ELSEVIER

Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: Simultaneous optimization and structure-based diversity

Jonathan S. Mason and Brett R. Beno

Computer-Assisted Drug Design, Structural Biology & Modeling, Department of Macromolecular Structure and Biopharmaceuticals, Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, NJ, USA, and Wallingford, CT, USA

New applications of fingerprints of multiple potential 4-point three-dimensional (3D) pharmacophores in combinatorial library design and virtual screening are presented. Preliminary results demonstrating the feasibility of a simulated annealing process for combinatorial reagent selection that concurrently optimizes product diversity in BCUT chemistry space and in terms of unique 4-point pharmacophores are discussed, and the advantage of using a customized chemistry-space derived for the library design is demonstrated. In addition, an extension to the multiple pharmacophore method for structure-based design that uses the shape of the target site as an additional constraint is presented. This development enables the docking process to be quantified in terms of the number and identities of the pharmacophoric hypotheses that can be matched by a compound or a library of compounds. The design of an example combinatorial library based on the Ugi condensation reaction and a serine protease active site is described. © 2000 by Elsevier Science Inc.

Keywords: pharmacophore fingerprints, BCUT *descriptors, library design, structure-based diversity, simulated annealing optimization, protein-site fingerprints, Chem-X, DiR*

Color Plate for this article is on page 538.

Corresponding author: Jonathan S. Mason, Structural Biology & Modeling, Department of Macromolecular Structure and Biopharmaceuticals, Bristol-Myers Squibb Pharmaceutical Research Institute, P.O. Box 4000, Princeton, NJ 08543, USA. Tel.: 609-252-4586; fax: 609-252-6030. *E-mail address*: jonathan.mason@bms.com (J.S. Mason).

INTRODUCTION

Methods for evaluating molecular similarity and diversity that are based on properties relevant to drug-receptor interactions and can be utilized for both ligands and receptors are central to many computer-assisted drug design (CADD) applications. These include virtual screening and combinatorial library design. For the latter, product-focused methods provide powerful capabilities and are often preferable to reagent-based approaches.[1] Product diversity is a key consideration in reagent selection for combinatorial libraries intended for screening-deck enhancement. In the absence of additional target-related information, a diverse set of N "drug-like" compounds has a larger likelihood of yielding hits than a set of N similar compounds. If information on active compounds is available, then the same metrics used to evaluate diversity can be used to quantify similarity. Currently, there are many different methods for optimizing diversity in the reagent selection process, which vary in the metric(s) used to evaluate diversity, the type of optimization algorithm used, and whether reagent or combinatorial product diversity is considered.[2] As each approach has characteristic strengths and weaknesses, it is advantageous to include multiple complementary methods when selecting reagents.

Two relatively new approaches for evaluating molecular diversity, DiverseSolutions (DVS)[3] BCUT metrics[4–6] and 4-point pharmacophore fingerprint analysis,[7–9] are attractive and potentially complementary, as they both incorporate information relevant to drug-receptor interactions, and both include shape information to at least a small extent. These descriptors can be calculated for whole products and used for virtual screening and to design libraries. Relevant BCUT chemistry-

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00073-5

space descriptors can be identified and rapidly calculated for very large libraries (>1 million products) using modest hardware, whereas the 3D pharmacophore fingerprints require substantially more calculation time (1–5 seconds per product) and for pragmatic reasons, have been used for smaller libraries where values for all products were calculated (<500,000 products). This article presents an optimization approach that calculates 4-point pharmacophores for potential combinatorial products "on the fly." The pharmacophore fingerprints additionally can be calculated complementary to a protein binding site and then used together with the ligand (product) fingerprints, with or without the binding site as a steric constraint, for virtual screening, library design, and docking (when using the site constraint). This article presents results from a method that includes the binding site as an additional constraint, producing fingerprints that can be used in the optimization procedure described, together with the BCUT metrics.

DVS BCUT metrics contain information regarding molecular connectivity/geometry and atomic properties such as hydrogen-bond donor/acceptor ability or polarizability that are known to be important for ligand-receptor binding interactions.[4–6] Typically, a small set (4–6 dimensions) of BCUT metrics that best describes the diversity of a set of compounds is selected and used to define a BCUT chemistry space. The BCUT chemistry space is then partitioned into cells, which allows a number of diversity-related tasks, including "void" identification, to be performed. In the context of reagent selection for combinatorial libraries, reagents are selected that, in combination, yield products that occupy a desired diversity/similarity distribution, such as the maximum possible number of BCUT chemistry-space cells.

BCUT chemistry-space descriptors are calculated from ligands only, whereas 3D pharmacophoric fingerprints can be calculated from both ligands and complementary to protein sites. The fingerprint is based on the 4-point pharmacophore methodology developed by Mason and coworkers.[7–9] A 4-point pharmacophore can be envisioned as a (possibly) irregular tetrahedron formed by connecting four different "features" [H-bond donors (D), H-bond acceptors (A), aromatic ring centroids (R), lipophilic regions (L), acidic atoms (C), basic nitrogen atoms (B)] found within a molecule. An individual 4-point pharmacophore is characterized by the identities of the vertices of the tetrahedron, and the set of intervertex distances (and an additional value if chirality is to be resolved). To reduce the number of possible 4-point pharmacophores to a manageable value and to reduce errors resulting from insufficient conformational sampling, interfeature distances are assigned to one of 7–10 bins depending on the ranges in which their values lie. Four-point pharmacophores generally are calculated for a large number of conformations per molecule and a composite fingerprint for all the valid conformations is used.

Both methods incorporate partitioning schemes and thus have the advantage of providing common frames of reference for rapidly comparing individual compounds, sets of compounds, and, in the case of pharmacophore fingerprints, ligands and protein receptor sites. As part of our efforts to extend in-house combinatorial library reagent selection technology, we explored the potential of simultaneously optimizing BCUT and 4-point pharmacophore diversity. Our objectives were to determine if both could be optimized simultaneously and to assess the feasibility of applying the procedure to reagent selection for large virtual libraries.

A simulated annealing algorithm[10] was selected to combine both components in a single optimization procedure. Simplicity and ease of implementation were the key factors affecting the choice, and multiple components can easily be included in the objective function using simulated annealing methodology.[11] In addition, simulated annealing has been used to perform reagent selection for combinatorial libraries based on 3-point pharmacophores[12] and other metrics.[11,13–15] Tests were run using a fully enumerated virtual library of 86,140 amide compounds constructed from carboxylic acids and primary amines present in the Available Chemicals Directory (ACD).[16]

Using average nearest-neighbor distances and Hopkins' statistic,[17–19] which evaluates the degree of clustering in a data set, the products of the optimized reagent sets are compared to the products of the original reagent sets. The diversity of the different product sets also is compared in terms of 4-point pharmacophore diversity.

The 3D pharmacophore fingerprints used in the optimization also could use pharmacophoric fingerprint information calculated complementary to protein sites.[7–9] The utility of comparing ligand fingerprints to protein site fingerprints to resolve selectivity has been reported,[7–8] and the resultant intersection fingerprint could be used in the optimization. In this article, we describe a new method that further refines this process, producing for a molecule a "fingerprint" of all the pharmacophores in a protein site that can be matched by docking the molecule into the site. Multiple docking orientations are evaluated systematically.

This new software that includes the shape of the target site in the pharmacophore-based analysis method is known as "Design in Receptor" (DiR).[9,20,21] The method provides new possibilities for docking, structure-based virtual screening, and library design through a novel quantification of target-based diversity, based on the systematically generated site-derived pharmacophore hypotheses. New modifications to the method have made it more effective for virtual screening and library design, and an example of its application in combinatorial library design is presented.

METHODS

BCUT Descriptors

Pearlman and Smith[4–6] introduced the use of "BCUT" metrics in a low-dimensional chemistry space using the DVS software.[3] A BCUT descriptor distills a large amount of information into a single number. This method utilizes four classes of matrices in which the diagonal matrix elements are based on computed physicochemical parameters related to ligand-receptor binding, including atomic charge, atomic polarizability, and atomic hydrogen-bond donor and acceptor abilities. The off-diagonal matrix elements are composed of topological information including 2D connectivity and/or interatomic distances, and scaling factors based on quantities such as exposed surface area are incorporated. The actual BCUT descriptors are the highest or lowest eigenvalues of these matrices for individual compounds. For example, a 3D charge BCUT for a particular compound could be defined as the highest or lowest eigenvalue of a matrix in which Gasteiger-Hückel charges for the individual atoms are the diagonal elements, and inverse distances (based on a single CONCORD[22] generated conformation) between the individual atoms form the off-diagonal elements. Generally, a large set of different BCUTs is calculated for a population of compounds, and a smaller subset that best defines the diversity of the population is selected to define the axes of a low-dimensional BCUT chemistry space.

The DVS software uses powerful cell-based methods to derive a chemistry space that optimally represents the diversity of a set of molecules. It also can rapidly choose diverse subsets or compare large datasets. Pearlman and Smith[4–6] introduced the use of a chi-squared-based "auto-choose" algorithm to accomplish the task of identifying a chemistry space tailored to a given population. This is particularly important with a focused population such as a combinatorial library, where diversity may be limited. Generally 4–6 BCUTs that provide maximum separation of the target population are identified. These BCUTs then serve as the axes of a low-dimensional chemistry space that subsequently is "binned" to produce cells. For example, a six-dimensional chemistry space divided into six bins per dimension contains 46,656 cells.

This method has been used for the design and analysis of large combinatorial libraries,[23,24] and several groups recently reported that subsets of BCUT metrics are able to cluster actives.[4,24] Pearlman and Smith[4–6] first presented this powerful concept of "receptor-relevant" subspaces, introducing a simple yet novel algorithm into DVS. This algorithm identifies a reduced set of dimensions (BCUT metrics) that conveys information relevant to receptor affinity. For example, angiotensin-converting enzyme inhibitors cluster in three of the six dimensions identified as important for all drugs (using the MDDR database[16]). Pearlman emphasized the importance of excluding metrics that are not "receptor-relevant" for a particular activity, to enable diversity in the "receptor-relevant" dimensions to be explored.

An important component of the DVS technology is nonlinear "binning" developed by Menard et al.[23] In this scheme, the axes of chemistry spaces are scaled, allowing inclusion of all structures within a set, while maintaining a reasonable distribution of compounds within cells. Nonlinear "binning" was subsequently implemented in DVS (from release 3.1.0).

Four-Point Pharmacophore Fingerprints

The ChemDiverse[25] module of the Chem-X[26] software calculates 3D pharmacophore fingerprints[7–9,27] consisting of multiple potential 3- and 4-point pharmacophores, systematically calculated, and incorporating conformational flexibility. For ligands, six pharmacophoric features that are likely to be important for drug-receptor interactions are automatically identified for each molecule through the use of atom types [for hydrogen bond donors, hydrogen bond acceptors, acidic centers (negatively charged at physiological pH 7), and basic centers (positively charged at pH 7)] and the addition of dummy atoms for hydrophobic regions and aromatic rings. For a protein site, complementary site points with associated pharmacophoric features are generated and the fingerprint is produced from these. All combinations of four pharmacophoric features are considered, together with 7 or 10 distance ranges[7] for each of the six distances, as illustrated in Figure 1.

A pharmacophore "fingerprint" indicates the presence or absence of all the theoretically possible combinations of features and distances (potential pharmacophores) and a chirality indicator can be added to applicable potential pharmacophores. About 2.3 million (7 distance ranges) or 9.7 million (10 distance ranges) 4-point potential pharmacophores are considered for each ligand or set of site points. The total number considered is less than the theoretical number of combinations because certain distance combinations are geometrically not possible.

Effective conformational sampling is required to create pharmacophore fingerprints for ligands. The method used is a customization of the Chem-X/ChemDiverse approach and incorporates "on-the-fly" generation of conformers with rapid evaluation of each conformation based on a steric contact check that rejects poor or invalid conformations. Using up to four rotamers per bond, the maximum analysis time is only 5 (random) to 15 (systematic) seconds on an R10000 250-MHz processor. The conformer generation process employs a systematic analysis, where possible, and a random analysis for very flexible molecules. Only the internal bump check is used in our implementation to eliminate unreasonable conformations, although additional rule-based tests or energy calculations could be applied. This procedure generates an information-rich fingerprint that includes a union of all the potential pharmacophores for sterically accessible conformations. Potentially important information about flexibility is encoded. Two molecules with similar functional groups connected by moieties of differing flexibility (one rigid and one flexible) will have different fingerprints (one small and one larger). However, if only a single low-energy conformation of each molecule is considered, the two molecules could have the same small fingerprint.

Pharmacophore fingerprints can be precalculated and stored in an efficient (compact) format and then searched very rapidly (S.J. Cho and J.S. Mason, unpublished results). A space-

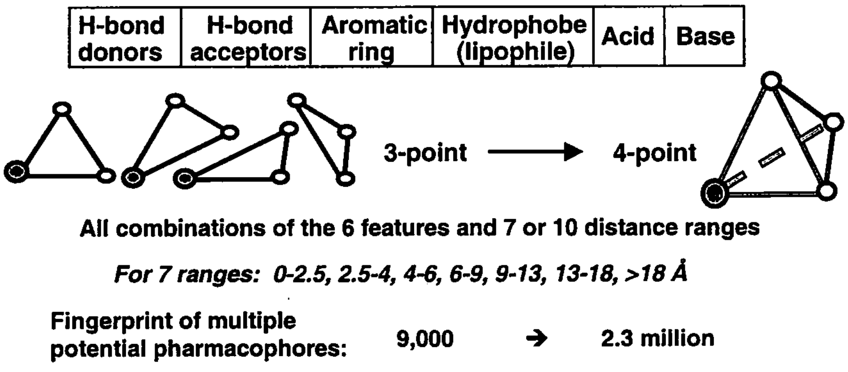

H-bond donors
H-bond acceptors
Aromatic ring
Hydrophobe (lipophile)
Acid
Base
3-point
4-point
All combinations of the 6 features and 7 or 10 distance ranges
For 7 ranges: 0-2.5, 2.5-4, 4-6, 6-9, 9-13, 13-18, >18 Å
Fingerprint of multiple potential pharmacophores:
9,000
2.3 million

Figure 1. Definition of multiple potential 3D pharmacophore fingerprints.

efficient format of one line of encoded information per compound is used (11 kB for 1,000 pharmacophores, compared to 58 kB ASCII and 1.4 MB binary formats in Chem-X). On a Silicon Graphics workstation (single R10000, 250-MHz CPU), it is possible to search more than 700K compounds per hour and extract the potential pharmacophores that each compound in the database has in common with the query compound. Fingerprint storage requirements and search times recently have been reduced further (B.R. Beno, unpublished results).

Relative to 2- and 3-point pharmacophores, significant increases in the amounts of shape information and resolution have been achieved using 4-point pharmacophores. This extension provides the ability to distinguish chirality, a fundamental requirement for many ligand-receptor interactions.[8,9] Previous studies[7–9,27] showed that the increased information (including shape and chirality) present in the tetrahedral 4-point pharmacophoric descriptions [compared to 2-point (distance) or 3-point (triangle) descriptions] normally is needed for molecular similarity studies (ligand-ligand, ligand-protein) that use only the fingerprint. The pharmacophore fingerprint method has an advantage over many other 2D and 3D similarity methods, because flexible and information-rich compounds such as small peptides can be used as input.

For an enzyme active site or a receptor site, complementary site points are added (as atoms, dummy atoms, or functional groups), and relevant pharmacophoric features are assigned to these points. The site points can be generated by geometric methods (as implemented in Chem-X/ChemProtein) or via energetic surveys of the site, using a variety of probe atoms (as implemented in the GRID program[28]). For example, a dummy atom site point that is assigned the hydrogen bond acceptor feature may be placed within hydrogen bonding distance of a sterically accessible N-H group within the receptor. The combined set of all site points represents a hypothetical molecule that is capable of binding to all available positions, and the fingerprint of potential pharmacophores is generated for this "molecule" in the same way as for a normal compound.

Pharmacophore fingerprints allow rapid comparisons of different ligands, ligands and protein sites, or different protein sites. Potential 3D pharmacophores that are shared, as well as those that are only present in one compound or protein site, can be readily identified. As all combinations of features and distance ranges between them are considered, the method provides a measure of diversity for both ligands and protein sites. Studies can be designed to identify a set of compounds that together explore all the pharmacophore hypotheses in the query molecule. This is preferable to other methods where a large flexible query molecule is used to select other equally large flexible molecules.

For virtual screening, the precalculated 3D pharmacophore fingerprints can be searched rapidly. This method has an advantage over many 3D methods in that it is not necessary to define a single hypothesis or conformation. An ensemble of pharmacophore hypotheses can be searched simultaneously. The input can be either a single ligand, an ensemble of ligands, a small peptide, or a protein binding site. Virtual screens for selecting a small number of compounds for focused biological screening have been successful in finding new hits in all four cases (unpublished in-house data).

For scoring and subset selection, both the total number of pharmacophores in each database compound and the number of pharmacophores each database compound has in common with a query compound can be used. Similarity indices, such as the Tanimoto coefficient, can be generated from these numbers.[7,8,29] The union of pharmacophore fingerprints for multiple active compounds also can be used as a search query. Using a Tanimoto coefficient to compute similarity, the fingerprint for a relatively flexible and promiscuous compound, such as a small peptide, can be used to select structures that are of comparable similarity to the query structure, but possess different pharmacophores in common with the query. In this case, the goal is to identify an ensemble of simple compounds that together express the largest possible number of the potential pharmacophores identified for the query compound. Each of the compounds in the ensemble thus contains a small number of pharmacophores in common with the query.

Another important application of the 3D pharmacophore fingerprint utilizes a user-defined feature type to force inclusion of a substructure of interest.[7,8] This creates a "relative" or "internally referenced" measure of diversity that has been extensively used to design combinatorial libraries that contain "privileged" substructures focused to 7-trans membrane G-protein coupled receptors.[7] Ligand-based diversity, centered around the privileged substructure of interest, is explored in this method.

As each "bit" of the pharmacophore fingerprint corresponds to an actual potential pharmacophore, the multiple potential 3D pharmacophore fingerprint technology provides a rapid method for 3D database searching (with multiple queries) and pharmacophore identification. It has been found to provide a complementary approach for molecular diversity applications to the DVS BCUT method (in house results and a published combined application of both for library subset selection[23]). Further details of all these methods have been published.[7–9,27,30]

Virtual Library Construction

For the DVS/pharmacophore concurrent optimization study, a small enumerated virtual library of 86,140 amide compounds was constructed from carboxylic acids and primary amines listed in the ACD.[16] Because this virtual library was intended solely for experiments involving diverse subset selection, no attempt was made to focus the set of reagents toward any particular biological target. The full set of available acids and amines in the ACD was filtered extensively using the set of SMARTS queries published by Hann and coworkers,[31] as well as several SLN filters routinely used in-house for reagent processing. Carboxylic acids containing amine or multiple acid moieties were removed, as were primary amines with carboxylic acid or multiple amine moieties. Duplicate compounds, and those with fewer than five or more than 60 heavy atoms, and metal-containing compounds were eliminated. After filtering, 5,219 carboxylic acids and 770 primary amines remained. From these, a set of 300 acids and a set of 300 primary amines were selected at random, and their structures were imported into SYBYL databases.[32] CombiLibMaker V4.2.2 (Combine Reactants mode)[33] was used to enumerate the virtual library of 90,000 compounds into 2D SLN strings, and CONCORD 4.0.2[22] was used to convert the virtual library SLN strings into 3D SDF format suitable for importing into Chem-X.[26] CONCORD errors occurred for products formed from 12 of the reagents. These were excluded from the set, and the final dimensions of the virtual library were 292 acids × 295 amines (86,140 products).

DVS Chemistry-Space Selection

The DVS 4.0.5 software package was used to calculate 3D hydrogen-suppressed BCUT values for the virtual library compounds and to select a BCUT chemistry space that optimally represented the diversity of the library. A maximum of eight chemistry-space dimensions was considered with the DVS "auto-choose" algorithm, and nonlinear scaling was applied to the metric values. Two charge, two polarizability, and one hydrogen-bond acceptor BCUTs (Figure 2) defined the resultant five-dimensional chemistry space. This chemistry space was partitioned into 10 bins per dimension, giving a total of 100,000 cells. The virtual library compounds occupied a total of 12,146 (12.1%) of the 100,000 cells in the library chemistry space. A list of compound names and single numbers representing the cell within the five-dimensional chemistry space that each compound occupied was exported from DVS to an ASCII file for use in the optimization procedure.

For comparison purposes, the cell occupancy of the virtual library compounds was also evaluated in a BCUT chemistry space derived from 3D hydrogen-suppressed BCUTs for a combined database of approximately 775,000 compounds from various sources, including the ACD, MDDR, and BMS databases. This chemistry space was six dimensional and was partitioned into seven bins per dimension to give a total of 117,649 cells. Of these, only 1,805 (1.5%) were occupied by the 86,140 library compounds. The implications of this for combinatorial library design purposes will be considered in the Discussion section.

Four-Point Pharmacophore Fingerprint Calculation for Library Design

Four-point pharmacophores for virtual library compounds were calculated as described earlier using the extensions to the Chem-X software recently reported by Mason et al.[7–9] Conformational sampling (reduced time of 4.5 seconds of systematic or 1.5 of random sampling using an R10000 250-MHz CPU) was performed for each compound. Although precalculating the pharmacophore keys for the entire virtual library was feasible in this case and would have dramatically decreased the time required for each iteration in the optimization procedure, it was decided to develop methodology to calculate pharmacophore keys "on the fly" to allow reagent selection from virtual libraries that are too large for practical precalculation of all pharmacophore keys.

A set of Perl and C-shell scripts and C programs were written to automate the process. To minimize the time required for each optimization step, pharmacophore keys were calculated in parallel during the optimization procedure as reagents were swapped into the active set. Routinely, 20 R10000 processors on a Silicon Graphics Origin 2000 were used. After the pharmacophore key for each compound was calculated, it was converted into a more compact encoded form for storage and processing as discussed earlier. Previously calculated keys were extracted from the saved set, rather than recalculated, during the course of the optimization.

Optimization Procedure

Due to its simplicity and ease of implementation, a simulated annealing (SA) algorithm was selected.[10–15] The overall procedure is implemented in a C program that runs on Silicon Graphics platforms. This optimizer is capable of including multiple components into the scoring function and controls the "on-the-fly" pharmacophore calculation process. For the optimization trials, an initial temperature factor (T) of 0.5 was held constant for the first 5% of the total requested iterations, and then decreased using the exponential function shown in Equation 1.

BCUT_Charge_1	bcut_gastchrg_S_invdist6_0.60_R_H
BCUT_Charge_2	bcut_gastchrg_S_invdist_3.00_R_L
BCUT_Hacceptor	bcut_haccept_S_invdist_0.60_R_H
BCUT_Polarizability_1	bcut_tabpolar_S_invdist2_1.00_R_L
BCUT_Polarizability_2	bcut_tabpolar_S_invdist_0.50_R_H

Each BCUT name describes the matrix and eigenvalue used to define the BCUT. For example:

bcut_gastchrg_S_invdist6_0.60_R_H

gastchrg_S	Products of atomic Gasteiger-Hückel charges and fractional surface area (S) on matrix diagonal
invdist6	Off-diagonal elements equal to inverse interatomic distances to the sixth power
0.60	Off-diagonal element scaling factor
R	Remove hydrogen atoms
H	BCUT metric is highest eigenvalue of matrix

tabpolar = tabulated atomic polarizability
haccept = atomic hydrogen-bond acceptor ability

Figure 2. BCUTS defining the 5-dimensional BCUT chemistry space identified for the virtual library.

$$T = 0.5 \cdot \exp(0.5 - [\mathit{current\_iteration}/(0.10 \cdot \mathit{total\_iterations})]) \quad (1)$$

At each iteration in the optimization procedure, an overall score for the currently selected set of products was calculated with using Equation 2:

$$\mathit{score} = \frac{O}{P} + \left(2 \cdot \frac{U_{pk}}{T_{pk}} + \frac{U_{pk}}{350000}\right) \quad (2)$$

The first term of the function shown in Equation 2 represents the diversity of a subset of products in the virtual library BCUT chemistry space as the ratio of filled cells (O) to total possible filled cells (P). The latter is the total number of combinatorial products in the subset (e.g., 400 for a product subset of amides created from 20 acid and 20 amine reagents). This component of the overall score is simple to calculate, but its use has an inherent danger. Because only the total number of occupied cells, and not the distribution of these cells within the chemistry space, is optimized, clustering of products could occur. A set of X products in Y different but tightly grouped cells would have the same BCUT chemistry-space diversity score as X products occupying Y cells distributed uniformly throughout the chemistry space.

The second term in the scoring function shown in Equation 2 evaluates 4-point pharmacophore (pharmacophore) diversity. This expression involves the total number of (potential) pharmacophores (T_{pk}) and the number of unique pharmacophores (U_{pk}) in the product subset. Here, U_{pk} is determined by counting the occurrence of an individual pharmacophore in a set of products only once, even if that pharmacophore is found in multiple products. The value for T_{pk} is determined by adding together the number of pharmacophores found in each product molecule in the set.

In a diverse set of molecules, the ideal ratio U_{pk}/T_{pk} as defined earlier should approach 1. A ratio of exactly 1 would indicate that no two compounds in the set had any pharmacophores in common. However, U_{pk} also should be as large as possible so that many different pharmacophores are represented in the set of molecules. Initial versions of the scoring function used the ratio U_{pk}/T_{pk} alone as the term for quantifying pharmacophore diversity. In practice, U_{pk}/T_{pk} could be substantially increased during the optimization. However, this was achieved by the selection of small, relatively rigid reagents, because these combined to produce products for which T_{pk} was small. As a result, an overall reduction in U_{pk} also was observed. This prompted the addition of an additional term ($U_{pk}/$350,000) to increase the value of U_{pk} during the optimization. The denominator term 350,000 was chosen by determining U_{pk} for the products of random selections of 20 × 20 reagent sets and selecting a value approximately three times larger than the largest U_{pk} value found. This number is strongly dependent on the size of the virtual library subset being optimized, as well as the structures of the products. In the final equation, the U_{pk}/T_{pk} term was scaled by a factor of 2.0. This scaling factor also was selected empirically based on the tendency of the $U_{pk}/$350,000 term to overwhelm the U_{pk}/T_{pk} term. Although this approach to scoring pharmacophore diversity is adequate for experimental purposes, additional work to develop a more general function is necessary.

Clustering Evaluation

Hopkins' statistic[17–19] was used as a simple way to evaluate the degree of clustering in the product sets. Shown in Equation 3, this compares the distances between actual data points and their nearest neighbors in a multidimensional space to the distances between hypothetical uniformly distributed points and their nearest-neighbor real data points. The hypothetical points are randomly selected within the bounds defined by the actual data, and the actual data points used in each comparison are a randomly selected subset of the full data set. In Equation 3, U_i values are the distances from the hypothetical points to the nearest actual data points, and W_i are the distances between the actual data points and their nearest neighbors. Hopkins' statistic values, H, close to 0.5 are indicative of uniform distribution of the data points, whereas values above 0.75 are strongly indicative of clustering within a data set. For this study, Hopkins' statistic was implemented in a C program. In practice,[18,19] Equation 3 is evaluated several times, and the results are averaged. For each data set examined, N was 40 (10% of the data set) and H was evaluated 10 times. Values averaged over the 10 evaluations (H_{av}) are reported in the text.

$$H = \frac{\sum_{i=1}^{N} U_i}{\left(\sum_{i=1}^{N} U_i + \sum_{i=1}^{N} W_i\right)} \quad (3)$$

Optimization Trials

Six separate optimization runs were performed. In each case, optimized sets of 20 acids and 20 amines (400 products) were selected. Optimizations were initiated from each of two random seed values (1,000 for trials A, C, and E, and 9,999 for trials B, D, and F) and continued for 2,500 iterations. Two of the optimization trials (A and B) included the combined BCUT/pharmacophore scoring function, two utilized only the BCUT cell-based diversity component (C and D), and two included only the pharmacophore term (E and F).

Protein-Site Fingerprints: Design in Receptor "DiR" Method

In contrast to pioneering methods such as DOCK for virtual screening and the extended version CombiDOCK[34] for combinatorial libraries, which are driven by the shape of the target site, with additional constraints possible for pharmacophoric features, the DiR method focuses on the pharmacophoric match and produces a "score" for a molecule that is a quantification of how much of the site-defined pharmacophoric diversity is matched. The systematic definition of complementary potential pharmacophores (defined from the position of complementary features such as those identified from energetic surveys of the site using probe atoms in the GRID program,[28] as discussed earlier) is used to quantify which pharmacophore hypotheses in the active site are matched for a particular ligand. The site is used as a shape constraint to reject any fits with bad steric contacts.

This new method thus enables the steric shape of a protein site to be used as an additional constraint in the comparison of pharmacophore fingerprints. DiR is included as a module of the 2000 release of the Chem-X software.[26] The DiR approach is equivalent to simultaneous 3D database searching using multiple 3D pharmacophoric queries and steric constraints, but requires only a single conformational sampling. A "fingerprint" for each molecule of protein site pharmacophores matched by any fitted conformation is output and can be used in the same way as, and combined with, normal molecule fingerprints. Complete pharmacophore fingerprints generated from the docked conformations of a ligand also can be output. Another possible output is a 3D database of structures fitted onto the site points. This can be used as input for other scoring methods and for force field minimization, etc.

The fingerprints output for each molecule of the matched site-derived pharmacophoric hypotheses enable the design of a screening set or a selection of reagents for a combinatorial library that optimize this matching, in terms of maximum coverage or of specific pharmacophores. The method thus can be used for designs that are enriched in a subset of site pharmacophores of interest. The "score" obtained for a ligand in a site does not attempt to quantify the potential interaction energy (this can be done separately using the saved fitted conformations), as with DOCK/CombiDOCK, but evaluates the number and identity of the pharmacophoric hypotheses that can be matched within the defined steric constraints. The requirement that the pharmacophoric match fits the shape of the target site clearly provides much additional information, and 2-, 3- and 4-point site potential pharmacophores can all be used to drive the process.

A novel measure of structure-based diversity is thus obtained. The pharmacophore fingerprint derived from the complementary site points quantifies the different pharmacophore hypotheses a ligand may match upon binding. In this way, it is possible to evaluate which ligands are able to fit in the site while matching at least one set of pharmacophoric features, and then to design a subset of ligands that match as many pharmacophoric hypotheses as possible. Additional constraints can be included to ensure that one or more groups of pharmacophore site points are included in each pharmacophore hypothesis used. This option enables any information about ligand binding modes to be used to bias the docking process.

Significant speed enhancements that make DiR practical for docking,[9] virtual screening, and library design now are incorporated in the software. These include the use of an atom-based bump check instead of grid-based surface maps and adaptive pharmacophores where a pharmacophore query is removed from consideration for a particular ligand after it has been matched once and the analysis for that structure is terminated if no more fits are possible. An example of this technology applied to combinatorial library design is reported later.

RESULTS AND DISCUSSION

Simultaneous Optimization of DVS BCUT Cell-Based Diversity and Four-Point Pharmacophore Diversity in Library Design

The progress of optimization trials A and B in terms of total score and the individual score components is shown in Figure 3. After 2,500 iterations, the total scores and individual components in each optimization reached plateaus indicative of minima. However, in only 2,500 iterations, it is unlikely that the global minimum was located during either trial. Modest increases in the individual score components occurred during the course of the optimizations, and each run converged to similar values in terms of total score and individual score components. However, the final reagent selections were quite different. The two trials selected only three acid and seven amine reagents in common.

In trials A and B, the products of the initial random selections of reagents occupied 327 and 324 cells in the BCUT chemistry space derived from the virtual library. The products of the optimized reagent sets from trials A and B filled 394 and 397 cells, respectively. This represents increases of 20%–23% over the initial values, and the final values were only 3–6 below the maximum possible number of 400, which was obtained when the optimization was done using only the BCUT diversity term in the scoring function (trials C and D).

Figure 4 shows the distribution of the products from the initial random reagent selection and the products of the optimized reagent sets from trials A and B in three two-dimensional subsets of the five-dimensional BCUT chemistry space. The products of the optimized reagents from each trial cover larger areas of the chemistry space than the products of the randomly selected reagents. This suggests that the subsets of products formed from the optimized reagents better express the diversity of the entire virtual library. However, neither the optimized products from trials A and B, nor the products of the randomly selected reagents, provide complete coverage of the area occupied by the virtual library, and there are areas covered by the products of the randomly selected reagents that are not covered by the optimized products.

As was pointed out by a reviewer, if active compounds are concentrated in areas of the chemistry space occupied by products of the randomly selected reagents but not occupied by products of the optimized reagent set, then the random selection is preferable. However, in this case, no activity information is available, and the objective was simply to choose a subset from the virtual library that best expresses the diversity of the whole virtual library. The scatter plots shown in Figure 4 suggest that the optimization procedure results in better coverage than the random selections.

Two additional metrics used to compare the products of the random and optimized reagent sets are average nearest-neighbor distances and Hopkins' statistic,[17–19] which evaluates the degree of clustering in a data set. The average nearest-neighbor distances in the five-dimensional DVS BCUT chemistry space were 0.79 and 0.81 for the products of the initial randomly selected reagents sets in trials A and B, respectively. Both optimization trials resulted in increases, and the final values were 1.18 and 1.13 for A and B, respectively. This represents an improvement in diversity relative to the products of the randomly selected reagents.

The average values of the Hopkins' statistic[17–19] (H_{av}) calculated for the products of the two sets of randomly selected reagents were 0.74 and 0.75, strongly indicative of clustering within the BCUT chemistry space. The optimized product sets from trials A and B had H_{av} values of 0.65 and 0.69. The reduction in H_{av} relative to the values obtained for the initial random selections represents a slight improvement. However, in both cases, H_{av} was larger than the value of 0.5 expected for uniform distributions, and some clustering within the chemistry space is likely.

To assess the effect of the pharmacophore term in the scoring function on the overall DVS BCUT diversity of the optimized products, two optimizations (trials C and D) were

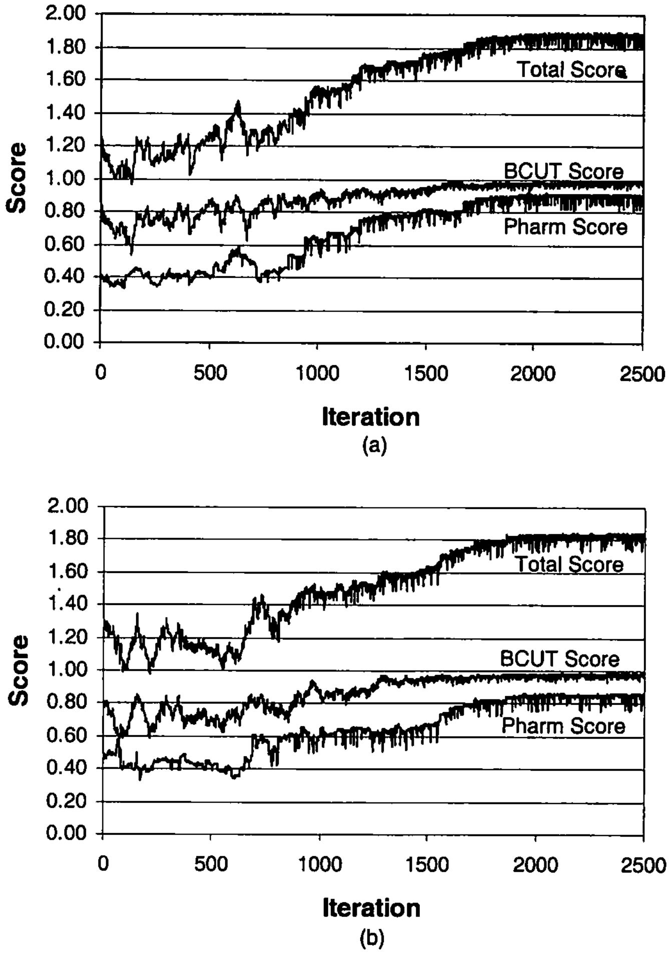


Figure 3. Plot of optimization progress for simulated annealing optimizations incorporating both BCUT and 4-point pharmacophore diversity into the scoring function. (a) Trial A. (b) Trial B.

performed using only the cell-based term in the scoring function. In each case, the products of the optimized reagents occupied 400 out of a possible 400 chemistry-space cells. The average nearest-neighbor distances were 1.23 and 1.15 for trials C and D, and the H_{av} values were both 0.67. Overall, this represents only a small improvement over the results from trials A and B, suggesting that inclusion of the pharmacophore scoring term does not strongly limit the BCUT chemistry-space diversity of the products of the optimized reagent sets.

The simple cell-based BCUT diversity metric used in combination with the optimization procedure appears capable of selecting reagents that, when combined, produce a more diverse set of products than do sets of reagents selected at random. It is interesting to note that the optimization trials in which only the pharmacophore score was used (trials E and F) selected sets of reagents that would yield products occupying only 291 and 276 cells in the BCUT chemistry space, respectively. Average nearest-neighbor distances also were reduced relative to the initial random selections, and the H_{av} values were significantly larger: 0.80 and 0.78 for trials E and F, respectively. Overall, this represents a substantial decrease in BCUT chemistry-space diversity relative to the initial random sets and suggests that the BCUT diversity term in the scoring function is important. Diversity in the BCUT chemistry space did not increase simply as a result of increasing pharmacophore diversity during the optimizations that included both terms in the scoring function.

Before presenting the effects of the optimization procedure on pharmacophore diversity, the importance of selecting a BCUT chemistry space in which a library is designed should be mentioned. Ideally, a "universal" BCUT chemistry space could be used, as it would allow for facile comparison of different combinatorial libraries.[23] However, if product diversity is of overriding importance, a BCUT chemistry space derived for the virtual library of interest should be used because the axes (BCUT metrics) of such a space are selected specifically to maximize the spread of compounds in the virtual library.[4–6] This is demonstrated by the distribution of virtual library compounds in the BCUT chemistry space derived for the library and the "universal" BCUT chemistry space derived for a combined database of ~750,000 compounds. In the library-specific space, the 86,140 products were distributed in 12.1% of the total cells, whereas in the "universal" chemistry space, these same products were concentrated in only 1.5% of the total cells.

The pharmacophore component of the scoring function used in the combined BCUT/pharmacophore diversity optimization trials includes two terms, U_{pk}/T_{pk} and $U_{pk}/350000$. The first attempts to bias reagent selection away from large, highly flexible molecules, whereas the second is designed to increase the total number of different pharmacophores represented in the products of the optimized reagent set. Due to the way in which U_{pk} and T_{pk} are included in the scoring function, optimizations that converge to very similar total pharmacophore

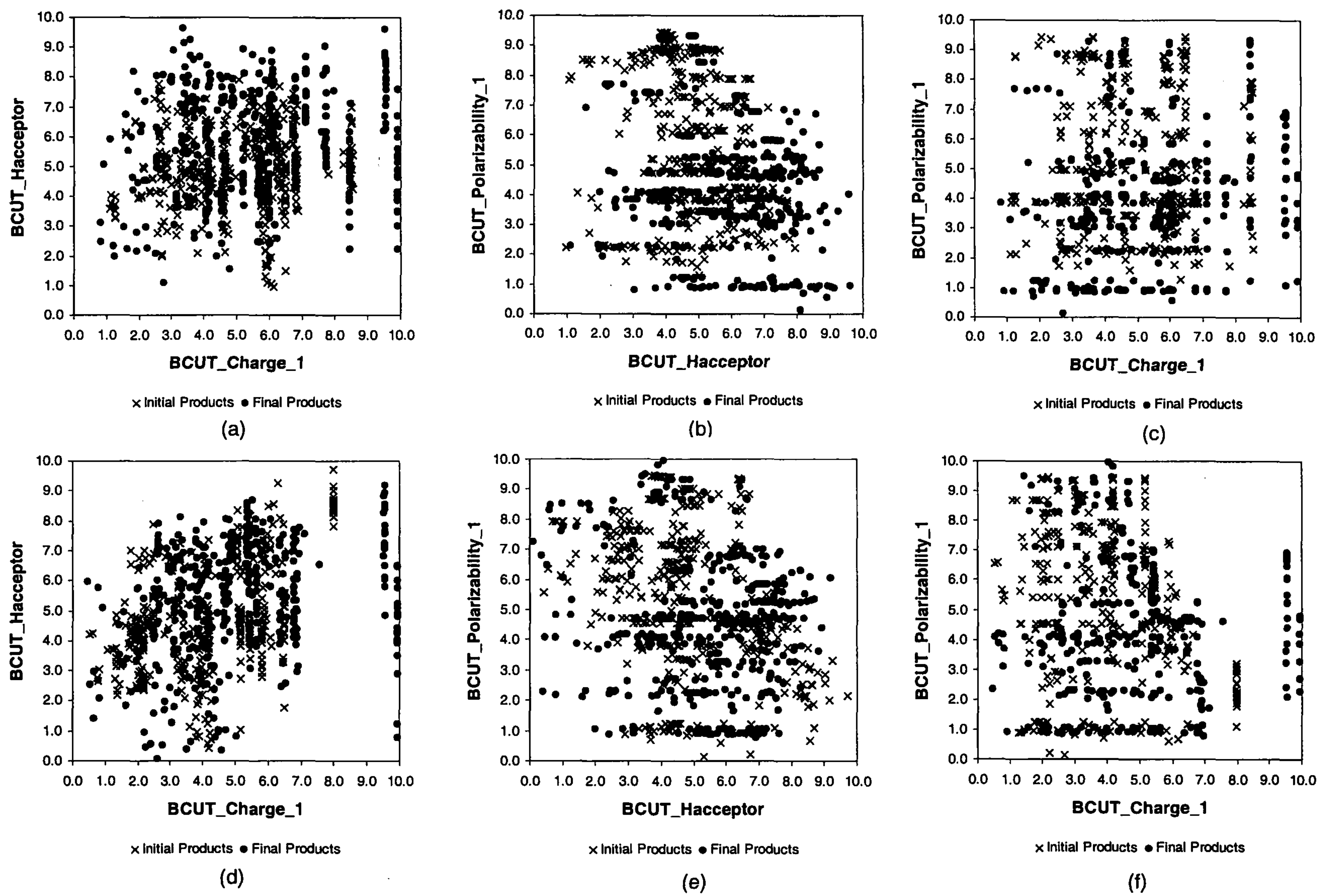


Figure 4. Distribution of combinatorial library products in two-dimensional subsets of the five-dimensional DVS BCUT chemistry space derived for the virtual library. Products of the optimized reagent sets are shown as black circles, and products of the original randomly selected reagents are denoted by black "X" symbols. Results from trial A (a, b, c) and trial B (d, e, f) are shown.

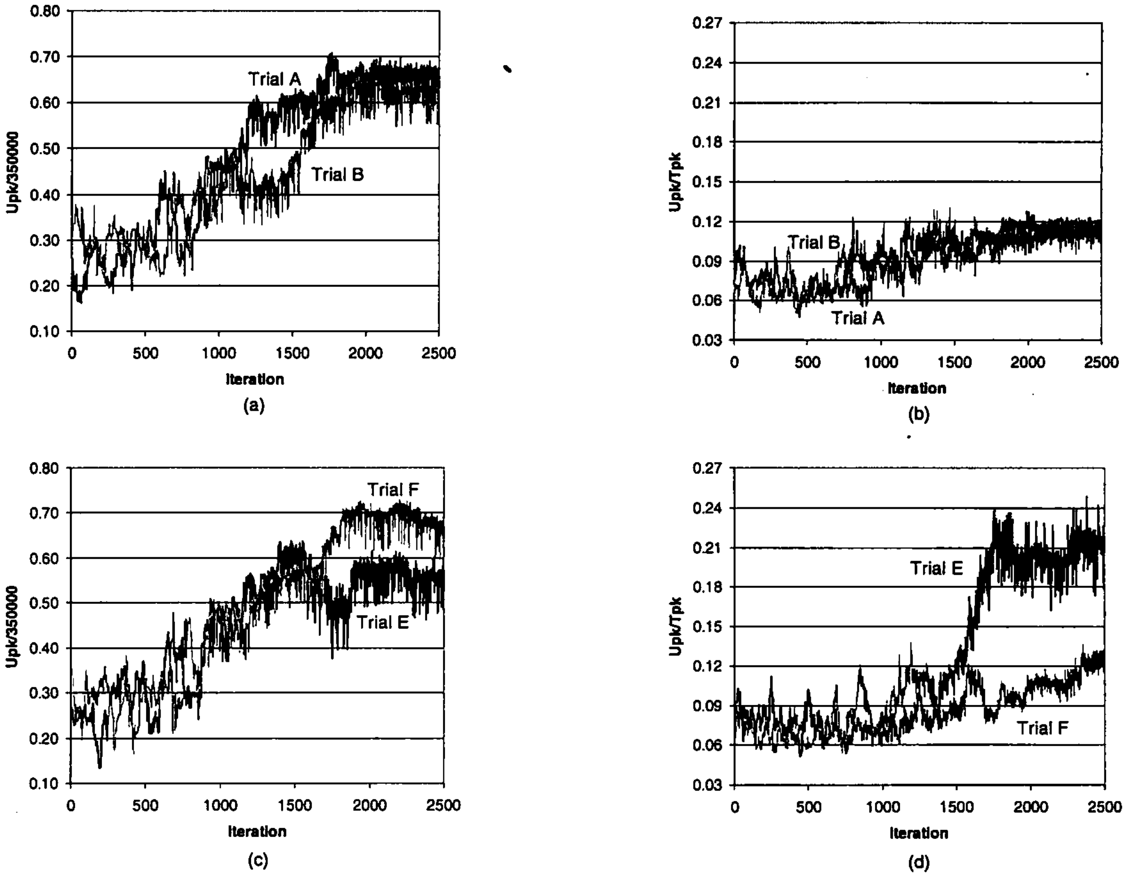


Figure 5. Changes in 4-point pharmacophore diversity terms during the optimizations for trials A and B (pharmacophore and DVS terms; a and b) and trials E and F (pharmacophore terms only; c and d).

scores can have significantly different final values for U_{pk} and T_{pk}.

Figure 5a shows the changes in $U_{pk}/350000$ for trials A and B, where U_{pk} is the number of unique pharmacophores found in the products of the reagents in the active set at iteration N. Substantial improvements in the number of unique pharmacophores were observed for all four optimizations trials.

Figure 5b show the changes in the values of U_{pk}/T_{pk} during the course optimization trials A and B. Also shown in Figures 5c and 5d are the results for trials E and F, which included only the pharmacophore term in the scoring function. All of the optimizations produced increases in the U_{pk}/T_{pk} term. The ratios of unique to total (U_{pk}/T_{pk}) pharmacophores for the initial product sets were 0.069 and 0.072 for trials A and B, respectively. After optimization, these values had increased to 0.117 and 0.112. Importantly, in trial A, U_{pk} increased by a factor of 2.55 whereas T_{pk} increased by a factor of only 1.52. In trial B, U_{pk} increased by a factor of 1.80, whereas T_{pk} increased by a factor of only 1.16.

Both combined BCUT/pharmacophore optimizations resulted in significant improvements in U_{pk}/T_{pk}. This clearly demonstrates that the simulated annealing approach combined with the pharmacophore diversity scoring function used can select sets of reagents yielding products that possess much larger numbers of unique pharmacophores than the product sets formed from randomly selected reagents. However, the largest increase U_{pk}/T_{pk} was observed for trial E, which included only the pharmacophore term in the scoring function. In this case, U_{pk} increased by a factor of 2.14 whereas T_{pk} *decreased* by 31%. Because the large increase in U_{pk}/T_{pk} was not observed for trial F, which also lacked the BCUT scoring term, this does not necessarily imply that BCUT and pharmacophore diversity are inversely correlated. However, the observation that the optimization procedure can converge to drastically different minima in the same number of iterations depending on the starting state suggests that improvements in the pharmacophore scoring term and possibly the optimization cooling schedule are necessary.

Is this approach practical for reagent selection for large combinatorial libraries? The most time-consuming stage of the optimization procedure is the "on-the-fly" calculation of the pharmacophores. When BCUT chemistry-space diversity is the only component of the scoring function, a 2,500-iteration optimization of a 20 × 20 reagent set requires <1 minute of CPU time on a Silicon Graphics R10000 machine. When both scoring components are included, the optimization requires approximately 16 hours of elapsed time and significantly more CPU time, because pharmacophores are calculated in parallel. This is a substantial time commitment, and a faster procedure is desirable.

The obvious solution is precalculation of pharmacophores for the entire virtual library. Good and coworkers[12] adopted this approach with their HARPick program, which used 3-point pharmacophores for combinatorial reagent selection. For small

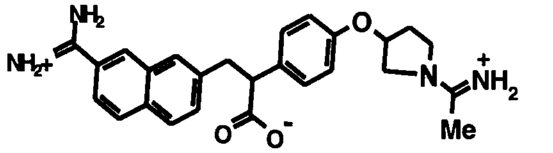


Figure 6. Daichii factor Xa ligand used for DiR validation studies.

virtual libraries (~250,000 combinatorial products), this is probably the best method. However, it is impractical for reagent selection from virtual libraries composed of 1,000,000 or more compounds.

We currently are in the process of optimizing our "on-the-fly" pharmacophore calculation procedure, which involves several C programs, C shell, and Perl scripts that convert the pharmacophores from the ASCII format output by Chem-X, to a smaller more manageable format, and store the encoded pharmacophores in a random-access database. This process could be greatly simplified with a single C program using the ChemLib interface to Chem-X that writes the pharmacophores directly in our modified format. In addition, we recently made significant improvements in our techniques for searching databases of encoded pharmacophores. The optimization procedure has not yet been modified to reflect these improvements.

Further gains could be achieved by reducing the conformational search time allotted to each combinatorial product molecule. Currently, either a maximum of 4.5 seconds of systematic searching or 1.5 seconds of random searching is performed, with a preliminary very short (<1 second) systematic search and results analysis performed for compounds that fall between the easily recognized classes of not too flexible (= systematic) and very flexible (= random). By performing only 1–2 seconds of random or systematic sampling and eliminating the preliminary systematic searching step, an additional decrease in the time required for each iteration could be realized.

Optimization of Protein-Site Pharmacophore Diversity in Combinatorial Library Design

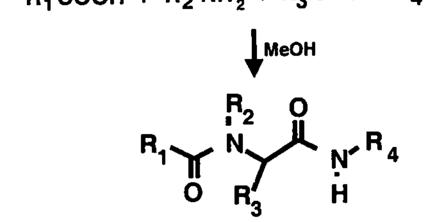
The extension of the pharmacophore fingerprints to protein binding sites was investigated using the DiR software.[20,21,26] An example of the DiR approach applied to combinatorial library design is presented later. The special fingerprints obtained were analyzed in a semimanual fashion in order to illustrate the type of information that can be extracted. They could be used independently or together with the ligand-based BCUT metrics in a simultaneous optimization as described earlier for combinatorial library designs in which the protein binding site is available.

This study used 4-point pharmacophores based on 23 complementary site points added to the factor Xa serine protease crystal structure.[35] These site points were selected based on GRID analyses,[28] using probes for hydrogen bond donors, acceptors, bases, acids, and hydrophobes. Pharmacophore feature points were added to regions identified by the GRID analyses, as previously described.[7] An extended set of 23 points was used for this work: 4 hydrogen-bond donors, 5 hydrogen bond acceptors, 4 basic, 2 acidic, 4 hydrophobic, duplicated as aromatic ring centroids. One hundred sixty-two atoms from the active site defined the shape and were used in a bump check [CPK (2/3 VDW) radii, maximum three bumps] to eliminate bad steric fits. A maximum number of 10,000 substructure matches and 300 hits per structure were allowed, with a tolerance of 1.5Å for the fitting of matching conformations to each site-point pharmacophore hypothesis.

To ensure that the docked structures, and thus the resultant fingerprints, were relevant, the site pharmacophores were forced to contain one S1 pocket hydrophobe/aromatic ring centroid point (from a choice of 3+3) and the S4 pocket hydrophobe or aromatic ring centroid point. An analysis of known trypsin-like serine protease x-ray structures of ligand complexes was performed to derive this restriction. This focused approach ensures that a "diversity" of matched site pharmacophores represents a realistic and desirable goal, of "reasonable" binding modes related to those experimentally found and thus expected to have a higher probability to give rise to biological activity. This restriction reduces the total number of site pharmacophores from 5,393 to 775 (using the seven distance ranges setting[7] and considering all distances in the 1–15Å range). Atom type group numbers 913 and 903 were used to require that one (and only one) of the S1 pocket hydrophobe/aromatic ring points (3+3) plus one of the S4 pocket hydrophobe/aromatic ring points were in the pharmacophore hypotheses. Using just the S1 pocket constraint reduces the total number of site pharmacophore hypotheses from 5,393 to 2,663.

A validation study of this approach to identify feasible binding models[9] showed that such pharmacophore queries could identify a binding mode for a known factor Xa inhibitor similar to the experimental one. The Daichii factor Xa inhibitor shown in Figure 6 was subjected to DiR analysis, and the results are shown in Color Plate 1. Color Plate 1A shows one of the matches of the Daichii factor Xa inhibitor with the DiR query pharmacophores; Color Plate 1B shows a comparison of this DiR match and the conformation from the x-ray crystallographic complex (DX5603, PDB:1FAX); and Color Plate 1C

Figure 7. The four-component Ugi condensation reaction used as a sample reaction for combinatorial chemistry together with the definitions of the 3 carboxylic acids (R_1), 2 amines (R_2), 3 aldehydes (R_3), and 24 isonitriles (R_4) (both stereoisomers used for the cyclohexyl isonitriles) used to build the virtual combinatorial library of 432 compounds for DiR analysis.



R1 = Me, Ph, CH2Ph
R2 = H, Me
R3 = Et, Ph, CH2Ph
R4 = (CH2)x-*m*-benzamidine
(CH2)x-*p*-benzamidine
(CH2)x-*m*-cyclohexyl
(CH2)x-*p*-cyclohexyl
X = 0,1,2,3

R4 = (CH2)$_X$-p-benzamidine	R1 = Me,	Ph,	CH2Ph
X=0, R3 = Et	0	4	4

R4 = (CH2)$_X$-m-benzamidine	R1 = Me,	Ph,	CH2Ph
X=0, R3 = Et	0	20	17
X=1, R3 = Et	0	23	35
X=2, R3 = Et	0	30	44
X=3, R3 = Et	0	55	64
X=0, R3 = CH2Ph	8	20	21
X=1, R3 = CH2Ph	22	27	35
X=2, R3 = CH2Ph	22	38	40

Figure 8. Scores, quantified in terms of the number of 4-point pharmacophore hypotheses in the factor Xa active site that were directly matched, for sample products in the Ugi virtual combinatorial library. R_2 was H for all these compounds.

shows 7 different matches of the Daichii factor Xa ligand with the DiR query pharmacophores. Three relevant site points are highlighted in each picture as point of reference.

For this study, the design of a combinatorial library based on the Ugi four-component condensation reaction[36] (Figure 7) to match a serine protease active site (e.g., factor Xa, thrombin) was investigated, using the factor Xa crystal structure.[35] Products were selected from a small virtual library of 432 products that was built using CONCORD[22] to generate the initial 3D structures. The products were constructed from four sets of reagents: 3 carboxylic acids (R_1), 2 amines (R_2), 3 aldehydes (R_3), and 24 isonitriles (R_4) as shown in Figure 7.

A DiR analysis was performed to identify which reagents could give products that match certain steric and pharmacophoric aspects of the binding site. For example, the position of substitution on a benzamidine containing fragment (targeted to the aspartate containing S1 pocket) and the length of other hydrophobic reagents (targeted to the S4 pocket) that produces suitable compounds can be evaluated quickly in terms of fits to the site-derived pharmacophore hypotheses. Both the number and identity of the matched site pharmacophore are known for each compound, and this allows optimal subsets and reagent combinations to be chosen to maximize the total number of different site pharmacophore hypotheses that are matched. The aim of maximizing this aspect of target-based diversity is to explore with the library the maximum possible area of the binding site and maximum number of potential binding modes and let the biological screening identify the most potent binder. As described earlier, only pharmacophore hypotheses that contain both S1 and S4 pocket hydrophobic or aromatic site points are considered. This ensures that the information in the resultant fingerprints of matched site pharmacophores has a relevance to expected binding modes for serine proteases and thus to potential biological activity.

Analysis of the fingerprints showed that *meta*-substitution on the benzamidine ring was optimal in terms of matching the most site pharmacophores, and that at least a phenyl R_1 (carboxylic acid) reagent was needed unless the R_3 reagent (aldehyde) was increased in size from an ethyl group to a benzyl group, allowing it to fill the S4 pocket instead of the R_1 reagent. Using this larger R_3 group enabled the products with R_1 methyl substituents to match some site points (8–22). Figure 8 illustrates the numbers of site pharmacophore hypotheses that were matched for some key products involving the benzamidine reagents. The value to the design of using both the phenyl and benzyl groups in the R_1 position was readily evaluated: when R_4 = *m*-benzamidine (x=0) and R_3=ethyl, an R_1 phenyl substituent results in matches for 20 site pharmacophore hypotheses, whereas a R_1 benzyl group matches 17 hypotheses. A simple logical OR operation on these two fingerprints gives a total of 26 pharmacophores, quantifying the value of including the benzyl group as well: 6 of the 17 hypotheses from the product in which R1 is a benzyl moiety are not matched by the product in which R_1 is a phenyl group. Thus, new information potentially may be obtained from the synthesis and testing of the products with R_1 = benzyl. Similarly, for the longer chain length benzamidines (x=1–3), in which the benzyl groups gave larger numbers of matches than the phenyl group, the value to the design of including the phenyl group could be quantified (between reagents with similar benzamidines, and relative to other products already selected). As may be expected, the phenyl group often matched few or no additional site pharmacophores, leading to questions such as the interest of using heterocylic rings such as pyridine, which could be readily evaluated to show combinations in which new matches were obtained. The value of different length *m*-benzamidine reagents (X=0,1,2) was similarly evaluated. An increasing number of site pharmacophore hypotheses are matched for each product as X increases from 0 (26 matched) to 1 (35 matched) to 2 (41 matched). By combining the products, there is an increase in the total number of matched hypotheses: the combination of X=0 (26) and X=1 (35) provides 44 unique matches, and this increases to 51 unique hypotheses matched when the product with X=2 is included. This confirms the incremental value of products containing all three reagents.

The purpose of the studies reported here are not, however, the isolated design of a combinatorial library, but to illustrate how protein site diversity can be evaluated and quantified in terms of pharmacophore fingerprints that could be incorporated into multimetric optimizations as described earlier. A reagent may be selected and be desirable for other reasons, for example, for BCUT chemistry space optimization or to modulate physicochemical/ADME (Absorption Distribution Metabolism Excretion) properties. Energetic scoring values for docked compounds also could be included (e.g., from DOCK/CombiDOCK). The pharmacophore keys described here enable the effect on the ensemble of possible binding modes explored to be quantified during the design.

During DiR analysis of this conformationally flexible virtual library of 432 compounds, 14,000 "hits" (matching fits that passed the steric bump check with the active site) from a total of more than two million evaluated fits were identified and stored in a results database. A total of 135 different site pharmacophore hypotheses were directly matched, i.e., were used as the "substructure" pharmacophore hypothesis to drive the fit into the site. Another 558 were indirectly matched, i.e., were matched within the defined fitting tolerance when a compound was fitted into the site based on the match to another hypothesis. The analysis took a longer than usual, with an average time of almost 1 minute per compound (compared with flexible docking times of 3–19 seconds[9]), because of the high flexibility and pharmacophoric richness of the compounds (an average of 10,000 conformations per molecule were sampled) and because all fits were saved to a database.

CONCLUSION

The work described here demonstrates that simulated annealing can be used to simultaneously improve the diversity of combinatorial products in a DVS BCUT chemistry space and increase the ratio of unique to total 4-point pharmacophores (pharmacophore diversity) in the product set. The gains with the reported implementation are modest, and a more robust pharmacophore diversity scoring function is necessary due to instability observed in the current function. We currently are examining alternatives as well as evaluating the possibility of using a distance-based BCUT diversity scoring scheme in combination with the pharmacophore function.

The 3D pharmacophore fingerprint method has been extended to include protein sites, the DiR extension including the shape of the target site in the analysis. Given the increasing amount of 3D structural information for targets, this provides timely new approaches for molecular similarity and diversity applications such as virtual screening and combinatorial library design. The method produces fingerprints that can be used independently or be used in combination with other descriptors such as the BCUTs and full ligand pharmacophore fingerprints described in this article. A simultaneous optimization of ligand-based diversity and target protein-based diversity thus is possible. The fingerprints allow the quantification of both ligand-based and structure-based diversity in terms of multiple 3D pharmacophore hypotheses. Several million hypotheses are systematically searched, stored, and compared. Protein binding sites are evaluated in terms of complementary pharmacophore points. Flexible and promiscuous compounds, such as small peptides, can be handled. Diverse sets of active compounds (e.g., screening hits) can be analyzed and an unlimited ensemble of hypotheses used for further searching, etc. The problem of needing to delineate a single or a small number of hypotheses is avoided. The pharmacophore fingerprint method provides a powerful 3D similarity (virtual screening) and library design tool that now can use protein-binding site information.

ACKNOWLEDGMENTS

The authors wish to thank Daniel Cheney, Andrew Good, Sung Jin Cho, and Yax Sun for their input into the development and application of the multiple potential pharmacophore fingerprint method, together with Keith Davies and Cathy White-Davies, and Professor Bob Pearlman for many useful discussions and the development of the DiverseSolutions software. We are also grateful to a reviewer for suggesting the use of Hopkins' statistic.

REFERENCES

1 Jamois, E.A., Hassan, M., and Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 63–70

2 For reviews, see Drewry, D.H., and Young, S.S. Approaches to the design of combinatorial libraries. *Chemom. Intell. Lab. Syst.* 1999, **48**, 1–20; Agrafiotis, D.K., Myslik, J.P., and Salemme, F.R. Advances in diversity profiling and combinatorial series design. *Molecular Diversity* 1999, **4**, 1–22

3 DiverseSolutions was developed by R.S. Pearlman and K.M. Smith at the University of Texas, Austin, and is distributed by Tripos, Inc., St. Louis, MO. Version 4.0.5 was used in this work

4 Pearlman, R.S., and Smith, K.M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 28–35

5 Pearlman, R.S. *DiverseSolutions user's manual.* University of Texas, Austin, TX, 1995

6 Pearlman, R.S., and Smith, K.M. Novel software tools for chemical diversity. *Perspect. Drug. Discovery Design* 1998, **9**, 339–353

7 Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C., and 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

8 Mason, J.S., and Cheney, D.L. Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores. *Proc. Pacific Symp. Biocomputing* 1999, **4**, 456–467

9 Mason, J.S., and Cheney, D.L. Library design and virtual screening using multiple 4-point pharmacophore fingerprints. *Proc. Pacific Symp. Biocomputing* 2000, **5**, 576–587

10 Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. *Numerical recipes in C: The art of scientific computing.* Cambridge University Press, New York, 1984, pp. 444–455

11 Agrafiotis, D.K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 841–851

12 Good, A.C., and 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

13 Agrafiotis, D.K. On the use of information theory for assessing molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 576–580

14 Agrafiotis, D.K., and Lobanov, V.S. An efficient implementation of distance-based diversity metrics based on K-D trees. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 51–58

15 Zheng, W., Cho, S.J., Waller, C. L., and Tropsha, A. Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: A novel computational tool for universal library design and database mining. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 738–746

16 MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577

17 Hopkins, B.A. New method of determining the type of

distribution of plant individuals. *Ann. Bot.* 1954, **18**, 213–226

18 Lawson, R.G., and Jurs, P.C. New index for clusteriñg tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.* 1990, **30**, 36–41

19 Fernåndez Pierna, J.A., and Massart, D.L. Improved algorithm for clustering tendency. *Anal. Chim. Acta* 2000, **408**, 13–20

20 DiR. http://www.oxmol.com/software/chem-x/dir/

21 Murray, C.M., and Cato, S.J. Design of libraries to explore receptor sites. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 46–50

22 Pearlman, R.S. *CDA News* 1987, **2**, 1–7; Balducci, R., McGarity, C.M., Rusinko III, A., Skell, J.M., Smith, K., and Pearlman, R.S. University of Texas at Austin. CONCORD is available from Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144

23 Menard, P.R., Mason, J.S., Morize, I., and Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 1204–1213

24 Schnur, D. Design and diversity of large combinatorial libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 36–45

25 http://www.oxmol.com/apps/pharmdiv/phexplain.shtml

26 Chem-X Software. Oxford Molecular Group, Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, UK

27 Mason, J.S., and Pickett, S.D. Partition-based selection. *Perspect. Drug. Discovery Design* 1998, **7/8**, 85–114

28 GRID software. Molecular Discovery Limited, West Way House, Elms Parade, Oxford OX2 9LL, UK; Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 1985, **28**, 849–857

29 Willett, P. *Similarity and clustering in chemical information systems.* Research Studies Press, Letchworth, 1987

30 Pickett, S.D., Mason, J.S., and McLay, I.M. Diversity profiling and design using 3D pharmacophores: Pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* 1996, **36**, 1214–1223

31 Hann, M., Hudson, B., Lowell, X., Lifely, R., Miller, L., and Ramsden, N. Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 897–902

32 Sybyl 6.5 software. Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144

33 CombiLibMaker, version 4.2.2. Developed by Pearlman, R.S., Stewart, E.L., Brusniak, M.K., Leong, M.K., and Smith, K.M. at the University of Texas, Austin, TX. Distributed by Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144

34 Sun, Y., Ewing, T.J.A., Skillman, A.G., and Kuntz, I.D. CombiDOCK: Structure-based combinatorial docking and library design. *J. Comput.-Aided Mol. Design* 1998, **12**, 597–604

35 Tulinsky, A., Padmanbhan, K., Padmanbhan, K.P., Park, C.H., Bode, W., Huber, R., Blankenship, D.T., Cardin, A.D., and Kisiel, W. Structure of human des9(1-45) factor Xa at 2.2 Å resolution. *J. Mol. Biol.* 1993, **232**, 947–966

36 Ugi, I., and Steinbruckner, C. Isonitriles. II. Reaction of isonitriles with carbonyl compounds, amines, and hydrazoic acid. *Chem. Ber.* 1961, **94**, 734–742