Pacific Symposium on Biocomputing 5:573-584 (2000)

LIBRARY DESIGN AND VIRTUAL SCREENING USING MULTIPLE 4-POINT PHARMACOPHORE FINGERPRINTS

Jonathan S. MASON, Daniel L. CHENEY
Computer-Assisted Drug Design, Department of Macromolecular Structure
Bristol-Myers Squibb Pharmaceutical Research Institute
PO Box 4000, Princeton NJ 08543, USA
Email: jonathan.mason@bms.com

The use of multiple potential 4-point three-dimensional (3-D) pharmacophores for the design of combinatorial libraries and for virtual screening is discussed. These 3-D pharmacophoric fingerprints can be calculated from both ligands and complementary to a protein site, with a common frame of reference, and can be very rapidly searched to identify common and different 4-point pharmacophoric shapes in compounds and protein sites. A new extension to the method for structure-based design is reported that uses the shape of the target site as an additional constraint. This enables the docking process, for example in library design and virtual screening, to be quantified in terms of how many, and which, pharmacophoric hypotheses can be matched by a compound or a library of compounds.

1 Introduction

Methods for molecular similarity and diversity that use properties relevant to drug-receptor interactions and that can be calculated for both ligands and receptors are needed for many computer-assisted drug design (CADD) applications. These methods need to be able to handle rapidly large numbers of structures, often of a relatively high conformational flexibility, with applications for analysis and design such as virtual screening and combinatorial chemistry library design.

3-D pharmacophore fingerprints,[1,2,3,4] consisting of multiple potential 3- and 4-point pharmacophores can be calculated systematically and with conformational flexibility for structures using software such as the ChemDiverse[5] module of Chem-X.[6] For ligands, the six pharmacophoric features (hydrogen bond donors, hydrogen bond acceptors, acidic centers, basic centers, hydrophobic regions and aromatic ring centroids) are automatically assigned to atoms or dummy atom centroids, whereas for a protein site, complementary site-points with associated pharmacophoric features are first generated and the fingerprint generated from these. A significant increase in the amount of shape information and resolution was found using 4-point pharmacophores, including the ability to distinguish chirality, a fundamental requirement for many ligand-receptor interactions.

The pharmacophore fingerprints (~10 and 2.3 million 4-point possibilities with 10 and 7 distance ranges respectively per feature-feature distance) give a common frame of reference for comparing different ligands and for comparing ligands to protein structures using the complementary potential pharmacophores. Applications to virtual screening and library design of these fingerprints are discussed below, together with the new use of pre-calculated fingerprints.

A new pharmacophore-based method known as "Design in Receptor" (DiR)[7,8] that includes the shape of the target site in the analysis provides new possibilities for docking and structure-based virtual screening and library design. The method enables a novel quantification of target-based diversity, based on the site-derived pharmacophore hypotheses. New modifications to the method to make it more effective for virtual screening and library design and example results for docking/virtual screening and combinatorial library design are presented below.

2 Methods

2.1 Generation of the 4-point pharmacophore fingerprints / Chem-X software

The Chem-X[5,6,7] software is a general molecular modeling package, with specialist optional modules such as ChemDiverse, 4-centre pharmacophores and Design in Receptor (DiR) that were used for this work. 3-D structures were generated for ligands using the CONCORD[9] program, and were read into a Chem-X/ChemDBS-3D database from an SD file using a customized parameterization file and fragment database to assign atom types; a single conformer was stored in the database, and conformational sampling done "on the fly". Six key 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 donor, hydrogen bond acceptor, acidic center (negatively charged at physiological pH 7) and basic center (positively charged at pH 7)] and the addition of dummy atoms [for hydrophobic regions and aromatic rings]. All combinations of four pharmacophoric features are considered, together with 7 or 10 distance ranges for each of the six distances, as illustrated in figure 1.

A pharmacophore "fingerprint" is thus generated that indicates the presence or absence of all the theoretically possible combinations of features and distances (potential pharmacophores); an additional chirality indicator can be added to applicable potential pharmacophores. About 2.3 million (7 distance ranges) and 9.7 million (10 distance ranges) 4-point potential pharmacophores are thus considered for each ligand or set of site-points.

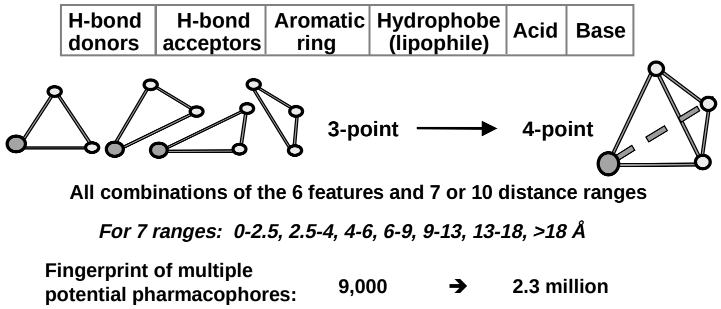

H-bond donors
H-bond acceptors
Aromatic ring
Hydrophobe (lipophile)
Acid
Base
3-point
4-point
All combinations of the 6 features and 7 or 10 distance ranges
For 7 ranges: 0-2.5, 2.5-4, 4-6, 6-9, 9-13, 13-18, >18 Å
Fingerprint of multiple potential pharmacophores:
9,000
2.3 million

Figure 1: Definition of multiple potential 3-D pharmacophore fingerprints

For ligands, effective conformational sampling is needed and is used in the generation of the pharmacophore fingerprints; the method used is a customization of the Chem-X/ChemDiverse method, based on an "on-the-fly" generation of conformers done at search time, with a quick evaluation of the conformation performed based on a steric contact check to reject poor or invalid conformations. Using up to 4 rotamers per bond the maximum analysis time is only 5–15 seconds on a Silicon graphics R10000 250Mz processor, using a systematic analysis where possible and a random analysis for very flexible molecules. Only this internal bump check was used to eliminate unreasonable conformations, although additional rule-based checks or energy calculations could be applied. This generates a relatively information rich fingerprint that includes as a union for all the sterically accessible conformations all potential pharmacophores; this encodes potentially important information about flexibility, two molecules with similar functional groups linked by a rigid and a flexible linker would have different fingerprints (one small, one larger), whereas if only the low energy conformation was used they could both have the same small fingerprint.

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 assigned to these points. The site-points can be generated by many different methods, such as geometric ones (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[10]). For example, a dummy atom site- point assigned the hydrogen bond acceptor feature may be placed at a hydrogen bonding distance from sterically accessible N-H groups. The combined set of all site-points represents a theoretical molecule that binds to all available positions, and the fingerprint of potential pharmacophores are generated for this “molecule” in the same way as for a normal compound.

Further details of all these methods have been previously described.[1,2,3]

2.2 Design in Receptor "DiR" Method

This new method enables the steric shape of a protein site to be used as an additional constraint in the comparison of the fingerprints of multiple potential pharmacophores of a protein site with a ligand; it will be a released module of the Chem-X software in 2000. The method is equivalent to simultaneous 3-D database searching using multiple 3-D pharmacophoric queries and steric constraints, but with only one conformational sampling being necessary. Speed enhancements were suggested and incorporated into the software to enable the method to be fast enough used for virtual screening and library design (see section 3.2.1). These included the use of an atom based bump check instead of grid-based surface maps and adaptive pharmacophores, that remove a pharmacophore query from consideration for a particular structure once it has been matched once and ends the analysis for that structure if no more fits are possible. The output is a database of structures fitted onto the site points, that can be used as input for other scoring methods or minimization etc., and a "key" (fingerprint) per structure and per database of which pharmacophore hypotheses have been matched.

The docking and library design example studies used 4-point pharmacophores based on 23 complementary site points added to the factor Xa crystal structure based on GRID analyses, with 162 atoms from the active site defining the shape and used in the bump check [CPK (2/3 VDW) radii, maximum 3 bumps]. 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 the site-point pharmacophore hypothesis.

3 Results

3.1 Pharmacophore fingerprint-based virtual screening and library design

Pharmacophoric fingerprints can be derived from either ligands or complementary site-points to a protein-sites and used to rapidly quantify which potential pharmacophores are in common, or different, between ligands, between ligands and protein sites, and between protein sites. As all combinations of features and distance ranges between them are considered, the method provides a measure of both ligand and protein-site diversity. Previous studies[1,3,4] have shown that the increased shape (and chiral) information present in the tetrahedral 4-point pharmacophoric descriptions [compared to 2-point (distance) or 3-point (triangle) descriptions] is normally needed for molecular similarity studies (ligand-ligand, ligand-protein) that use only the fingerprint. The pharmacophore fingerprint method has the advantage over many other 2-D and 3-D similarity methods that it is possible to use as input information from flexible and information rich compounds such as small peptides; the goal of the studies can then be to find a set of compounds that together explore all the pharmacophore hypotheses in the input molecule, rather than a single compound.

The pharmacophore fingerprints can be pre-calculated (with conformational sampling) and stored in an efficient format, and then searched directly very rapidly (S.J. Cho and J.S. Mason, unpublished results); a space efficient format of one line of encoded information per compound is used (11KB for 1000 pharmacophores, compared to 58KB ASCII and 1.4MB binary formats in Chem-X). On a Silicon Graphics workstation (single R10000, 250Mhz cpu) it is then possible to search more than 700K compounds per hour, including writing a file of which potential pharmacophores are in common for each compound.

The results of the comparison of the pharmacophore fingerprints can be used as part of the scoring and subset selection process; both the numbers (per compound and in common) of potential pharmacophores and the actual pharmacophores can be used. Similarity indices such as the Tanimoto coefficient can be generated from the numbers,[3,4] whilst the new fingerprints generated of the actual common pharmacophores can be used to select a set of structures that have similarity to a target structure (e.g. using Tanimoto coefficient) but with each structure having this similarity with different potential pharmacophores. An example of where this would be useful is with relatively flexible and promiscuous compound such as a small peptide as input, where the goal is not to find similarly promiscuous compounds but an ensemble of simpler compounds that together express as many as possible of the target potential pharmacophores. This can be achieved by selecting from compounds without too large a total number of potential pharmacophores the structure with most common potential pharmacophores, then continuing down the list only selecting further compounds if their common pharmacophore fingerprint contains new (e.g. 10) potential pharmacophores relative to the union of the common fingerprints of all those already selected. This process effectively excludes from the fingerprint of the reference compound the potential pharmacophores of each compound selected, forcing further compounds selected to match different potential pharmacophores.

Another important modification to the 3-D pharmacophore fingerprint method is to force one of the features to be a group or substructure of interest.[3,4] 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[3]. 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, i.e. features, 3-D distances between them and for 4-point pharmacophores an optional flag for chirality, searching the fingerprints provides for this level of resolution a very rapid method for 3-D database searching and pharmacophore identification.

The multiple potential 3-D pharmacophore fingerprint method has been found to provide a complementary approach for molecular diversity applications to the DiverseSolutions BCUT method[11,12] that is based on atomic/molecular properties important for ligand-receptor interactions. Combined applications include subset selection[13] and combinatorial library design (B.R. Beno and J.S. Mason, unpublished results, see reference 14 for library design applications of BCUTs).

3.2 Structure-based virtual screening and library design

Pioneering methods such as DOCK and the extended version CombiDOCK[15] for combinatorial libraries have been successfully used for virtual screening and are driven by the shape of the target site, with additional constraints possible for pharmacophoric features. The DiR method focuses on the pharmacophoric match, using the systematic definition of potential pharmacophores to provide a method 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. The resultant pharmacophore fingerprint can be stored and combined as for ligand-derived fingerprints, enabling the design (of a virtual screening set, of compounds for the combinatorial library) to optimize the matching of all possible site-derived pharmacophoric hypotheses, or to enrich a subset of interest. The "score" thus obtained for a ligand in a site does not attempt to quantify the potential interaction energy (this can be done separately), as with DOCK/CombiDOCK, but which and how many pharmacophoric hypotheses can be matched within 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 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 quantifying different pharmacophore hypotheses a ligand may match upon binding. It is thus possible to evaluate which ligands are able to fit in the site whilst matching at least one set of pharmacophoric features, and thence to design a subset of ligands that match as many pharmacophoric hypotheses as possible. Additional constraints can be set to ensure that at least one of a group of pharmacophore site-points is included in each pharmacophore hypothesis used.

3.2.1 Docking and Virtual screening

The DiR method provides a rapid method to flexibly dock compounds into a protein site evaluating multiple pharmacophore hypotheses. To enable the method to run fast enough for virtual screening and library design a special option of “adaptive” pharmacophore queries was incorporated into the software. A pharmacophore query is removed from consideration for a particular structure once it has been matched once and ends the analysis if no more fits are possible. This avoids the potentially time-consuming docking of multiple conformers to a single pharmacophore hypothesis, yet still enables a quantification of target-based diversity coverage (how many site pharmacophore hypotheses a ligand can match).

An example of the DiR method is the docking of factor Xa inhibitors to the active site of the x-ray crystal structure. A typical ligand such as a Daiichi inhibitor (see figure 2A) could be docked in 3 seconds, having sampled 384 conformations and identifying 7 matches fitted to the active site (see figure 3A); all 4-point pharmacophore hypotheses that contained at least one feature from the S1 and S4 pockets were sampled.

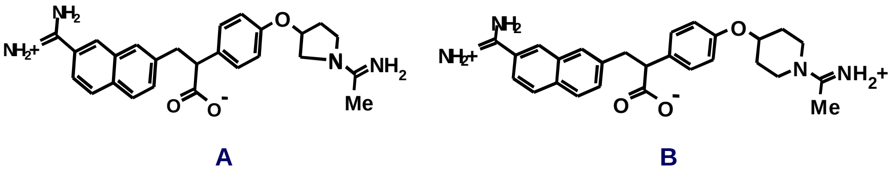


Figure 2: Daichii Factor Xa ligands used in DiR docking studies

A related compound (DX5603: 6 → 5-membered cyclic guanidine, see figure 2B) was evaluated with increased conformational sampling: 27,000 conformations of DX5603 were sampled in 19 seconds. The six resultant matches are shown in figure 3B, one of which is very close to the published x-ray structure of the ligand-enzyme binding complex (PDB: 1FAX, see figure 3C). Both studies used adaptive pharmacophore queries and atom-based bump-checking.

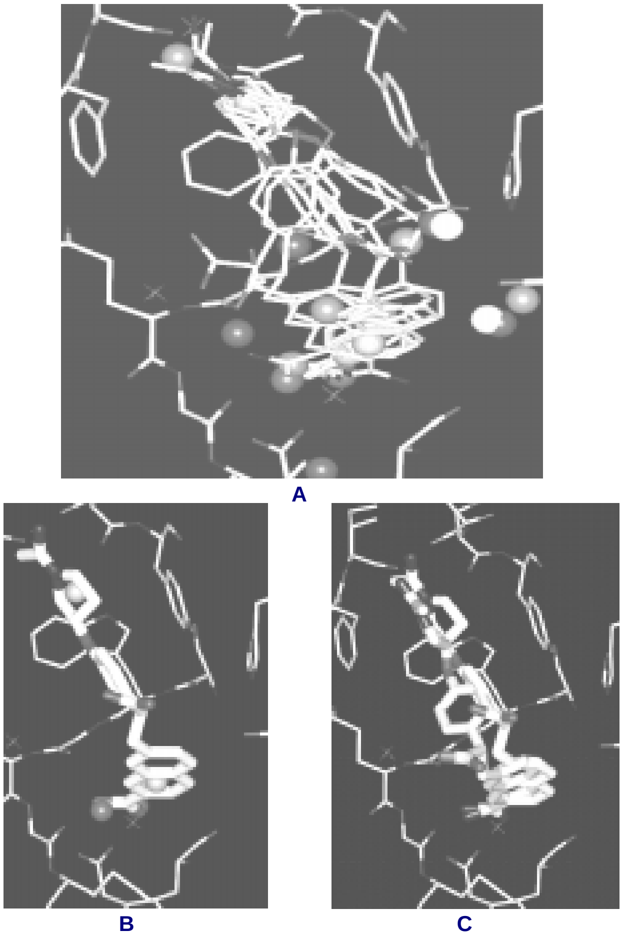
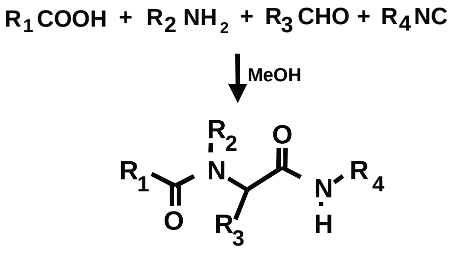


Figure 3: DiR docking studies of Daichii factor Xa inhibitors [A = 7 matches for structure in figure 2A, B = one of 6 matches for the DX5603structure in figure 2B, C = comparison of DiR hit and x-ray complex of DX5603 (PDB: 1FAX)].

3.2.2 Combinatorial library design

An example of this is the design of a combinatorial library based on the Ugi four-component condensation reaction (see figure 4) to match serine protease active sites (e.g. Thrombin, Factor Xa).

A small virtual library of 384 products was built using CONCORD to generate the starting 3-D structures, with 3 carboxylic acids (R1), 2 amines (R2), 3 aldehydes (R3) and 12 isonitriles (R4), as shown in figure 4.



R1 = Me, Ph, CH_2Ph
R2 = H, Me
R3 = Et, Ph, CH_2Ph
R4 = $(CH2)_X$-*m*-benzamidine
$(CH2)_X$-*p*- benzamidine
$(CH2)_X$-*m*-cyclohexyl
$(CH2)_X$-*p*- cyclohexyl
X = 0,1,2,3
→ 384 compounds

Figure 4: The 4-component Ugi condensation reaction used as a sample reaction for combinatorial chemistry together with the definitions of the 3 carboxylic acids (R1), 2 amines (R2), 3 aldehydes (R3) and 12 isonitriles (R4) used to build the virtual combinatorial library of 384 compounds for DiR analysis.

The addition of complementary site-points for the target protein site was achieved through the use of GRID maps. A DiR analysis was then performed to identify which reagents could give products that match certain steric and pharmacophoric aspects of the binding site. 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 quickly evaluated, in terms of fits to site-derived pharmacophore hypotheses

Thus, using 4-point pharmacophore hypotheses that were constrained to force at least one complementary point from the S1 and S4 pockets, it could be shown that meta-substitution on the benzamidine ring was optimal, and that at least a phenyl R1 (carboxylic acid) reagent was needed unless the R3 reagent (aldehyde) was increased in size from an ethyl group to a benzyl group and could thus fill the S4 pocket. Using this larger R3 group enabled the products with R1 only a methyl to match some site points (8-22). Some of the results of the number of site pharmacophore hypotheses matched per product are shown in figure 5.

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 5: 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; R2 was H for all these compounds

During the DiR analysis of this very conformationally flexible virtual library of 384 compounds, a total of 14,000 “hits” (matching fits that passed the steric bump check with the active site) were found and stored to a results database, from a total of 2 million fits evaluated. A total of 135 different site pharmacophore hypotheses were directly matched, i.e. were used as the “substructure” match 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 average time of almost one minute per compound (compared with flexible docking times of 3-19 seconds in 3.2.1.) because of the high flexibility and pharmacophoric richness of the compounds (average of 10,000 conformations per molecule were sampled) and because all fits were saved to a new database.

By using the pharmacophore key stored for each compound that details which site pharmacophore hypotheses were matched, optimal subsets and reagent combinations can be chosen to maximize the total number of site pharmacophore hypotheses that are matched. The aim of maximizing this aspect of target-based diversity is to explore with the library the maximal amount of the binding site and number of potential binding modes.

4 Conclusion

The 3-D pharmacophore fingerprint method (based on a systematic analysis for multiple 4-point potential 3-D pharmacophores), now extended with the DiR method to include the shape of the target site in the analysis, provides new approaches for molecular similarity and diversity applications such as virtual screening and combinatorial library design. The methods allow the quantification of both ligand-based and structure-based diversity, in terms of multiple 3-D pharmacophore hypotheses. Multiple hypotheses are stored and compared (routinely up to 10 million are checked for per structure), enabling the method to handle flexible and promiscuous compounds such as small peptides, or diverse sets of screening hits, for which a single or a small number of hypotheses cannot easily be delineated. The method provides a powerful new 3-D similarity (virtual screening) and library design tool.

Acknowledgements

The authors wish to thank Sung Jin Cho, Andrew Good and Brett Beno for their input into the development and application of the multiple potential pharmacophore method.

References

1. J.S. Mason and S.D. Pickett, “Partition-based Selection”, Perspectives in Drug Discovery and Design, 7/8:85-114 (1998).
2. S.D. Pickett, J.S. Mason and I.M. McLay, “Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ)”, *J. Chem. Inf. Comput. Sci.* 36:1214-1223 (1996)
3. J.S. Mason, I. Morize, P.R. Menard, D.L. Cheney, C. Hulme and R.F. Labaudiniere, “A 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. ,* 42:3251-3264 (1999)
4. J.S. Mason and D.L. Cheney, “Ligand-Receptor 3-D Similarity Studies Using Multiple 4-Point Pharmacophores”, *Proc. Pacific Symposium on Biocomputing*, 4:456-467 (1999).
5. K. Davies and C. Briant, http://www.netsci.org/Science/Combichem/feature05.html
6. Chem-X Software, Oxford Molecular Group, Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, U.K.
7. http://www.oxmol.com/prods/chem-x/dir/dir.html
8. C.M. Murray and S.J. Cato, “Design of libraries to explore receptor sites”, *J. Chem. Inf. Comput. Sc.i*, 39:46-50 (1999).

9. R.S. Pearlman,*CDA News,* **2**, 1-7 (1987); , R. Balducci, C.M. McGarity, A. Rusinko III, J.M. Skell, K. Smith and R.S. Pearlman, University of Texas at Austin, CONCORD is available from Tripos Inc, St. Louis, Missouri.

10. Molecular Discovery Limited: West Way House, Elms Parade, Oxford OX2 9LL, U.K.

11. R.S. Pearlman and K.M. Smith, “Novel software tools for chemical diversity”, *Perspect. Drug. Discov. Design,* 9:339-353 (1998).

12. R.S. Pearlman and K.M. Smith, “Metric validation and the receptor-relevant subspace concept”, *J. Chem. Inf. Comput. Sci.*, 39:28-35 (1999).

13. P.R. Menard, J.S. Mason, I. Morize and S. Bauerschmidt, “Chemistry space metrics in diversity analysis, library design, and compound selection”, *J. Chem. Inf. Comput. Sci.*, 38:1204-1213 (1998).

14. D. Schnur, “Design and diversity of large combinatorial libraries using cell-based methods”, *J. Chem. Inf. Comput. Sci.*, 39:36-45 (1999).

15. Y. Sun, T.J.A. Ewing, A.G. Skillman and I.D. Kuntz, “CombiDOCK: structure-based combinatorial docking and library design”, *J. Comput.-Aided Mol. Des.* 12:597-604 (1998).