean "AND" across 100 bitmaps took 0.3 s or less, depending somewhat on the average bitmap size, as did applying a Boolean "OR." Realizing these speeds does require maintaining a cache of pointers to memory locations of recently "visited" bitmap regions, however.

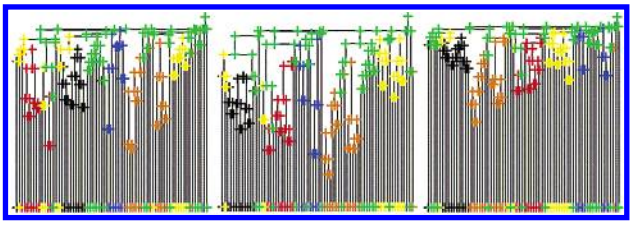

Figure 3. Hierarchical clustering (complete linkage method) for the set of β-blockers (highlighted in black), type I and III antiarrhythmics (green and yellow, respectively), benzamides (red), K^+-channel openers (blue), and phenothiazines (orange) compiled by Mannhold et al.[29] The pharmacophore multiplets used were based on 25 conformations of each compound generated using directed tweak. (A) Triplets similarities assessed in terms of the Tanimoto coefficient. (B) Triplets similarities assessed in terms of the Stochastic Cosine. (C) Augmented triplet similarities assessed in terms of the Stochastic Cosine.

Stochastic Cosine and Augmented Triplets. A set of 68 structurally diverse drugs spanning six pharmacological classes (class I and III antarrythmics, β-blockers, phenothiazines, benzamides, and K^+-channel openers) was originally compiled by Mannhold et al.[29] for use in evaluating molecular lipophilicity predictions. Here it is used to illustrate application of the stochastic cosine similarity measure for comparing pharmacophore multiplets (Figure 3). Triplet bitmaps were created utilizing the five default feature types (donor and acceptor atoms, hydrophobic centers, positive nitrogens, and negative centers) alone (Figure 3A,B) or augmented with associated acceptor and donor site points (Figure 3C). Pairwise Tanimoto (Figure 3A) or stochastic cosine similarities (Figure 3B,C) were then used as input to agglomerative hierarchical clustering using the complete linkage method.

Most of the pharmacological classes fell into one or two pharmacophore clusters under this similarity measure. This is not surprising for the structurally "tight" classes such as benzamides and phenothiazines, since structural similarity usually implies pharmacophoric similarity. It is more remarkable for the two types of antiarrhythmics and for the K^+-channel openers, because these classes are much more structurally diverse.

Shifting from the Tanimoto coefficient to the stochastic cosine produced a general reduction in the numerical dissimilarity values (increase in similarities), as indicated by the shorter basal steps and increased "white space" in the corresponding dendrograms (Figure 3A vs Figure 3B). A small qualitative improvement in the homogeneity of the clusters accompanies this increase in effective dynamic range as well as some changes in the relationships among classes.

As expected, the clustering obtained with augmented triplets is similar in quality to that seen for "regular" triplets (Figure 3C vs Figure 3B). The overall pattern of clustering is somewhat different, however, particularly with regards to the distribution of type I antiarrhythmics and the relationships among classes. More significant is the shift overall toward lower pairwise similarities which produces the elongated bottom "legs" and decreased white space in the dendrogram. This reflects their partial "quartet" nature as well as the intrinsically reduced discriminating power of extension points with respect to centroid hydrogen bonding features; it is even more pronounced when Tanimoto similarities or "regular" quartets are used. The loss in discriminatory power reflects,

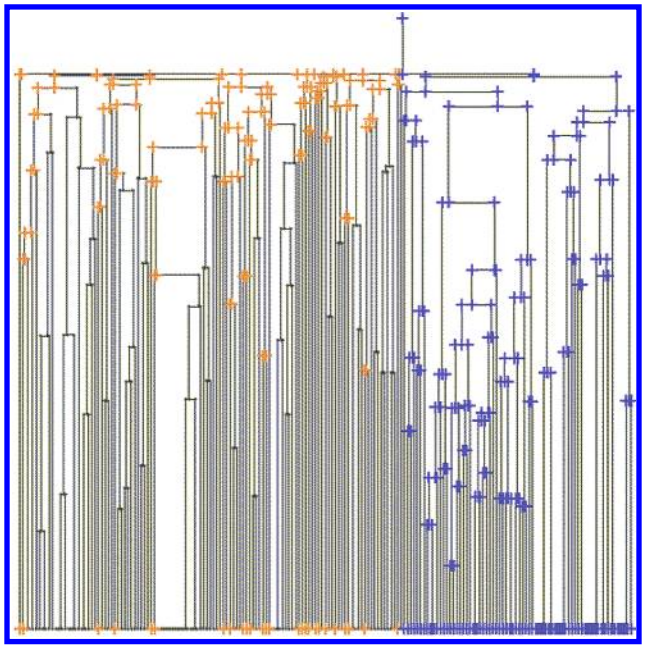

Figure 4. Hierarchical clustering of pharmacophore triplet bitmaps for 96 compounds from a GSK3 QSAR data set[30] together with 91 structurally similar compounds drawn from the NCI anticancer screening database.[27] Basal nodes corresponding to the GSK training set are highlighted: blue crosses indicate inhibitors with characterized IC50s, whereas orange ones indicate inactive Novo Nordisk analogues. Nonhighlighted basal nodes correspond to NCI decoy compounds.

at least in part, the torsional ambiguities of terminal rotors such as hydroxy and amino groups. Work is underway on treating extension point features as constraints—e.g., as a range of angles—as a way to address this problem.

A similar but more focused analysis was carried out on a set of 96 glycogen synthase kinase-3 (GSK3) inhibitors and analogues originally compiled for a quantitative structure−activity relationship (QSAR) analysis.[30] This was comprised of 70 active compounds and 26 less active analogues included in the study for the sake of completeness. An additional 91 were drawn from the NCI anticancer screening database[27] that exhibited a Tanimoto similarity of 0.65 or greater to the compounds in the training set with respect to standard UNITY substructural fingerprints. Twenty-five conformers were produced for each of the 187 compounds, and the triplet bitmaps so obtained were used for hierarchical clustering, with the results shown in Figure 4. The actives (blue crosses) all cluster together to the right in the dendrogram, whereas the inactive analogues (orange crosses) are distributed among the unhighlighted nodes corresponding to decoy compounds from the NCI database.

Multiplet Hypotheses for Estrogen Receptor Antagonists. One hundred random conformers were generated for each of 20 estrogen antagonists taken from the open literature (Figure 5),[31,32] and the 100 highest-scoring bits for several multiplet bitmap "flavors" derived from this training set were incorporated into the corresponding bitmap hypotheses. Compounds comprising a large fraction (385 000 compounds) of the Novo Nordisk Compound Database (NNCD) were then ranked with respect to the Tanimoto similarity of each to the hypothesis for the analogous multiplet, giving the results shown in Figure 6. The NNCD contains a total of 434 known estrogen antagonists (as characterized by their ability to displace estradiol), including eight compounds from the training set (Figure 5). Multiplets considered include standard triplets (Figure 6A,B); triplets augmented with a phenolic privileged substructure (Figure 6C); triplets augmented with hydrogen bond donor and acceptor extension points (Figure 6D); and standard quartets (Figure 6E,F). The target bitmaps for compounds in the database were constructed using one (Figure 6A,C−E) or 100 (Figure 6B,F) conformers for each target compound.

The top (red) curve in each panel shows the degree of enrichment found for actives with respect to the number expected for random selection (0.13%) as a function of the number of top-ranking NNCD compounds considered (note that the top rank order plotted is 1000, corresponding to only 0.26% of the database). The middle (black) curve shows the cumulative number of actives recovered at each rank. The bottom (green) curve represents the number of actives recovered based on pharmacophore multiplet similarity which exhibit a UNITY fingerprint Tanimoto similarity of 0.85 or greater to at least one compound from the training set. These actives would also be recovered by a simple substructural similarity search and include those eight compounds from the training set also found in the NNCD. The ranks obtained for the training set compounds themselves are indicated by green circles for those falling in the 1000 compounds most similar to the corresponding hypothesis.

The simple triplet hypothesis using a single target conformer obtained from CONCORD works remarkably well, recovering nearly 50% of the actives in the first 1000 "hits" (Figure 6A). Of these, just under half are structurally similar to the training set. A small bias in favor of the training set compounds themselves is evident in that six of the eight training set compounds present among the 434 actives in the NNCD are ranked within the top 200 overall and among the "best" 100 actives; the expected result would be 1.8 ($100\times(8/434)$) were there no bias *among actives*. A seventh makes it into the top 1% of "hits," whereas the training set compound least similar to the triplet hypothesis (**14**) falls near the 2% level.

Increasing to 100 the number of conformers contributing to the *target* bitmaps reduces the specificity with respect to inhibition by a factor of about 3 (Figure 6B), i.e., only about a third as many bitmaps of known actives appear among the 1000 most similar to the hypothesis. The number of substructural analogues "hit" is even more sharply reduced, however, the proportion dropping below 30%. In fact, the recovery of compounds from the training set in this case is unbiased—a single example versus an expected 1.4. That this reflects target flexibility is suggested by the lack of change in rank for the rigid pyrroloindolizine agonist NNC 45-0095[33,34] (Figure 7), indicated in Figure 6 by black squares. Other NNCD actives, such as those derived from diaryl chromanes analogous to NNC 45-0781[35,36] (Figure 7), are more flexible and therefore more difficult to recover. Stochastic hypothesis generation has yet to be implemented; doing so will enable a direct test of this rationale, since application of Stochastic Cosine similarity is expected to improve recovery if such is the case. Preliminary experiments (data not shown) indicate that increasing the number of high-scoring bits included in the hypothesis also serves to increase the number of flexible actives retrieved. Somewhat surprisingly, using an asymmetric similarity measure that discounts

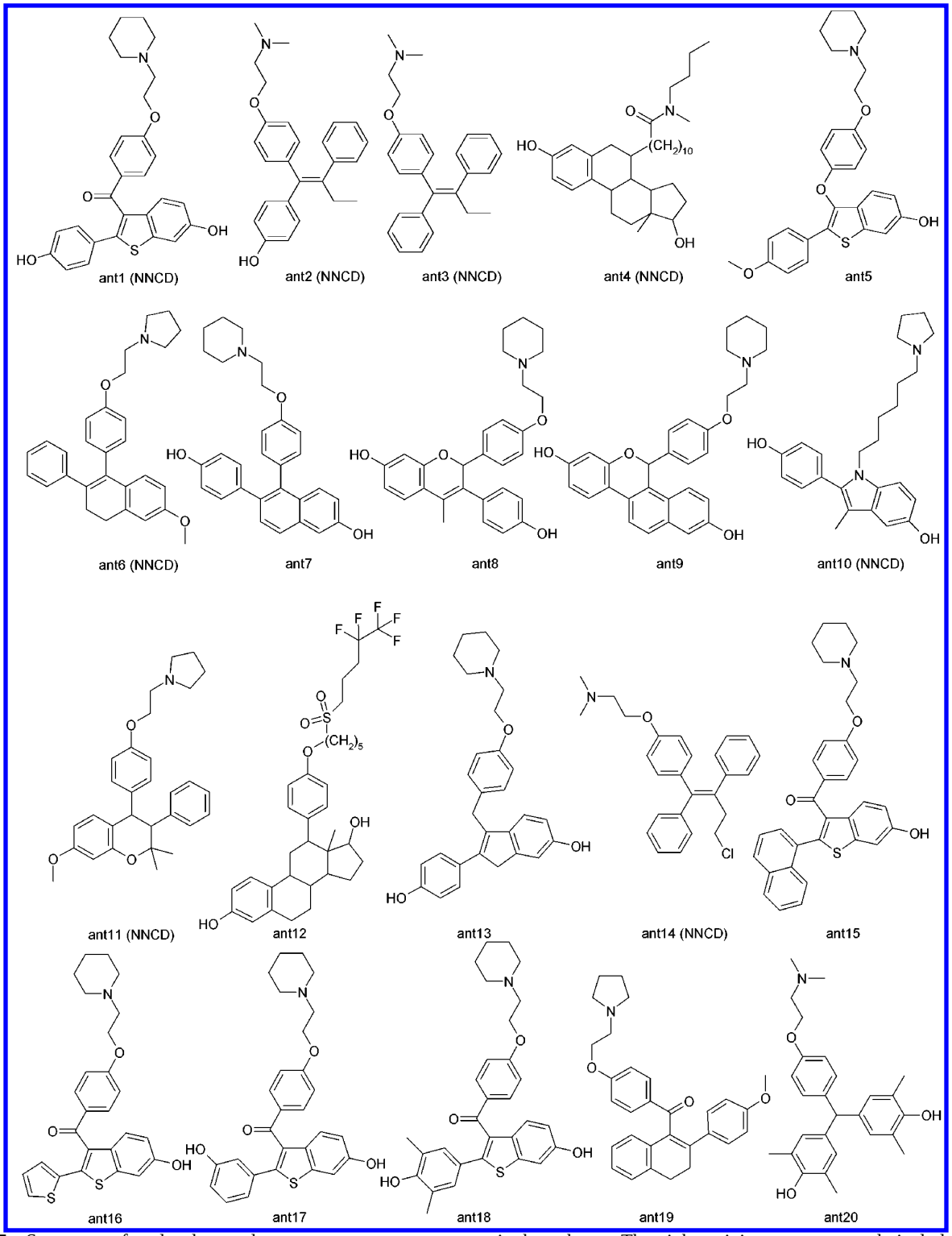

Figure 5. Structures of molecules used to construct estrogen antagonist hypotheses. The eight training set compounds included in the Novo Nordisk compound database are indicated by an appended NNCD.

the number of bits set in the target but not in the hypothesis did not improve performance substantially (data not shown).

Adding a phenol as a privileged substructure (Figure 6C) reduces recovery somewhat, particularly for training set compounds. It should be noted, however, that four of those compounds (**3**, **6**, **11**, and **14** in Figure 5) are methoxyphenyl derivatives lacking the privileged substructure altogether. The standard **10**, which is a phenol but does not appear in the top 1000 bitmap "hits", does make it into the top 1%. Going to the more generalized case in which triplets are augmented with hydrogen bond donor and acceptor extension points (Figure 6D) gives results intermediate between those for simple triplets and those augmented with phenol as privileged substructure.

Figure 6E,F shows recovery profiles obtained when the fourth multiplet vertex is fully generalized so as to obtain full-fledged quartet bitmaps. The proportion of actives structurally similar to the training set is reduced somewhat from that seen for triplets for the top 200−300 hits, with a concomitant reduction in overall recovery of actives of about

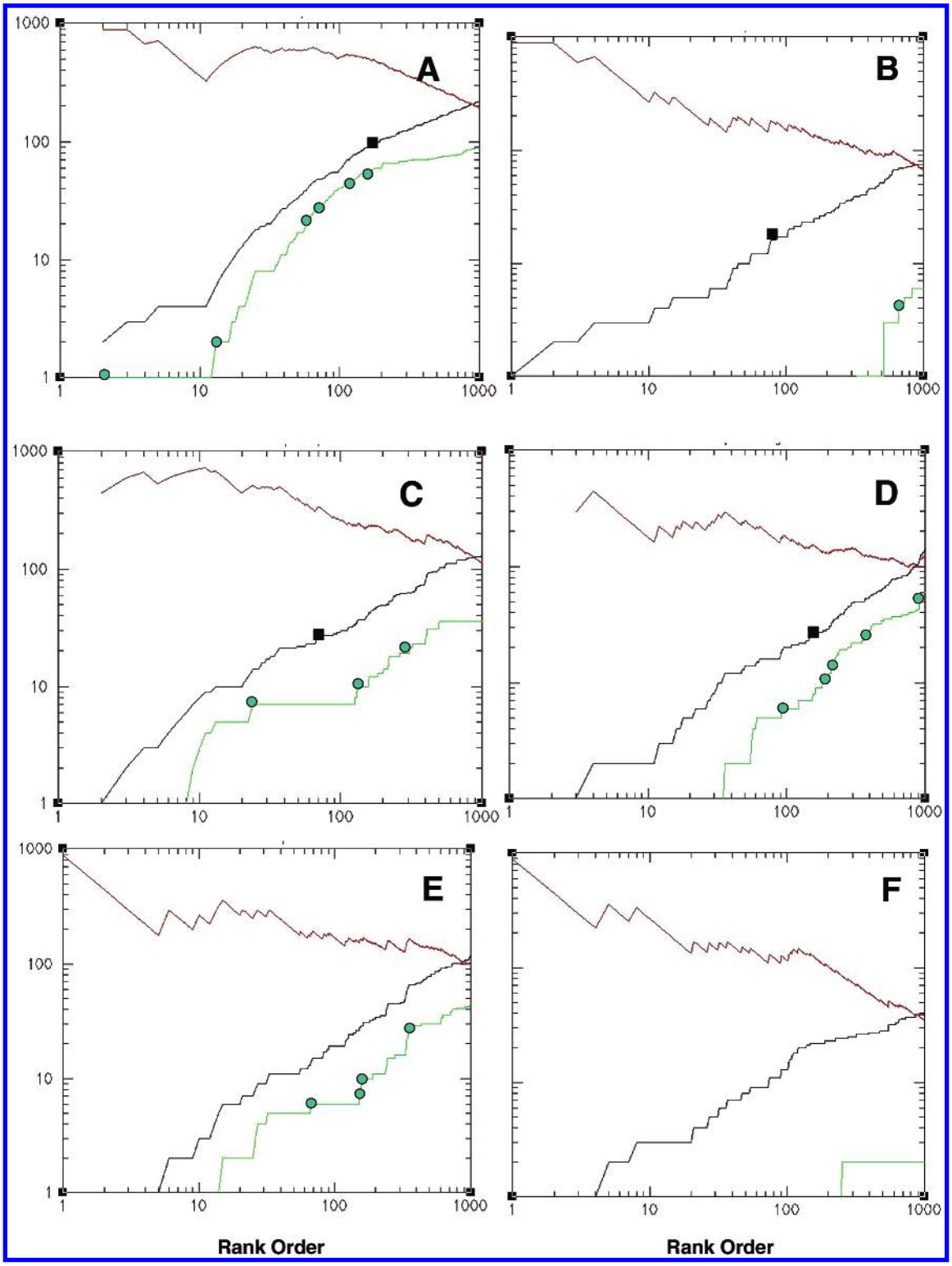

Figure 6. Recovery and enrichment curves for the 434 known estrogen antagonists in a 385 000-compound subset of the Novo Nordisk database for 100 bit multiplet hypotheses obtained from the training set shown in Figure 5, with 100 conformers considered for each compound in that training set. The upper (red) curves indicate cumulative enrichment as a function of rank order. The middle (black) lines show the respective recoveries of known actives. The lower (green) curves indicate the number of compounds recovered at the given rank which have a Tanimoto similarity greater than or equal to 0.85 to any compound in the training set with respect to UNITY substructural fingerprints. Note that both scales are logarithmic, with the top rank (abscissa) corresponding to 0.26% of the full 385 000 compound database surveyed. (A and B) Pharmacophoric triplet bitmap hypotheses. (C) Hypothesis built from triplets augmented with a phenolic privileged substructure. (D) Hypothesis built from triplets augmented with donor and acceptor extension points. (E and F) Pharmacophoric quartet hypotheses. Target bitmaps for structures in the database were based on contributions from single conformers (A, C, D, and E) or on 100 random conformers (B and F). Black squares indicate the lowest (best) rank for an analogue of NNC 45-0095, and green circles indicate the ranks for compounds from the training set.

40%. Otherwise the triplet and quartet profiles are surprisingly similar, whether the target multiplets are generated from one (Figure 6A vs Figure 6E) or from 100 (Figure 6B vs Figure 6F) conformations. It bears noting, however, that

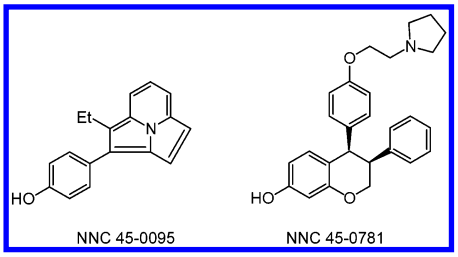

Figure 7. Representative hydroxychromane and pyrroloindolizine estrogen antagonists found in the NNCD.

Table 1: Composition of the 200 Highest Ranking Compounds

				training set		
multiplet	conf[a]	actives	enrichment	per se	analogues[b]	NC 45-0095 analogue[c]
triplet	1	104	461	6	61	193
	100	28	124	0	0	75
privileged	1	49	217	2[d]	13	70
augmented	1	29	129	2	11	151
quartet	1	34	151	3	11	2532
	100	23	102	0	0	2493

[a] Number of conformers contributing to each target bitmap. [b] Number of compounds with a UNITY fingerprint Tanimoto similarity of 0.85 or more to some compound in the training set. [c] Rank of the analogue with the highest Tanimoto similarity to the respective hypothesis. [d] Of the four containing the phenolic substructure used in generating the bitmaps and hypothesis.

binning ranges used are likely better suited to triplets than to quartets[12,17] and that the use of a static hypothesis rather than its stochastic equivalent probably penalizes the latter more than the former.

The distribution of training set compounds across the rank order parallels that seen for their structural analogues. Hence the ranks of the eight also found in the NNCD ran from a minimum of 2 and a median of about 100 for single conformer triplets up to a minimum rank of about 2000 and a median just slightly higher. Details of the recoveries of the training set compounds and compounds similar to them among the 200 bitmaps most similar to the hypothesis are given in Table 1, along with the rank of the most pharmacophorically similar NC 45-0095 analogue. The latter is a measure of the method's ability to "lead hop," since the pyrroloindolizine agonist bears little structural similarity to the compounds making up the training set. In interpreting the data in Table 1, it should be borne in mind that this represents the crème de la crème of the hits—the top-scoring 0.05% of the database. This format allows direct comparison literature results such as those reported by Pickett et al.[17] for searching against multiplet fingerprints for individual fibrinogen receptor antagonists.

DISCUSSION

One concern was that, at the end of the day, multiplets complex enough to be informative would be so specific as to be essentially a very complicated and expensive means of assessing substructural similarity. In fact, however, the clustering and virtual screening results seen here indicate substantial complementarity to methods based on similarity in substructural fingerprints. Augmentation with donor and acceptor extension points is one way to productively incorporate more pharmacophoric information into triplets without risking the potential over-specification involved in going all the way to quartets, but more work along these lines is needed.

The virtual screening results shown here are qualitatively similar to those obtained from "classical" flexible 3D searches, but the application of multiplet hypotheses differs from such searches in several key ways. Such hypotheses have elements of partial match search queries, in that no individual bit need be matched in any particular target, but they are "fuzzier" in that there is no need to prespecify how many need to match. Furthermore, the implicit partial match constraint in pharmacophore multiplet hypotheses applies across the groups of features represented by the multiplets, rather than to groups of individual features. Then, too, a single multiplet hypothesis can—and, if based on high-throughput screening results, usually will—represent overlapping ligand interaction modes with the target protein binding site, which is difficult or impossible to achieve in "classical" discrete 3D queries. This is complemented by the ability to use hierarchical clustering of bitmaps to identify truly disjoint binding mode classes within a brace of active compounds.

The false positive rates seen here are comparable to those seen with discrete queries, but the nature of the "bad hits" is different. Errors for multiplet hypotheses likely reflect misidentification of separated constituent pharmacophore multiplets in inactives that, in active compounds, form more complex pharmacophoric structures. Errors in discrete queries, on the other hand, more often reflect the presence of "bad" steric elements or interfering features. Given that the ultimate goal here is not so much to pick out individually good ligands as to characterize libraries and lead-hop to structurally distinct ligand classes, the false negative rates observed are not a severe problem. This may not be the case in less propitious circumstances, however.

Although effective, the multiplet hypotheses we generate are quite robust to false positives. Nor does over-training appear to be a problem, at least in the case of the estrogen receptor (Figure 6): hypotheses do exhibit some bias in favor of actives in the training set over those outside it but only to a modest degree. Increasing the number of conformer multiplets folded into a single bitmap reduces both specificity and bias substantially, particularly when multiple conformers of targets are allowed—the chances of hitting a conformation similar to one found in a training set compound get small quickly. It is hoped that the use of stochastic hypotheses and similarity coefficients will help ameliorate this problem.

Except in the case where privileged substructures absent from some of the actives were used, more than half of the training set compounds present were found within the top-ranking 1% of the NNCD. Interestingly, the major effect of increasing pharmacophore multiplet complexity was a sharp compression of the rankings for the eight training set compounds in the NNCD, at least on a logarithmic scale: the range was 4000-fold (2−8000) for triplets vs 50- and 65-fold for augmented triplets and quartets, respectively.

The data available to us for the estrogen receptor were precisely the kind needed to carry out the analysis presented here—a thorough biochemical characterization of a very large data set only incidentally overlapping with the well characterized, publicly available training set but containing a substantial number of structurally diverse actives. This meant that no artificial "seeding" of actives or selection of decoys was needed. The results should be taken as being illustrative

rather than definitive, however, confined as they are to a single biochemical target whose natural ligand is fairly flat and rigid, and involving as they do only a restricted range of parameter variation. Preliminary experiments indicate that consolidating some of the bins used at longer distances, for example, can enhance the recovery of actives. Once the full effects of bin size, differential weighting of features, and the intensity of conformational sampling on performance have been characterized, it should be possible to extend studies described here to other systems to determine the general advantages and disadvantages of full flexibility *vis à vis* fixed torsional intervals and of triplets *vis à vis* quartets and other more complex multiplets, among other questions of general interest in the pharmacophore community.

The primary goal here was to develop an infrastructure capable of storing pharmacophore multiplets compactly enough and manipulating them quickly enough to make large-scale analyses practical. The approach described here represents a substantial realization of those goals. Moreover, as the database searching example makes clear, the system is flexible enough to be a powerful tool for characterizing the behavior of pharmacophore multiplets in general. The speed and memory requirements are such that quite thorough sampling (in the thousands to tens of thousands) is practical enough to support library characterization and design as well as 3D database searching.

Research is currently underway on optimization of bin sizes, feature weights, and conformational sampling in the hope of improving performance further and illuminating the advantages (or lack thereof) of fully flexible conformational sampling.

ACKNOWLEDGMENT

The authors would like to acknowledge the helpful suggestions provided by the anonymous reviewers, especially their provision of a few particularly cogent references. We would also like to thank the organizers of the Chemical Structure Association Trust meeting at Noordwijkerhout for giving us the opportunity to present this material.

Supporting Information Available: Pharmacophore macro definitions from SYBYL 6.8.1. The copyright for the Supporting Information is held by Tripos, Inc. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(2) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(3) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for lead Discovery. *J. Biomolec. Screen.* **1996**, *1*, 65−73.

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

(5) 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.

(6) 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.

(7) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.

(8) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(9) Mason, J. S.; Beno, B. R. Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: Simultaneous optimization and structure-based diversity. *J. Mol. Graphics Mod.* **2000**, *18*, 438−451.

(10) Eksterowicz, J. E.; Evensen, E.; Lemmen, C.; Brady, G. P.; Lanctot, J. K.; Bradley, E. K.; Saiah, E.; Robinson, L. A.; Grootenhuis, P. D. J.; Blaney, J. M. Coupling structure-based design with combinatorial chemistry: application of active site derived pharmacophores with informative library design. *J. Mol. Graph. Model.* **2002**, *20*, 469−477.

(11) Chen, X.; Rusinko, A., III; Young, S. S. Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(12) 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: New York, NY, 2001; pp 399−428.

(13) 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.

(14) SYBYL, SYBYL Selector, and UNITY are distributed by Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

(15) Good, A. C.; Lewis, R. A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926−3936.

(16) 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.

(17) 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.

(18) CONCORD was developed by R. S. Pearlman, A. Rusinko, J. M. Skell, and R. Baducci at the University of Texas, Austin, and is distributed exclusively by Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

(19) CONFORT was developed by R. S. Pearlman and R. Balducci at the University of Texas, Austin, and is distributed by Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

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

(21) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(22) Addition of a special "null" feature to the augmentation list acts as a sort of "wild card" and overrides such exclusion.

(23) The Dictionary of Pharmacological Agents is compiled by Chapman and Hall/CRC, London. The searchable version used here was prepared and distributed by Tripos.

(24) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(25) Willett, P.; Winterman, V. A Comparison of Some Measures of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.

(26) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(27) National Cancer Institute, Bethesda MD.; http://dtp.nci.nih.gov/.

(28) The number of distance bins used was reduced to 12 in this case specifically so that 32-bit versions could be generated.

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

(30) Nærum, L.; Nørskov-Lauritsen, L.; Olesen, P. H. Scaffold Hopping and Optimization towards Libraries of Glycogen Kinase-3 Inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1525−1528.

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

(32) 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.

(33) Jørgensen, A. S.; Jacobsen, P.; Christiansen, L. B.; Bury, P. S.; Kanstrup, A.; Thorpe, S. M.; Bain, S.; Nærum, L.; Wassermann, K. Synthesis and Pharmacology of a Novel Pyrrolo[2,1,5-*cd*]indolizine (NNC 45-0095), a High Affinity Non-Steroidal Agonist for the Estrogen Receptor. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 399−402.

(34) Jørgensen, A. S.; Jacobsen, P.; Christiansen, L. B.; Bury, P. S.; Kanstrup, A.; Thorpe, S. M.; Bain, S.; Nærum, L.; Wassermann, K. Synthesis and Estrogen Receptor Affinities of Novel Pyrrolo[2,1,5-*cd*]indolizine Derivatives. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2383−2386.

(35) Bury, P. S.; Christiansen, L. B.; Jacobsen, P.; Jørgensen, A. S.; Kanstrup, A.; Nærum, L.; Bain, S.; Fledelius, C.; Gissel, B.; Hansen, B. S.; Korsgaard, N.; Thorpe, S. M.; Wassermann, K. Synthesis and Pharmacological Evaluation of Novel *cis*-3,4-diaryl-hydroxychromanes as High-Affinity Partial Agonists for the Estrogen Receptor. *Bioorg. Med. Chem.* **2001**, *10*, 125−145.

(36) Christiansen, L. B.; Wenckens, P. S.; Bury, P. S.; Gissel, B.; Hansen, B. S.; Thorpe, S. M.; Jacobsen, P.; Kanstrup, A.; Jørgensen, A. S.; Nærum, L.; Wassermann, K. Synthesis and Biological Evaluation of Novel Thio-Substituted Chromanes as High-Affinity Partial Agonists for the Estrogen Receptor. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 17−19.

CI025595R