

A Self-Organizing Algorithm for Molecular Alignment and Pharmacophore Development

DEEPAK BANDYOPADHYAY, DIMITRIS K. AGRAFIOTIS

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

Received 23 July 2007; Revised 29 August 2007; Accepted 11 September 2007

DOI 10.1002/jcc.20854

Published online 12 November 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We present a method for simultaneous three-dimensional (3D) structure generation and pharmacophore-based alignment using a self-organizing algorithm called Stochastic Proximity Embedding (SPE). Current flexible molecular alignment methods either start from a single low-energy structure for each molecule and tweak bonds or torsion angles, or choose from multiple conformations of each molecule. Methods that generate structures and align them iteratively (e.g., genetic algorithms) are often slow. In earlier work, we used SPE to generate good-quality 3D conformations by iteratively adjusting pairwise distances between atoms based on a set of geometric rules, and showed that it samples conformational space better and runs faster than earlier programs. In this work, we run SPE on the entire ensemble of molecules to be aligned. Additional information about which atoms or groups of atoms in each molecule correspond to points in the pharmacophore can come from an automatically generated hypothesis or be specified manually. We add distance terms to SPE to bring pharmacophore points from different molecules closer in space, and also to line up normal/direction vectors associated with these points. We also permit pharmacophore points to be constrained to lie near external coordinates from a binding site. The aligned 3D molecular structures are nearly correct if the pharmacophore hypothesis is chemically feasible; postprocessing by minimization of suitable distance and energy functions further improves the structures and weeds out infeasible hypotheses. The method can be used to test 3D pharmacophores for a diverse set of active ligands, starting from only a hypothesis about corresponding atoms or groups.

© 2007 Wiley Periodicals, Inc. J Comput Chem 29: 965–982, 2008

Key words: SPE; stochastic proximity embedding; distance geometry; pharmacophore modeling; pharmacophore development; molecular alignment; conformational analysis; self organizing

Introduction

Pharmacophore methods allow the abstraction of a set of active molecules by a three-dimensional (3D) geometric arrangement of chemical groups. Such methods are key to drug discovery efforts, since they power 3D database searches and allow the discovery of new ligands dissimilar in 2D structure to currently known actives but retaining the activity (scaffold hopping¹) and surmounting physical, chemical, and/or metabolic liabilities (lead evolution²). Pharmacophores have also helped side-effect prediction with models for HERG K⁺-channel binding.³ Gillet et al.⁴ describe the *de novo* design of ligands fitted to a 3D pharmacophore hypothesis. Pharmacophore methods have been combined with structure-based docking methods,^{5–8} and additionally with molecular mechanics/dynamics methods^{9,10} and clustering techniques.¹¹ Pharmacophores have also been developed for ADME/tox predictions.¹²

With the availability of receptor structural information, pharmacophores are increasingly being derived by receptor-based

methods; however, for many important therapeutic target classes, such as G-protein coupled receptors (GPCRs), the target protein structures or models are not available, and pharmacophore modeling is one of the easiest methods to characterize the active molecules and quantify their structure-activity relationships. Pharmacophore-based QSAR models guide subsequent lead-optimization efforts to obtain favorable chemical and metabolic properties without losing activity.

Pharmacophore perception methods attempt to extract common steric and electronic features from a set of active molecules, while pharmacophore development methods represent them as points or groups at certain distances to each other in 3D, and decide between competing pharmacophore hypotheses which one best explains the observed activity patterns. Once extracted, a pharmacophore serves as a core set of features representing the important part of a molecule for a chemist, as a fil-

Correspondence to: D. K. Agrafiotis; e-mail: dagrafio@prds.jnj.com

ter to determine or predict activity of new compounds, and as a search query to a 3D conformational database to identify novel putative ligands.

The method we describe in this article involves iterative conformational search and pharmacophore alignment that starts from 2D connection tables and 1D pharmacophores (inter-molecular atom or group correspondences), and produces an alignment of 3D structures that helps elucidate the pharmacophore in 3D. When the conformational analysis and alignment method converges, which happens when the supplied pharmacophore correspondence is plausible, both the geometries of the individual molecules and of the alignment are correct. The 3D pharmacophore can then be extracted from the geometry of the corresponding pharmacophore groups in the alignment. If multiple trials return different plausible alignments, this provides several alternative viable pharmacophore hypotheses.

Related Work

Research in the field of pharmacophore perception, development/elucidation and use in drug design has seen a renaissance in the last few years, and the state of the art and several recent technological advances are summarized in the excellent book by Güner.¹³ Several commercial software products such as Accelrys' Catalyst, Schrödinger's Phase, CCG's MOE and offerings from MDL, Tripos, Chem-X and other companies implement pharmacophore detection and 3D database searching. Specialized commercial pharmacophore software packages are also emerging, such as Inte:Ligand's LigandScout,^{14,15} which uses a polynomial-time matching algorithm for faster pharmacophore detection and development.

We review mostly the advances in the ligand-based methods, which take as input a group of known active ligands. Methods that incorporate structural information from the receptor binding site into the derivation of the pharmacophore are not covered, since these methods can be cast as a ligand-based pharmacophore problem by including distance constraints to external points in the active site, or by adding excluded volumes where atoms from the binding site would sterically hinder the presence of a ligand atom. Several excellent reviews of virtual screening methods cover receptor and ligand-based pharmacophores.^{16–22}

An example of a mature automated method for ligand-based pharmacophore design is Catalyst from Accelrys Inc.^{23,24} It combines a multi-conformer database generated by an algorithm called poling, and a flexible search mechanism encoded in the HipHop²⁵ and Hypogen²⁶ methods that generates pharmacophores by finding the geometries of shared sets of chemical groups between the conformations of each active compound, and removing the sets that are also shared by many of the inactive compounds. Catalyst includes a 3D database search program, and has been used extensively to design and optimize new ligands.^{27–29}

Some new directions in pharmacophore research involve the representation of the pharmacophore and how it is perceived and developed.

Bernard et al.³⁰ compute a representation called a 2D conformationally-sampled pharmacophore that does not require molec-

ular alignment. They express the conformational space available to each molecule as a 2D probability distribution of distances and bond angles between selected atoms. Then they find the overlapping probability distribution, which reveals possible pharmacophore hypotheses.

Several methods try to combine information from ligands and the receptor or convert one to the other. Pseudoreceptor modeling methods such as Yak³¹ and ProPose³² construct peptidic active site models around a small molecule or aligned ensemble, which are found to compare favorably with crystal structures. Pseudo-ligand approaches such as Virtual Ligand³³ go the opposite route, deriving an idealized constellation of potential ligand sites that interact with binding site residues. Renner et al.³⁴ designed inhibitors of a HIV protein-RNA interaction using a fuzzy pharmacophore model with potential pharmacophore points and associated probabilities³⁵ and a correlation vector representation for molecular similarity searching.³⁶

Some of the areas of emerging research into pharmacophore development address concerns about the speed and the versatility of current methods. When a pharmacophore development method is to be used with large sets of highly flexible actives, both speed and accuracy/coverage become very important. If the flexibility is accounted for by storing single or multiple low-energy conformations and using matching techniques for rigid structures, the method may not find some bound conformations that are not low-energy structures but would have significant strain energy if they were unbound.³⁷ Also, the number of available conformations for each flexible molecule can be very large, and increase both the combinatorial complexity of alignment and the number of generated hypotheses requiring further analysis and filtering.³⁸ Kristam et al.³⁹ have compared different conformational analysis techniques for the pharmacophore development and 3D database search tasks in the Catalyst software. On the other hand, if conformations are matched by performing conformational analysis on-the-fly during alignment, speed of the conformational analysis and alignment methods becomes an issue. Some examples of iterative but slow approaches are the genetic algorithms used in systems like GASP⁴⁰ and GALAHAD.^{41,42} Another method, based on multi-objective optimization of an evolutionary algorithm and Pareto ranking, aims to generate a manageable number of diverse but plausible hypotheses.⁴³

We present an attempt at this problem, by simultaneously and iteratively performing conformational analysis and alignment, in effect casting the slower alignment procedure as a faster conformational analysis procedure for which a fast method (SPE) is available. The basic idea of ensemble distance geometry was pioneered by Sheridan et al.⁴⁴ to derive the nicotine pharmacophore automatically. Here, we apply their basic principles to derive an ensemble version of a nonlinear geodesic embedding, not a strictly distance-geometry-driven one.⁴⁵

Method

First, we briefly describe the method of 3D structure generation by self-organization,⁴⁶ based on Stochastic Proximity Embedding (SPE).^{45,47} SPE is a fast and scalable method for producing

low-dimensional Euclidean maps that preserve the intrinsic dimensionality and nonlinear geometry of complex high-dimensional observation spaces. When applied to conformational sampling, SPE generates atomic coordinates in a 3D space that best preserve a set of distance and chiral constraints derived from the covalent structure of the molecule.

Input and Preprocessing

The molecules to be aligned are stored in an input SMILES or SDF file. If the SDF file for a molecule contains 3D coordinates, they are used as initial coordinates for the alignment. There are several types of constraints for each molecule in the alignment:

- *Internal distance constraints* connect two atoms within the same molecule, specifying the indices of the two atoms and a minimum and maximum allowed distance.
- *External distance constraints* connect an atom of a molecule to an external point with known coordinates, typically one from a protein binding site taken from a crystal structure or homology model. External distance constraints also specify the range of distances that are allowed between the atom and the external point.
- *Pharmacophore distance constraints* connect pairs of atoms in two different molecules, specifying a minimum (by default set to zero) and a maximum distance between them (set according to the tolerance for aligning that point, by default 0.05 Å). Pharmacophore points are specified in a defined order for each molecule, and corresponding pharmacophore points are paired up between all n pairs of molecules in which they occur, yielding ${}^n C_2$ distance constraints.

Two variants of pharmacophore constraints require special handling:

- *Centroid constraints* occur when an aromatic ring or hydrophobic region in a molecule acts as a pharmacophoric group. Such a group is represented in SPE by the centroid of the atoms that it contains, which is calculated from their coordinates at each iteration of the algorithm and not stored as an additional pseudo-atom.
- *Angle constraints* occur when planar aromatic rings or single hydrogen-bonded atoms have a normal or direction vector that must match between corresponding pairs of pharmacophore atoms or groups; it is not enough that the point or centroid coordinates match. Angle matching is implicit for pharmacophore groups specified as aromatic rings, and may be explicitly specified for hydrogen-bond donors and acceptors (whose directionality may also be modeled using a pair of points, or ignored). When two matching atoms or groups have an angle constraint, SPE will try to minimize the angle between them by treating it as its distance equivalent, the arc length. The detailed formulas for coordinate updates, distance error evaluation, and gradient for an angle constraint will be given later in this section.

Our method for alignment applies the SPE algorithm to an ensemble of molecules by treating it as a single molecule whose

components are concatenated, but not bonded, to each other. For this purpose, sequence numbers of the atoms in each successive molecule are offset by adding the total number of atoms encountered in all the previous molecules, so that they form a single sequence. The same offsets are also applied to the sequence numbers of the atoms referenced in all the constraints, so that the constraints refer to the same atoms in the concatenated molecule as they do in the original ones.

Modified Stochastic Proximity Embedding

After the initial preprocessing stage, SPE is run on the aggregate molecule using a variant of the earlier algorithm of Xu et al.⁴⁶ The specific stages where our algorithm differs from the original one are highlighted in the description below:

- *Distance geometry rules*-Geometric rules derived from standard covalent geometry set plausible ranges for quantities such as bond lengths, bond angles, dihedral angles, tetrahedral volumes (or planarity, when it is nearly zero), steric overlap, and others. These constraints are added to SPE's distance geometry engine along with the externally supplied constraints. Since steric overlap is allowed between atoms in different molecules that are being aligned, all such pairs of atoms are excluded from the steric overlap test by assigning a minimum distance bound of zero rather than the sum of their van der Waals radii.
- *Distance geometry embedding step*-Embedding is an iterative step of SPE where randomly picked constraints are applied to the current coordinates to produce new coordinates. The original SPE chooses from three different constraint types to embed: pairwise distance, tetrahedral volume, and external distance. The probability that a type of constraint is applied at each iteration is the fraction of that type of constraints times a bias parameter.

When performing alignment, we add pharmacophore constraints as a fourth type of constraint to choose from during each embedding step. A bias parameter is also provided for the new constraint type, to ensure that pharmacophore constraints get chosen even when they are outnumbered by distance or volume constraints. We also modify the distance constraint type to make it component-aware, i.e., to store the distance matrices only for individual component molecules rather than the whole ensemble, and always pick a pair of atoms from within the same molecule to check for constraint satisfaction.

- *Handling of centroid coordinates*-When a single atomic pharmacophore point is selected for an embedding step, coordinates of the atoms on both molecules are moved towards each other if their distance exceeds the upper bound, or away from each other if it is smaller than the lower bound. When a pharmacophore group with many atoms is selected, its centroid is computed on the fly, and the coordinates of the two matching centroids on the two molecules are used to compute the direction and magnitude of the motion that will be applied to all atoms in both groups.
- *Handling of angular/direction vector motion*-Direction vectors are associated with directional hydrogen-bonding atoms or

planar groups such as aromatic rings that form a pharmacophore point. Since SPE is based on distance geometry and does not incorporate direction vectors or angle constraints, we needed to extend it to handle angular motion. Similar to the minimum and maximum distance bounds attached to a distance constraint, we introduce a minimum (zero) and maximum (default $\pi/18$) angle that each pair of direction vectors can make. Deviations between direction vectors are considered as violations of this constraint only if:

1. Both direction vectors are well-defined; for aromatic rings, this means that all the atoms are almost coplanar, and not almost collinear. For hydrogen-bond donors and acceptors, this means that the bond lengths are nearly correct in both molecules, so the current direction vectors can be compared.
2. The distance between the centroids of both rings is smaller than the upper bound for matching a pharmacophore constraint, i.e., there is no violation of a distance constraint, or the centroid positions nearly coincide. Only in this case are angle violations considered; this resolves the problem of combining pharmacophore distance and angle violations into a single value.

Angular twist motions of a pharmacophore group are applied by rotating each atom about the line passing through their centroid and parallel to the cross product of the two direction vectors, which we call the axis of the rotation. Denote the distance of each atom from the axis as its radius; then multiplying an angle by this radius gives the arc length that each atom must move. We use the arc length of an angle deviation to calculate its equivalent distance motion, and the arc length of the maximum allowed angle as the maximum allowed distance. Then we use the same formula as SPE uses for updating coordinates using distance constraints, to compute the arc length and hence the angle of the final rotation to be applied.

One difference between the way translations and rotations are applied is that rotations are applied fully to only one of the groups of atoms, rather than equally to both groups. The chosen group is the one that occurs later in the constraint file, and hence has a higher index number in the list of angular constraints; this ensures consistency, and minimizes the chances of an alignment failing because of large moves that scatter many quasi-aligned vectors towards one misaligned one.

- *Distance function error evaluation*-Distance geometry methods report the goodness of an embedding by evaluating the extent to which constraints are violated; the value of the distance function error is used to choose good conformations if a force field is not in use, as well as in the refinement postprocessing step. Violations of distance, volume, and external constraints are evaluated as before, and violations of pharmacophore point constraints are evaluated in the same way as for distance constraints. Thus, the violation of a pharmacophore distance constraint is the difference between the squared distance between the pair of atoms or group centroids and the squared maximum distance allowed, expressed as a fraction of the squared maximum distance. Violations of all the con-

straints are squared and summed up to get the final distance geometry error reported. For pharmacophore points or groups that have an associated direction vector, a mismatch between these vectors when all other constraints are satisfied needs to be factored into the function evaluation. Eligibility criteria used are similar to those for angular motion aforementioned (the direction vector is well-defined, and the points already satisfy the distance constraints). Eligible pairs of direction vectors that diverge by more than a maximum amount (default $\pi/18$) are converted to arc lengths and assigned a violation using the same formula as for distances.

- *Distance function gradient*-The postprocessing steps of distance geometry refinement and force-field energy minimization need to compute the gradients of the distance and energy functions at the current values of the embedded coordinates, so that they can push each function towards its minimum. For a pharmacophore constraint whose atoms and groups are farther apart than the allowed maximum distance, its gradient is computed in a similar way to the gradient of a distance function.

We also derive formulas for the gradient of the angle between two direction vectors that are constrained to align with each other. We calculate the gradient at each point that can rotate during the motion that aligns the direction vectors, i.e., each point on the second in the pair of pharmacophore groups being aligned. The formula is derived for groups such as planar aromatic rings whose direction vectors are bidirectional, since they are plane normals and can be lined up with another direction vector in two ways. In such a case, we take the gradient not of the angle θ between these two direction vectors, but of $\alpha = 1 - \cos^2 \theta$, which converges to zero when $\cos \theta$ becomes either 1 or -1 , i.e., the vectors are either parallel or antiparallel to each other. This ensures that the least required motion is applied. The resulting gradient is given by the following formula, where V is a basis vector in the plane of the first ring calculated as the cross product of the axis of rotation and normal vector of the first ring (see derivation in the Appendix):

$$\frac{\partial \alpha}{\partial X} = \frac{2(V \cdot X) \{ (V \cdot X)X - \|X\|^2 V \}}{\|X\|^4}$$

Unidirectional vectors such as directional hydrogen bonds, which can line up with matching vectors in only one way, are modeled by adding an extra point constraint in our current implementation. However, if the angle θ is parameterized in terms of the positions of neighboring atoms, then expressions for the motion, evaluation and gradient of a unidirectional vector can be derived.

Postprocessing and Output

After a run of SPE, the embedded 3D conformations are observed to be nearly in their lowest energy states, and SPE is found to sample low energy conformations if it is run for a sufficient number of steps.⁴⁶ This continues to be the case with the

SPE-based pharmacophore alignment: possible alignments are sampled, and both the conformations and the alignments found look plausible. However, to produce the best possible alignment we use two postprocessing steps: geometric refinement and energy minimization. Since we start from good conformations, both these steps usually converge in 10 or fewer iterations.

Below, we describe the changes made in the postprocessing and output stages of the algorithm.

- *Distance geometry refinement*—The use of a modified gradient function that is aware of molecular components and pharmacophore constraints, as described in Modified Stochastic Proximity Embedding Section, enables the existing refinement code for SPE to work with pharmacophore alignment.
- *Energy minimization*—Apart from a modified gradient function, there are two changes and one performance optimization that need to be made for energy minimization to work with pharmacophore alignment:

1. No electrostatic or van der Waals interaction is allowed between atoms in different components. Such atoms are added to an existing list of atom pairs that do not interact with each other.
2. Rigid realignment is needed to make the pharmacophore points overlap with each other, after being separated by relaxation of the rest of the structure ensemble. Rigid realignment is done by least squares superposition of the corresponding points to recover the optimal translation and rotation. We extend this process somewhat to accommodate the ring normals, which must also remain aligned in the final alignment. We convert each ring normal into an additional single point, 1 Å away from the ring centroid, before alignment. If both directions of the ring normal were to be converted into points, there would be a combinatorial problem of matching 2^N combinations for this pair of points to correctly align the ring if it occurs in N structures. Thus, we place only one point for each ring, whose direction was initially chosen as the side of the ring that faces away from the centroid of all pharmacophore points. This worked to some extent because of our assumption that the geometries being realigned after minimization were already prealigned, and during minimization the rings did not flip over or otherwise change orientation *vis-à-vis* the centroid of all points. However, we did observe rare cases where the wrong orientation was chosen, leading to parallel rather than aligned rings. Thus, an additional and stronger criterion was chosen that works even in cases where the rings change orientation around the centroid of all points; *viz.*, to look at the curvature of a series of three successive pharmacophore points that includes the centroid of the ring being aligned. Since the pharmacophore points are already aligned, one would expect this curvature to be the same across molecules, as is indeed observed. Thus the point corresponding to the normal is placed consistently on the inner or outer side of the ring based on local curvature, *i.e.* by checking the sign of the dot product of the normal with the cross product of vectors joining pairs of those three points, and flipping the normal if necessary.

3. To improve performance of the minimization on large molecular ensembles being aligned, we break up the ensembles and submit each molecule for independent minimization. Essentially this is the same thing that happens when we try to minimize the entire ensemble, since atoms in different molecules do not interact energetically; however, the time complexity of the problem, and the number of steps required for convergence, are much lower for minimizing several smaller molecules than one large ensemble. Step 2 (rigid realignment) is then run on the separately minimized molecules. Currently this optimization is triggered if we are aligning more than three molecules and running minimization as a postprocess.

- *Output of molecules*—Instead of outputting the entire ensemble as a single molecule, the components are split up and written individually in an SD file.

Pharmacophore Extraction and Rendering

The alignment based on pharmacophores can be visualized as such, or converted to the conventional representation of a pharmacophore as a set of spheres at certain points or regions. In our representation, such spheres are placed at pharmacophore points, and are assigned radii indicating the RMS deviation in the coordinates of that point with respect to the centroid of all instances of that point in the alignment. The radii of spheres corresponding to hydrophobic regions and aromatic rings are increased by the average distance between each representative centroid point and the atoms it represents. In addition, aromatic rings are planar and have two normal vectors corresponding to the top and bottom faces of their plane.

We now describe the pharmacophore sphere extraction and visualization process in more detail. To compute the point radii, first we calculate the RMSD for each pharmacophore point's coordinates within each trial, with respect to the centroid of these coordinates. Before the coordinates can be averaged between trials, we have to determine which trials produced unique aligned conformations that should be considered separately. For this, a vector of pairwise distances between pharmacophore points or centroids between trials is built, and vectors are considered unique if and only if they differ in two or more pairwise distances by more than 15%. This criterion is an attempt to ensure robustness, in distinguishing truly unique alignments that change several pairs of distances from alignments that differ in only one pairwise distance or by less than 15%, which are observed to be essentially similar.

To calculate the radii of pharmacophore spheres within a set of unique trials, first the sets of averaged pharmacophore coordinates from each trial are rigidly aligned using least squares superposition. Then the RMSD of each pharmacophore point in this cross-trial rigid alignment is computed with respect to its centroid, to which is added the maximum of the earlier computed RMSDs for that point in any of the trials. Thus, if \mathbf{x}_{ijk} denotes the position of the i -th pharmacophore point on the j -th molecule in the k -th trial, where $i \in \{1 \dots N_{\text{phore}}\}$, $j \in \{1 \dots N_{\text{mol}}\}$, and $k \in \{1 \dots N_{\text{trial}}\}$, $J_i^1 \dots J_i^{N_{\text{mol}}}$ denotes the indices of

the N_{mol}^i molecules in which the i -th pharmacophore occurs, and

$\mathbf{z}_{\text{avg}}^*(i, k) = \frac{1}{N_{\text{mol}}^i} \sum_{j=j_1^i}^{N_{\text{mol}}^i} \mathbf{x}_{ijk}$ denotes the averaged position of the i -th point in the k -th trial over all molecules where it occurs, then we can define the averaged superimposed position as:

$$\mathbf{z}_{\text{avg}}(i, k) = \begin{cases} \mathbf{z}_{\text{avg}}^*(i, k), & \text{if } k = 1 \\ \text{superimpose} \left(\mathbf{z}_{\text{avg}}^*(i, k) \text{ on } \mathbf{z}_{\text{avg}}^*(i, 1) \right), & \text{if } k > 1 \end{cases}$$

from which:

$$\text{rmsd}(i, k) = \sqrt{\frac{1}{N_{\text{mol}}^i} \sum_{j=j_1^i}^{N_{\text{mol}}^i} (\mathbf{x}_{ijk} - \mathbf{z}_{\text{avg}}^*(i, k))^2}$$

$$\text{center}(i) = \frac{1}{N_{\text{trial}}} \sum_{k=1}^{N_{\text{trial}}} \mathbf{z}_{\text{avg}}(i, k)$$

$$\text{radius}(i) = \sqrt{\frac{1}{N_{\text{trial}}} \sum_{k=1}^{N_{\text{trial}}} (\mathbf{z}_{\text{avg}}(i, k) - \text{center}(i))^2} + \max_{k=1}^{N_{\text{trial}}} (\text{rmsd}(i, k))$$

Pharmacophore points representing aromatic rings are visualized in a slightly different way from those representing atoms or hydrophobic groups. Two additional quantities to be visualized include the averaged symmetric normal vector of each ring and its deviation within each trial and across trials. The aggregation of ring normal vectors differs from that of point coordinates in two key ways. First, the average computed on the normal vectors is a geodesic average on their representation as points on a unit sphere, which is computed using the successive weighted averaging technique described by Nathan Reed on the DevMaster.net forum.⁴⁸ Just as the average of a series of N numbers can be written as a recursive weighted pairwise average, $\text{average}(x_1 \dots x_N) = ((N-1) \times \text{average}(x_1 \dots x_{N-1}) + x_N)/N$, similarly the average of N vectors can be computed by successively finding $\text{average}(x_1 \dots x_i)$ as the point on the unit sphere that subdivides the arc between $\text{average}(x_1 \dots x_{i-1})$ and x_i in the ratio $(i-1):1$. The deviation of the vectors within each trial is also measured within the unit sphere, as the mean distance between the average normal vector and each contributing vector in a direction perpendicular to the average normal:

$$\text{hasNormal}(i, k) = \{j \in \{1 \dots N_{\text{mol}}\} | \mathbf{x}_{ijk} \text{ has a normal}\}$$

$$\text{angdev}(i, k) = \frac{1}{|\text{hasNormal}(i, k)|} \times \sum_{j \in \text{hasNormal}(i, k)} \left\| N_{ijk} - N_{ijk} \cdot \text{geoNorm}(N_{ijk}) \right\|$$

Using the same technique as in the rigid realignment after minimization, these averaged normal vectors are then represented as

proxy points, one for each ring, placed in the direction pointing out from the centroid of all points towards the centroid of the ring. The rotation and translation found during the least squares superposition of pharmacophore points between trials is then applied to this proxy point, and the new normal vector is inferred by subtracting the transformed centroid from the transformed proxy point.

Finally, the transformed normals are geodesically averaged across trials, and deviation between them across trials is found analogously to point coordinates, as the RMSD of transformed average normals added to the maximum deviation seen in any trial. Denoting the transformed normals as N_{ik}^* , we get:

$$\text{angdevtrials}(i) = \frac{1}{N_{\text{trial}}} \times \sum_{k \in \{1 \dots N_{\text{trial}}\}} \left\| N_{ik}^* - N_{ik}^* \cdot \text{geoNorm}(N_{ik}^*) \right\| + \max_{k=1}^{N_{\text{trial}}} (\text{angdev}(i, k))$$

To visualize each normal vector and its deviation, two techniques are used. First, the normal vector is surrounded by a cone whose apex half-angle α is taken such that $\sin \alpha$ equals the deviation (which can never exceed 1). Thus, the deviation is expressed as the radius of the circular cross-section of the cone where it crosses the unit sphere. This cone and vector is drawn on both sides of the sphere, since the vector is bidirectional. Also in parallel, the pharmacophore sphere radius is scaled down in the direction of the normal vector by an amount equal to its deviation. Thus, an aromatic ring that is nearly perfectly aligned across trials is drawn as a flat disc with thin cones surrounding both normal vectors; a badly aligned one that distributes normal vectors roughly equally in angular space will show up as a spheroid with aspect ratio $\sin(45^\circ) \approx 0.71$ and cones with 45° . Since plane normals are treated as bidirectional when taking geodesic averages, larger deviations tending to 1 are avoided.

Datasets for Validation

To test our pharmacophore alignment method, we provide several cases where pharmacophores have been derived from the literature, some where the bound conformations are known from the structures of some of the ligands in a protein binding site, and some where the method samples multiple known alignments or bound conformations consistent with each selected pharmacophore. The selected datasets of compounds are summarized in Table 1.

Adamantane Fragments

This is a synthetic case created to test the molecular alignment functionality, where the pharmacophore points chosen have no chemical or therapeutic significance. The adamantane molecule, which has a bridged cyclic structure, is split into two components, the cyclohexane ring and the Y-shaped 3-ethylpentane. Individually, cyclohexane can adopt the chair, boat or twist-boat conformations (with the boat and twist-boat being significantly higher in energy, but geometrically distinct), and 3-ethylpentane

Table 1. Numerical Summary of Performance of the Alignment Program on Pharmacophore Case Study Datasets.

Dataset	Molecules	Features	RMSD	Error	T_{spe}	T_{min} (s)	Ref. Steps	Min. Steps
Opioids	14	3	0.44	23–307	0.6s	11	3–5	100–500
P-gp binders	7	4	1.65	12–1418	0.4s	15	2–27	175–800
Dopamine receptor substrate/inhibitors	8	3	0.81	8–46	0.3s	2.5	1–23	80–350
HIV-1 protease inhibitors	15	4	0.98	922–437,360	0.8s	112.5	23–200	150–1000
HIV-1 protease structure	15	17 ^a	0.99	148–165,077	0.7s	99	154–200	171–1000
HIV-1 integrase inhibitors	5	3	0.83	7–16	0.2s	3.5	4–86	150–400

The columns are as follows: dataset name, number of molecules, number of pharmacophore points, the largest RMSD for any pharmacophore point in the final alignment (RMSD), the range of distance constraint errors (which includes pharmacophore constraint errors, hence can be huge), the average time per embedding using SPE (T_{spe}), the average time per embedding including DG refinement and energy minimization (T_{min}), and finally the range of iterations taken for DG refinement and energy minimization to converge (Ref. Steps and Min. Steps, capped at 200 and 1,000, respectively).

^a4 ligand-based from previous model; 13 structure-based derived from steric and H-bond interactions shown in Figure 5 from Wang et al.⁴⁹

can adopt several conformations with specific chiral configuration at the branching point. Chirality is made visible by attaching a different halogen atom to each endpoint and the branching point of the chain, as shown in Figure 1. However, when they are aligned with adamantane by considering the carbons with the same halogen substituents as pharmacophore points, the chair (and occasionally, twist-boat) conformation of cyclohexane and two compact chiral conformations of 3-ethylpentane are produced. This reduces to one conformation (aligned with adamantane) when an additional constraint is added to tie the branching point of the Y in adamantane and 3-ethylpentane. The substitution of hydrogens by halogen atoms to mark corresponding pharmacophore points does not change the geometry of adamantane or its components, as the halogens are covalently bonded to carbons and cannot have ionic or hydrogen bond interactions with any other atom.

The aligned conformation shown is attained even without distance or energy minimization, both of which converge in just a few iterations (typically 2–10 for distance, and 10–50 for energy).

Opioids

The compound morphine derived from opium poppies is a potent analgesic, but is also highly addictive and damages the nervous system. Morphine achieves its analgesic effect by binding to a set of native receptors in the body, called the opioid receptors. Deriving potent painkillers that also bind to an opioid receptor but have fewer side effects and addictive tendencies is an active area of cheminformatics research. Opioid receptors are often membrane proteins, without determined structures, and thus drug design targeting opioid receptors is a classic application of ligand-based pharmacophores.^{50–55} Pharmacophores have been developed that attempt to capture the analgesic activity of opioids without most of their undesirable side-effects and addictiveness.⁵⁶ Also, several authors have shown that different 3D pharmacophores specifically account for binding to opioid receptor sub-types such as μ , δ , and κ .^{52,57–64} This makes the elucidation

of a pharmacophore for opioid activity a challenging task. Bernard et al.³⁰ among others, have studied the opioid pharmacophore by a computational method that considers the overlap of probability distributions of distances and angles between groups, and showed that several different 3D pharmacophores are feasible.

For our case study we considered 13 opioids: naturally occurring morphine, codeine, and synthetic heroin (acetylated morphine) that are analgesic but also addictive and toxic to some degree; MET-enkephalin and norepinephrine, respectively a peptide and a small molecule that are endogenous ligands of the body's opioid receptor; and eight synthetic morphine derivatives, all analgesic, some potentially toxic/addictive and some not, as shown in Figure 2. Morphine occurs twice in our dataset as it is represented by two SMILES from two different sources; though included inadvertently, the duplicate morphine was left in our dataset since it is of interest to see if the two copies attain the same conformation and chirality under pharmacophore constraints.

The pharmacophore tested consists of an aromatic ring, a hydroxyl group attached to the ring, and basic nitrogen (highlighted in Fig. 2). According to Messer,⁵⁶ this is the essential scaffold of chemical features for opioid activity, and all other groups including the hydrophobic ring system near the basic nitrogen can be dispensed with and the analgesic activity retained. In fact, several compounds such as fentanyl, meperidine/demerol, and methadone lack even the hydroxyl group and still retain analgesic activity, so we allow this group to be missing in these compounds.

P-Glycoprotein Binders

The human phosphoglycoprotein (P-glycoprotein, P-GP) functions as a cellular efflux pump for drugs and other toxic or foreign substances, and helps in maintaining the blood-brain barrier, among other functions. Pearce et al.^{65,66} first described the P-glycoprotein binding pharmacophore in a series of analogs of the natural product alkaloids reserpine and yohimbine that

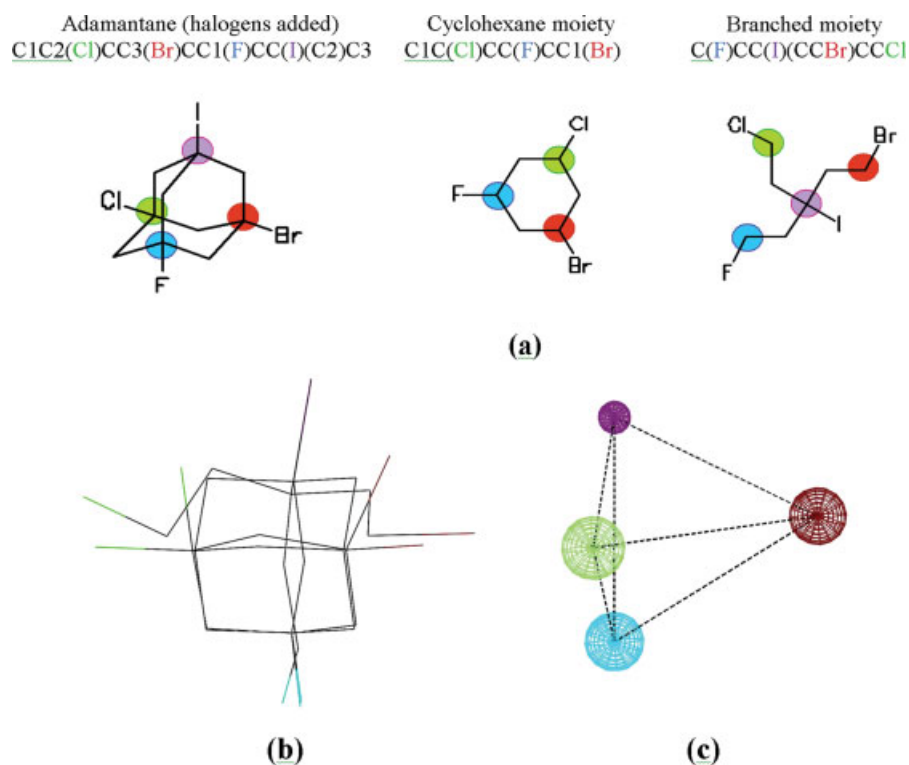


Figure 1. (a) Substituted adamantane and its fragments; (b) unique alignment of the fragments; (c) the derived “pharmacophore,” which is nearly independent of different random number seeds and parameters such as whether refinement and minimization were used; alignment shown includes refinement and minimization selected.

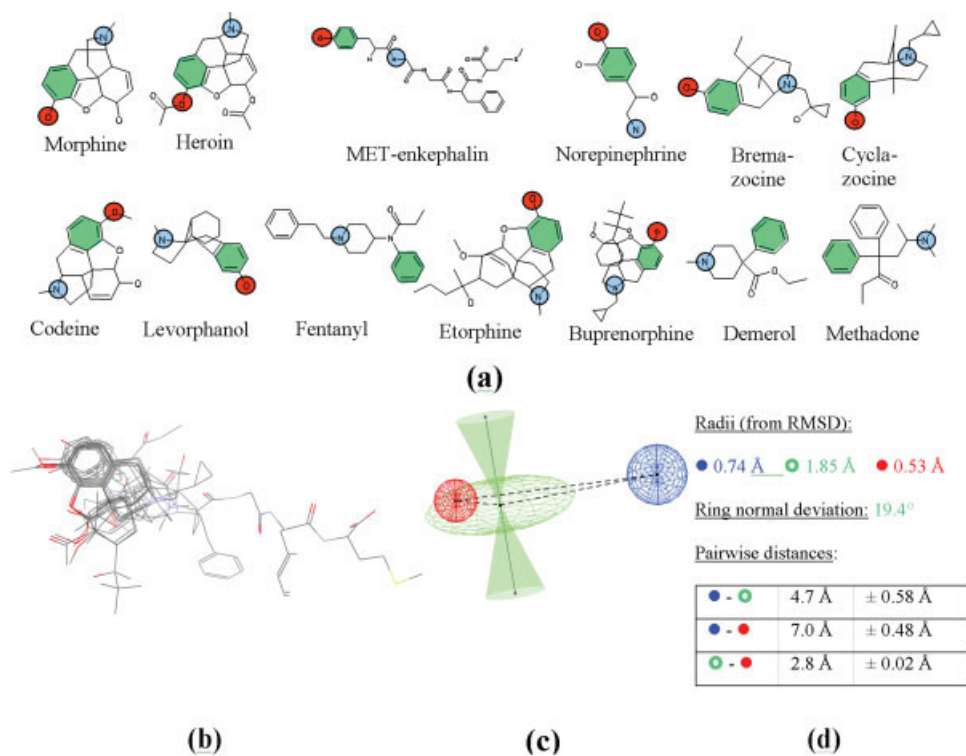


Figure 2. (a) The morphine analogs dataset; (b) representative alignment (after refinement and minimization); (c) pharmacophore derived from the alignment; (d) measurements of pharmacophore point radii and pairwise distances.

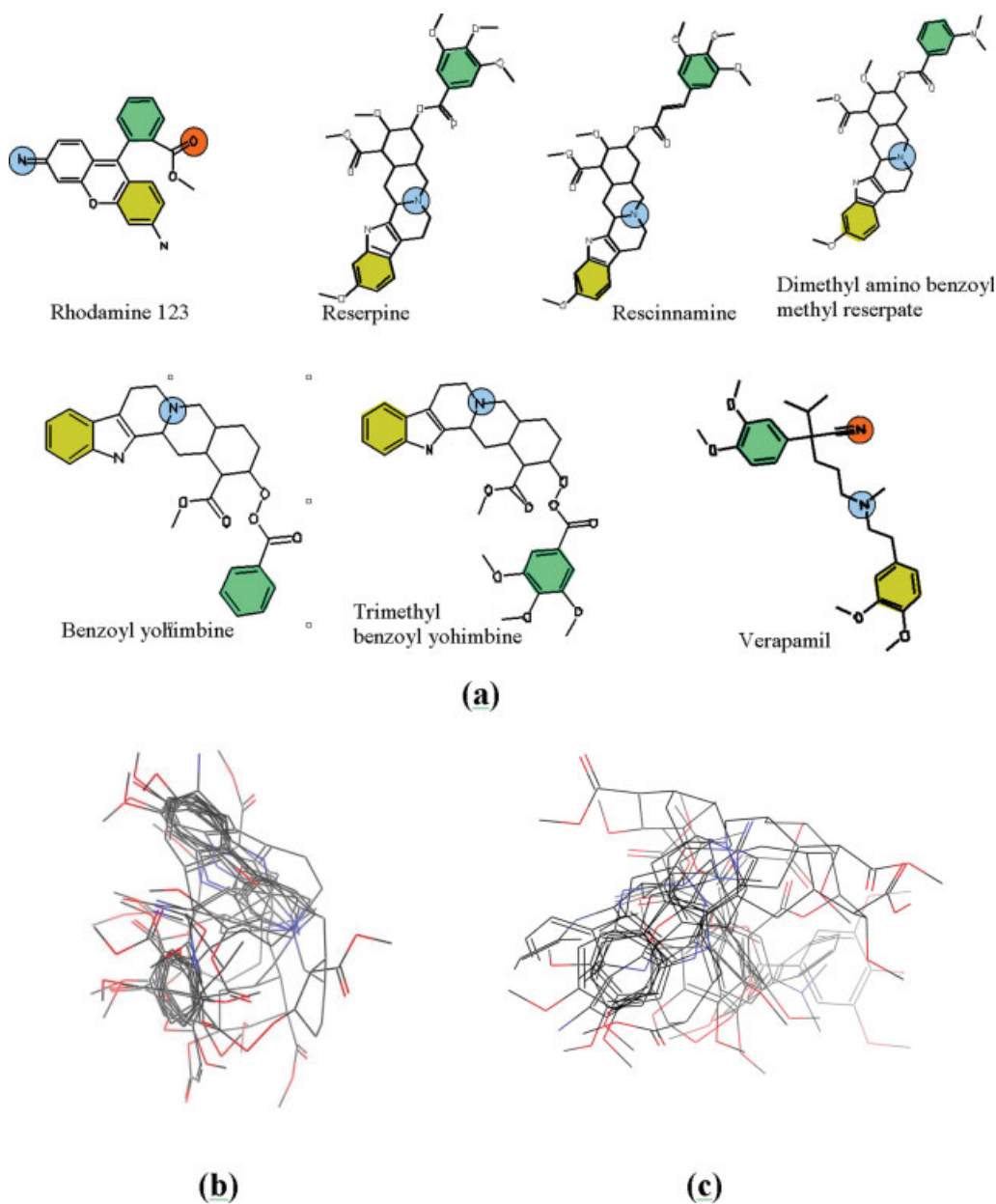


Figure 3. (a) The P-glycoprotein binder dataset; (b) alignment after refinement; (c) alignment after refinement and minimization. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

inhibit P-glycoprotein activity. Since then, several other inhibitor series have been studied,⁶⁷ and several refined pharmacophore models have been proposed.^{68–72}

We consider the seven P-GP binders shown in Figure 3, with five of them from the reserpine and yohimbine analog series of Pearce et al.,⁶⁵ and two (verapamil and rhodamine 123) being actives from a different series studied by Varma and Hou⁶⁸ and others. The pharmacophore for P-GP binding integrates these two datasets, incorporating the two aromatic rings and a basic nitrogen group found from the reserpine/yohimbine, and

a hydrogen bond acceptor occurring in verapamil and rhodamine 123.

Dopamine Analogs

Dopamine is a brain neurotransmitter with many functions, the most well-known among them being mediation of the reward system, which causes a sensation of pleasure when we repeat behaviors that have been beneficial in the past. Dopamine binds to a series of dopamine receptors in the brain, which are also

activated by some narcotic substances such as cocaine (from coca leaves), lysergic acid (from ergot fungus), and LSD (lysergic acid diethylamine, a synthetic derivative). Dopamine is also a precursor to the opioid hormone norepinephrine; apomorphine, a morphine derivative, also binds to dopamine receptors. Two more dopamine receptor agonists that we include in the list of actives are methylphenidate, sold as the attention deficit/hyperactivity disorder (ADHD) drug Ritalin, and mazindol, used as an appetite suppressant and cocaine addiction therapy before being recently discontinued.

The pharmacophore for nonselective dopamine receptor activity contains an aromatic ring, an H-bond donor/acceptor adjacent or proximal to the ring (hydroxyl group in dopamine), and a positively charged H-bond donor (basic nitrogen in dopamine).⁷³

HIV-1 Inhibitors

The HIV-1 protease (HIV1p) is one of the most studied and best represented protein structures in the PDB; crystal structures are available with HIV1p bound to several protease inhibitors (PIs). Several pharmacophore models were proposed for HIV1p inhibitors. Wang et al.⁴⁹ postulated a three-point model with one carbonyl H-bond acceptor and two hydroxyl H-bond donors. Pandit et al.⁷⁴ hypothesized a four-point model with one hydroxyl H-bond donor neighbored by one hydrophobic and two aromatic groups, with excluded volumes from the receptor structure and partial matching of one among four features (two HBAs, one HBD, one hydrophobic) added for specificity. Using multiple protein crystal structures and MD simulation, Meagher et al.⁷⁵ were able to derive a “consensus of consensus” pharmacophore with eight points: two H-bond donors near the catalytic aspartates, four aromatic/hydrophilic groups around the periphery of the active site, and two aromatic/hydrophobic groups near the center of the active site.

In modeling the HIV1p pharmacophore, we compare the three-point ligand-based pharmacophore model and the structure-based model with hydrogen-bonding and excluded volume constraints from Wang et al.⁴⁹ External coordinate constraints are used to model reported proximities to active site residues, and also exclusion spheres derived from them.

For HIV-1 integrase (HIV1i) inhibitors, the earliest pharmacophore proposed by Nicklaus et al.⁷⁶ comprised a H-bond acceptor (carbonyl O in Caffeic Acid Phenethyl Ester, CAPE, the first natural product HIV1i inhibitor found) interacting with two more acceptors (O or N) and shielded by an excluded volume sphere around the carbonyl carbon in CAPE. An updated pharmacophore model⁷⁷ consisting of a hydrophobic aromatic group, two hydrogen-bond acceptors and one donor helped identify several novel inhibitors.^{78,79} A dynamic pharmacophore model of HIV-1 integrase inhibition was produced by Deng et al.⁸⁰ by combining pharmacophore modeling with MD conformational sampling of the protein.

Results

We summarize the alignments of the test datasets found by our algorithm along with the derived pharmacophores in Figure 2

through Figure 9. We also draw these pharmacophores within the same figures, as well as quantify them by listing measurements of the derived pairwise distance values and their standard deviations, the pharmacophore sphere radii, and the angles made by and uncertainties in ring normal vectors. We discuss the contents of the derived pharmacophores and measurements in some detail, comparing them to what is known from the literature. We give some metrics of the algorithm’s performance for the test cases, including distance error refinement and energy minimization.

Opioids

As seen in Table 1, the morphine analog/opioid pharmacophore-based alignment was very sharp, with at most 0.44 Å RMSD within any of the aligned pharmacophore points when averaged across all trials. The alignment of the rings and ring normals was almost perfect after refinement, and deviated a bit more after minimization. The extracted pharmacophore consists of the hydrogen bond donor/acceptor hydroxyl (colored red), lying on or slightly above the plane of the aromatic ring (colored light green), whose normal deviates on average by 19.4° from its mean position within the alignment (see Fig. 2). The basic nitrogen atom (colored blue) is somewhat above this plane and at a distance of 4.7 Å from the ring centroid and 7.0 Å from the hydroxyl oxygen.

The pharmacophores found match well with what is known in the literature. For example, the specific pharmacophore for δ -opioid binding proposed by Loew and coworkers⁵² contains the aromatic ring, basic nitrogen and a hydrophobic group, and distances between the ring centroid and nitrogen are 4.5 Å. The distances and orientations between the aromatic ring and its attached hydroxyl correspond to their covalent bonding to each other.

As an aside, we also notice that in all trials that we considered, the two copies of morphine in the dataset superimposed almost exactly, with the correct chiral conformation. This validates the underlying conformer generation algorithm of SPE, which produces reasonable embeddings of a highly chiral molecule by enforcing the volume constraints, and the implementation of the newly added pharmacophore constraints, which do not break the correct chirality.

P-glycoprotein Binders

P-glycoprotein, being a cellular multi-drug efflux pump, is known to have a large and flexible active site that can adapt to bind a large variety of substrates. Its promiscuity and flexibility imply that its many substrates bind in several different binding modes; indeed, more than one pharmacophore hypothesis has been proposed with nearly equal weight,⁶⁸ and ensemble⁸¹ and conformationally-sampled³⁰ pharmacophores have also been proposed to accommodate the plurality of binding modes in P-GP. In this context, the finding of three unique and valid configurations from 10 trials with refinement, and two unique conformations after refinement and minimization, supports the heterogeneity of this pharmacophore (Figs. 3–5). The pharmacophores after refinement and minimization have large deviations in the aligned coordinates and in the normal vectors of aligned rings (over 45°), which hints that even the alignments being merged because

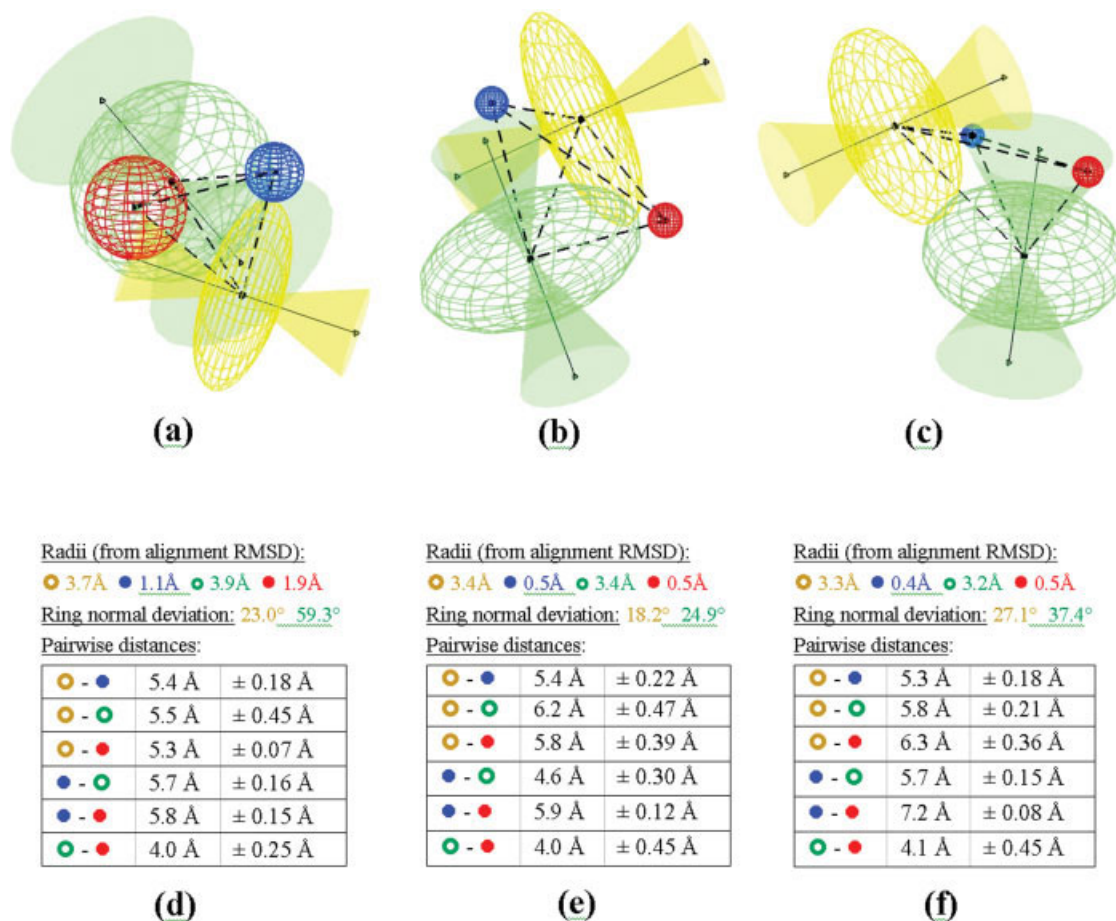


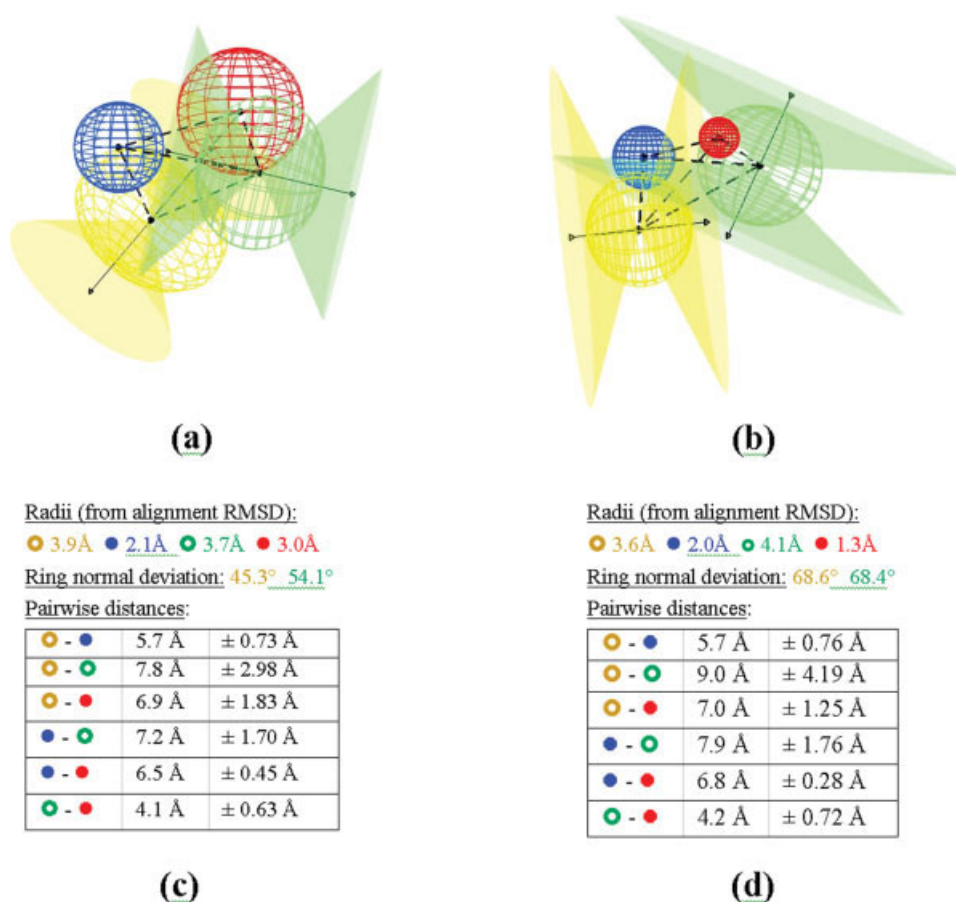
Figure 4. P-glycoprotein binder pharmacophores derived from the alignment in Figure 3(b) obtained when SPE was run with refinement. There were five unique pharmacophores found in 10 trials, two of which were filtered out because of overlapping pharmacophore spheres. In (a)–(c) we show the remaining pharmacophores, their point radii, and pairwise distances. Notice that (b)–(c) are much better pharmacophores than (a), but all three unique pharmacophores conserve the distances between the yellow aromatic ring and the blue basic nitrogen, and between the green aromatic ring and the red H-bond acceptor. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

of similar pairwise distances are really distinct, and that energy-minimized alignments do not need to have the rings aligned. Though the different unique alignments have different values for most pairwise distances, the distance between the green aromatic ring and the red H-bond donor/acceptor, and that between yellow aromatic ring and the blue basic nitrogen, stays relatively fixed across all alignments. This indicates the invariance of these distance pairs across the different pharmacophore models proposed.

Dopamine Analogs

The dopamine receptor binders considered in our dataset, both substrates and inhibitors, were strongly aligned using the aromatic ring, attached hydrogen bond donor/acceptor and distal basic nitrogen as shared features (see Fig. 6). The alignment was

almost perfect when only distance geometry refinement was used for postprocessing, with sphere radii under 0.5 Å and aromatic ring normal deviation of 6.7°; this increased to 18.1° after energy minimization, with a corresponding threefold increases in sphere radii for the two hydrogen-bonding pharmacophore points. Overall the pharmacophore fit the compounds being aligned well, though they were a mix of both substrates and inhibitors, for which different pharmacophores are reported in the literature. For dopamine receptor substrates, the model mentioned by Kubinyi⁷³ does not enumerate these distances, but the numbers we derive do seem plausible when inserted into their model. For dopamine uptake inhibitors, the pharmacophore model developed by Wang et al.⁸² specifies a nitrogen atom, aromatic center and carbonyl group, and the distance between the nitrogen and the aromatic ring in that model is 5–7 Å (we find it to be 5.0 ± 1.19 Å). The carbonyl group serves as the H-bond



Radii (from alignment RMSD):

● 3.9Å ● 2.1Å ● 3.7Å ● 3.0Å

Ring normal deviation: 45.3° 54.1°

Pairwise distances:

● - ●	5.7 Å	± 0.73 Å
● - ●	7.8 Å	± 2.98 Å
● - ●	6.9 Å	± 1.83 Å
● - ●	7.2 Å	± 1.70 Å
● - ●	6.5 Å	± 0.45 Å
● - ●	4.1 Å	± 0.63 Å

Radii (from alignment RMSD):

● 3.6Å ● 2.0Å ● 4.1Å ● 1.3Å

Ring normal deviation: 68.6° 68.4°

Pairwise distances:

● - ●	5.7 Å	± 0.76 Å
● - ●	9.0 Å	± 4.19 Å
● - ●	7.0 Å	± 1.25 Å
● - ●	7.9 Å	± 1.76 Å
● - ●	6.8 Å	± 0.28 Å
● - ●	4.2 Å	± 0.72 Å

Figure 5. P-glycoprotein binder pharmacophores derived from the alignment in Figure 3(c) obtained when SPE was run with refinement and minimization. There were four unique pharmacophores found in 10 trials, two of which were filtered out because of overlapping pharmacophore spheres. In (a)–(b) we show the two remaining pharmacophores, and in (c)–(d) their point radii and pairwise distances. Notice that both unique pharmacophores conserve the distances between the yellow aromatic ring and the blue basic nitrogen, and between the green aromatic ring and the red H-bond acceptor, and these distances are similar to those in Figure 4. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

donor/acceptor (red) in our pharmacophore for cocaine and methylphenidate, and thus the distances between this group and the ring (3.4–6.1 Å) and nitrogen (2.2–4.5 Å) specified by Wang et al. are close to what we determine (3.2 ± 0.66 Å to the ring, and 5.5 ± 0.76 Å to the nitrogen).

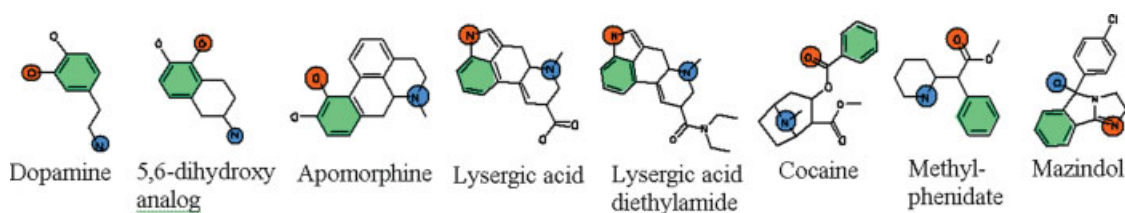
HIV-1 Integrase

The pharmacophore for HIV-1 integrase inhibition developed by our program (see Fig. 7) has the two O/N acceptors closer together at a distance of 2.53 ± 0.14 Å, with distances to the carbonyl O acceptor as 7.5 ± 1.86 and 7.2 ± 1.48 Å; this range matches or partially overlaps the respective distance ranges of 2.548 ± 0.3 Å, 9.053 ± 0.4 Å, and 8.711 ± 0.4 Å mentioned by Nicklaus et al.⁷⁶ Their model also contains an exclusion sphere centered on the carbon atom of the carbonyl group from

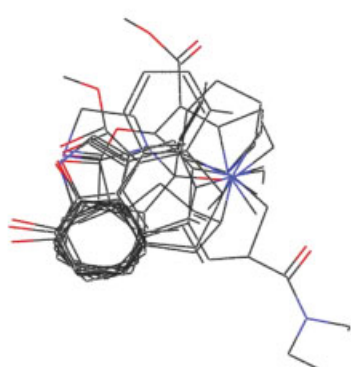
CAPE, to ensure the orientation of the carbonyl group; this is an internal excluded volume constraint that is not supported by our current system, but can be easily incorporated in the future.

HIV-1 Protease

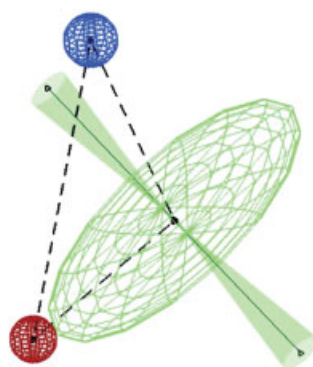
The HIV-1p inhibitors in our dataset are aligned using two sets of pharmacophores: ligand-based and structure-based. The ligand-based pharmacophore comprises four constraints: two for the directional carbonyl group/H-bond acceptor, and one each for the hydroxyl groups/H-bond donors. There is some ambiguity about which hydroxyl corresponds to which other one, since Wang et al. do not differentiate them.⁴⁹ This can be resolved either during the pharmacophore point assignment pre-processing step, or using the heuristic that the closer hydroxyls or the ones on the same ring system as the carbonyl are in correspon-



(a)



(b)



(c)

Radii (from alignment RMSD):

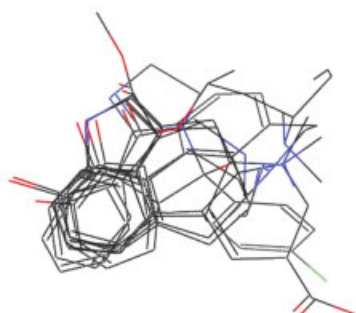
● 0.36Å ● 2.54Å ● 0.42Å

Ring normal deviation: 6.7°

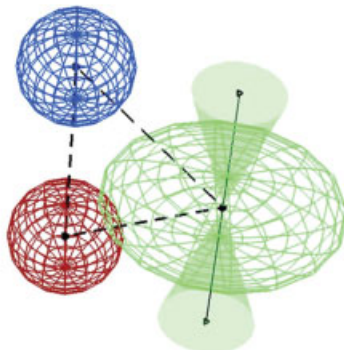
Pairwise distances:

● - ●	3.0 Å	± 0.19 Å
● - ●	5.2 Å	± 0.18 Å
● - ●	4.3 Å	± 0.17 Å

(d)



(e)



(f)

Radii (from alignment RMSD):

● 1.02Å ● 2.20Å ● 1.05Å

Ring normal deviation: 18.1°

Pairwise distances:

● - ●	3.2 Å	± 0.66 Å
● - ●	5.5 Å	± 0.76 Å
● - ●	5.0 Å	± 1.19 Å

(g)

Figure 6. Dopamine receptor substrate/inhibitor dataset. Top row: (a) eight active compounds, corresponding groups colored red (H-bond donor/acceptor), green (aromatic), or blue (H-bond donor); Middle row: (b) alignment after refinement; (c) pharmacophore; and (d) measurements from (b); Bottom row: (e) alignment after refinement and minimization; (f) pharmacophore, and (g) measurements from (e). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

dence (“near” hydroxyls, and equivalently for the “far” hydroxyls). There are only a couple of cases where it is not obvious in 2D what the near and far hydroxyls should be, and these we assigned by trial and error, or intuition. A general solution to matching either of multiple possible candidates for a correspondence involving the detection and reinforcement of emergent

interactions was earlier explored by Agrafiotis and Xu in the context of protein docking (unpublished data), but did not work reliably in all cases. Further work in this area is ongoing, and will be presented in due course.

Starting with the above ligand-based pharmacophore points, the structure-based pharmacophore adds hydrogen-bonding

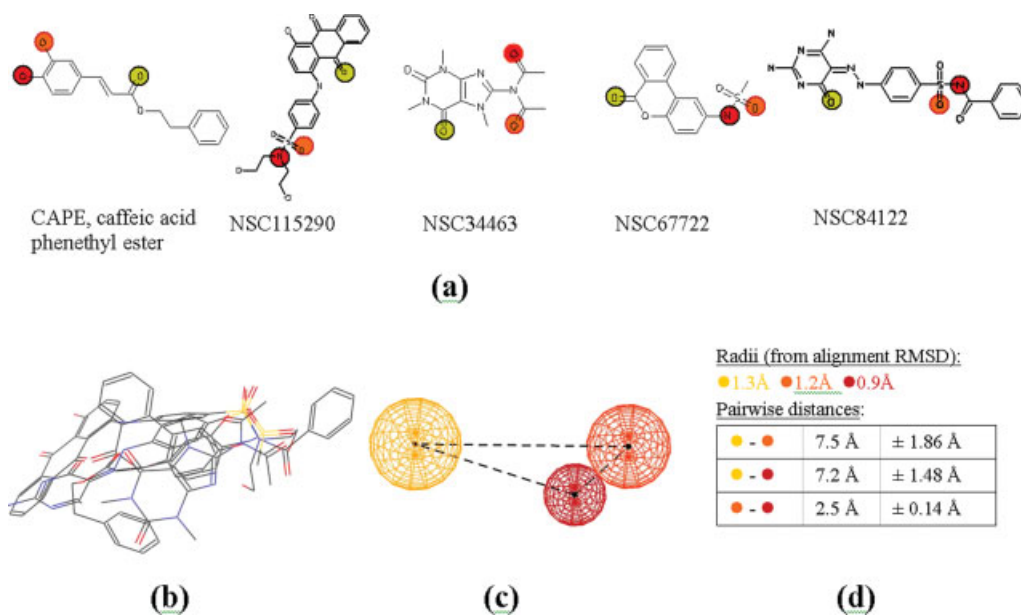


Figure 7. (a) Some HIV-1 Integrase inhibitors, including natural compound CAPE (caffeic acid phenethyl ester), and synthetic compounds identified in various screens as listed in the text; (b) alignment after refinement and minimization; (c) extracted pharmacophore; (d) measurements. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

contacts and excluded volume spheres around the key interacting residues of the active site. The excluded volumes are added to deny specific interacting atoms on one of the small molecules (NSC32180) access to the steric volume of specific interacting residues on the protein, rather than denying it to all atoms of all ligands automatically. The latter choice is closer to what many people think of as excluded volumes, and it needs an easy extension of the current system to implement them. The 15 external constraints are given a frequency bias of 16, i.e., are 16 times more likely to be picked than based on their relative count alone, to give them parity with distance, volume and pharmacophore constraints.

The alignment and extracted pharmacophores and measurements for the ligand and structure-based pharmacophores, shown in Figures 8 and 9 respectively, show remarkable similarity. The main difference, not visible in these figures, is that the different trials of a structure-based alignment have nearly the same orientation and fill the same region of space (i.e., the active site), while the ligand-based alignment trials tend to have more diverse orientations and are spread through space. Both alignments result in nearly the same distance and radii measurements for the developed pharmacophores. External constraints do improve the alignment, but the underlying pharmacophore can still be determined using no protein structure information.

Discussion

As seen in Table 1, the embedding and alignment part of our method is very fast, though the process takes significantly longer when energy minimization is introduced. The quality of the

aligned conformations generated without energy minimization is acceptable in a few cases (e.g. the morphine and dopamine analogs), but the majority of cases need energy minimization to generate realistic geometries for the compounds being aligned. We achieve some speed-up by minimizing all molecules in the alignment separately and then realigning them. The need for energy minimization can be obviated if smaller rigid parts of each molecule are embedded in pre-minimized geometries, and rearranged at flexible regions such as rotatable bonds during embedding. Such a fragment-based embedding approach called Self-Organizing Superposition (SOS) has been developed in our group,⁸⁵ and can be combined with the method presented in this paper to produce a pharmacophore alignment that does not need to be minimized. SOS is nearly an order of magnitude faster than SPE, and produces much better geometries.

It is notable that the method can accommodate both receptor and ligand-based pharmacophore constraints, or a combination of the two. The framework of conformational analysis based on SPE is broad enough, and robust enough, that a new “pharmacophore” distance constraint type can be added to adapt it for alignment rather than conformer generation, without major changes to the current embedding algorithm. Thus, we demonstrate the extensibility of the SPE algorithm to new problems amenable to a self-organizing solution. We are able to cast the pharmacophore development/molecular alignment problem in terms of conformational analysis, and thus take advantage of a fast algorithm with high sampling capacity (SPE) that exists for conformational analysis. It has been previously demonstrated that the SPE algorithm samples conformational space well^{46,84} and by extension we can claim that the present algorithm should sample “alignment space” well.

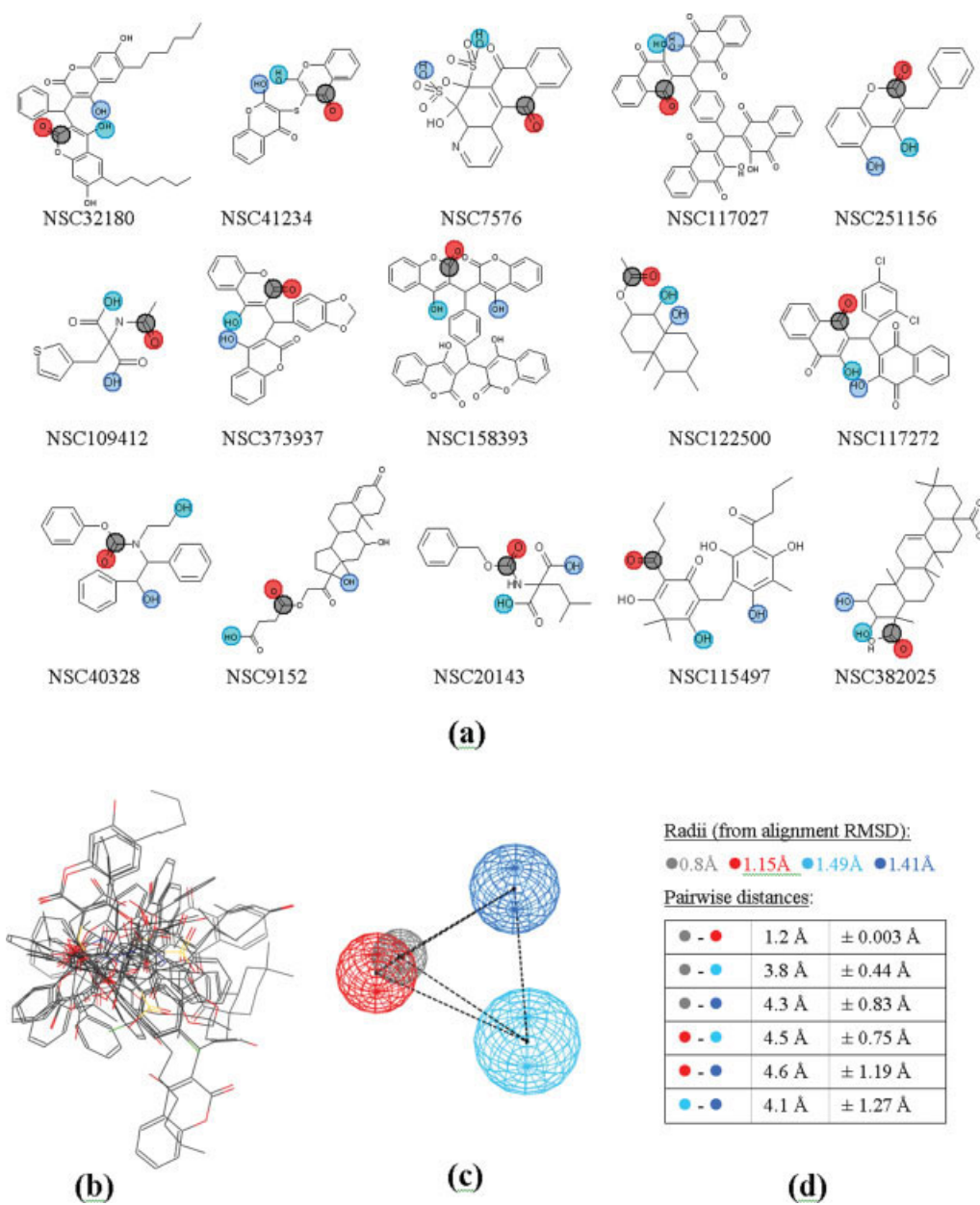


Figure 8. (a) 15 HIV-1 protease inhibitors discovered in a 3D search, from Wang et al.⁴⁹ The pharmacophore is specified as correspondences between a carbonyl, along with a near and a far hydroxyl group on each compound. No structural constraints from the active site are used; (b) sample alignment with this correspondence; (c) extracted pharmacophore; and (d) measurements. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The results presented on well-studied cases serve as a validation of the method's accuracy for aligning small and medium-sized sets of chemical compounds using a variety of constraints. We do not present an independent validation of the derived pharmacophores using 3D database search queries, or by comparison with other pharmacophore development algorithms. The focus of this paper is on the methodology, and such large validation or comparative studies will form the subject of a future article.

The alignment program presented has a requirement that correspondences between pharmacophore groups be well-defined across each molecule in the alignment. Such correspondences can be resolved prior to the program being run, by defining all the chemical groups of each type in each molecule, and formulating a hypothesis (exhaustively, if necessary) as to which groups in one molecule correspond to which ones in others. It is tempting to allow for many-to-one or many-to-many matching,

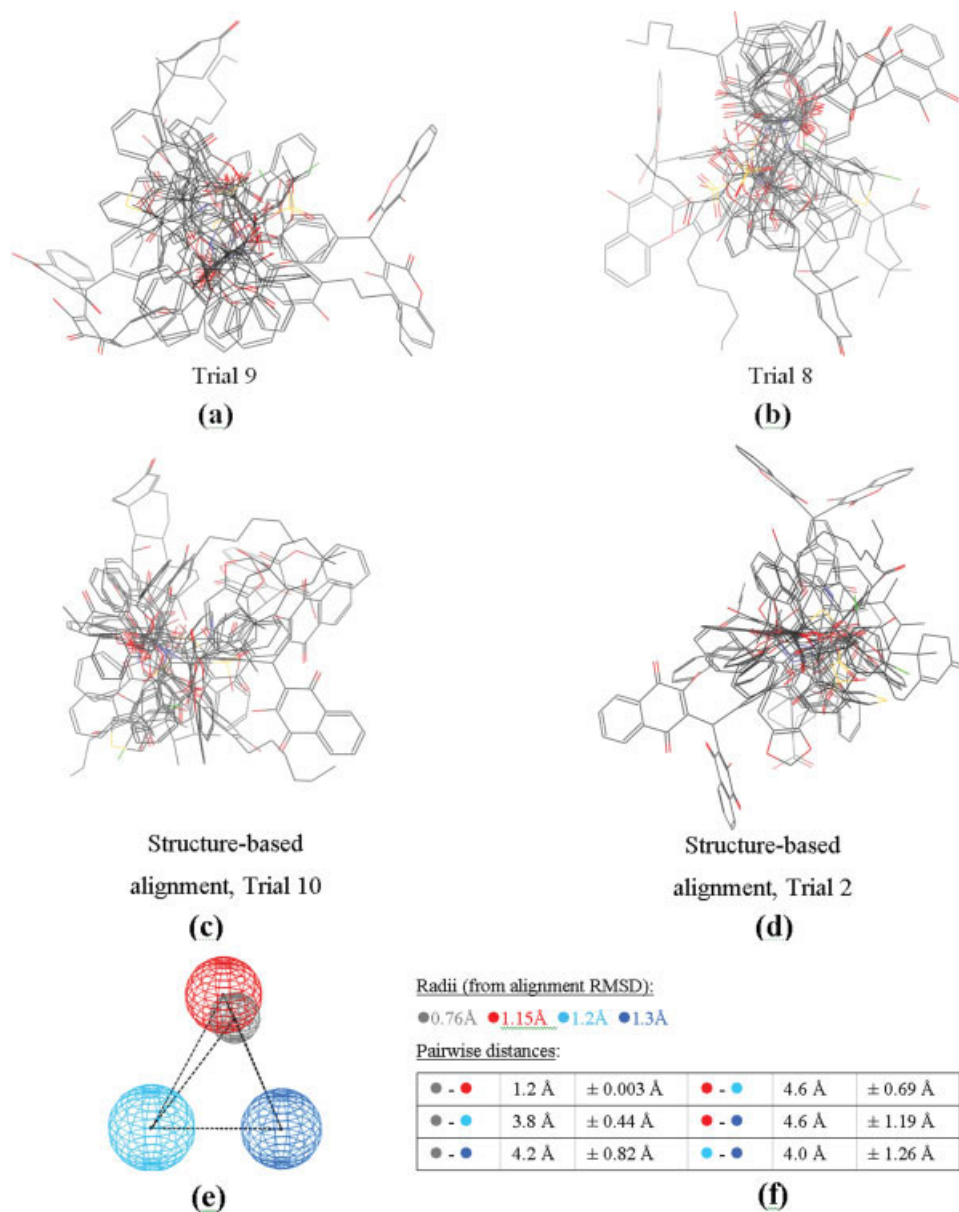


Figure 9. Alignments with highest (a,c) and lowest (b,d) energies derived from HIV-1 protease inhibitor candidates of Wang et al., ligand-based (a,b) and structure-based (c–d) pharmacophores, and refinement/minimization; (e) the derived structure-based pharmacophore; (f) measurements. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

for example specifying that hydroxyl 1 and 2 are interchangeable and equivalent (as seems to be the case in the HIV1 protease example); that any two out of three hydroxyls in a molecule A can be its pharmacophore points; or that one of the two carbonyls in molecule B matches one of the five carbonyls in molecule C. We have tried such an approach in the past in the context of protein-ligand docking using SPE, where several protein atoms may interact with each ligand atom, and obtained mixed results (Xu and Agrafiotis, unpublished data). The method seemed to often get stuck into suboptimal correspondences, and

did not exploit other, more plausible combinations. We are considering improvements to the algorithm to alleviate this problem.

Conclusion

We have presented an algorithm for aligning multiple molecules specified as 1D (SMILES) or 2D (connection tables) using shared feature hypotheses specified as corresponding groups on

each molecule. This is a novel application of a self-organizing method previously developed for conformational analysis or 3D structure generation. We cast the task of alignment as a conformational analysis of the molecular ensemble (an idea previously explored using distance geometry). The resulting implementation is useful as a fast way to thoroughly and stochastically sample the molecular alignment space, and to develop 3D pharmacophores from 1D correspondences so that one may assess and rank competing hypotheses by the fit and quality of the alignment, and use them as database search queries. Case studies using nontrivial alignments of multiple molecules by ligand and structure-based pharmacophores demonstrate the method's speed, accuracy and utility. We feel that this method provides a useful new tool for computer-assisted drug design.

Appendix: Derivation of Gradient of a Bidirectional Angle Vector During Ring Rotation

Assume that the two rings have normal vectors \hat{n}_1 and \hat{n}_2 respectively making an angle θ between them. Then the atoms of the second ring are allowed to rotate about the axis χ given by $\hat{n}_1 \times \hat{n}_2$ in order to align the normal vectors. Denote the vector in the plane of the first ring perpendicular to both \hat{n}_1 and χ as the *basis* vector $V = \chi \times \hat{n}_1$. Now at any point X on the second ring (on or near its plane), the local angle θ it needs to rotate to align with the first ring can be expressed as $\cos \theta = (V \cdot \hat{X}) / (V \cdot X) / \|X\|$. So if we set $\alpha = 1 - \cos^2 \theta$, which goes to zero when $\cos \theta$ is ± 1 so that the rings are aligned with vectors either parallel or antiparallel, and take derivatives w.r.t. X on both sides, we get $\partial \alpha / \partial X = \partial / \partial X (1 - \cos^2 \theta) = \partial / \partial X [1 - \{(V \cdot X) / \|X\|\}^2]$. We use the substitution and quotient rules for differentiation, and work out the derivatives of $V \cdot X$ and $\|X\|$ by breaking X and V into coordinates: $\partial(V \cdot X) / \partial X = V$ and $\partial \|X\| / \partial X = X / \|X\|$. Finally we get the following formula for the gradient:

$$\frac{\partial \alpha}{\partial X} = \frac{2(V \cdot X) [(V \cdot X)X - \|X\|^2 V]}{\|X\|^4}$$

References

- Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew Chem Int Ed Engl* 1999, 38, 2894.
- Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D.; Spellmeyer, D. C.; Miller, J. L. *J Med Chem* 2000, 43, 2770.
- Aronov, A. M.; Goldman, B. B. *Bioorg Med Chem* 2004, 12, 2307.
- Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. *J Chem Inf Comput Sci* 1994, 34, 207.
- Hindle, S. A.; Rarey, M.; Buning, C.; Lengauer, T. *J Comput Aided Mol Des* 2002, 16, 129.
- Joseph-McCarthy, D.; Thomas, B. E., IV; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. *Proteins* 2003, 51, 172.
- Daeyaert, F.; de Jonge, M.; Heeres, J.; Koymans, L.; Lewi, P.; Vinkers, M. H.; Janssen, P. A. J. *Proteins* 2004, 54, 526.
- Goto, J.; Kataoka, R.; Hirayama, N. *J Med Chem* 2004, 47, 6804.
- Moitessier, N.; Henry, C.; Maigret, B.; Chapleur, Y. *J Med Chem* 2004, 47, 4178.
- Wang, J.; Kang, X.; Kuntz, I. D.; Kollman, P. A. *J Med Chem* 2005, 48, 2432.
- Steindl, T. M.; Crump, C. E.; Hayden, F. G.; Langer, T. *J Med Chem* 2005, 48, 6250.
- Yamashita, F.; Hashida, M. *Drug Metab Pharmacokinet* 2004, 19, 327.
- Güner, O. F., Ed. *Pharmacophore Perception, Development, and Use in Drug Design*; International University Line (IUL Biotechnology Series): La Jolla, CA, 2000.
- Wolber, G.; Langer, T. *J Chem Inf Model* 2005, 45, 160.
- Wolber, G.; Dornhofer, A. A.; Langer, T. *J Comput Aided Mol Des* 2007, 20, 773.
- Lemmen, C.; Lengauer, T. *J Comput Aided Mol Des* 2000, 14, 215.
- Mason, J. S.; Good, A. C.; Martin, E. J. *Curr Pharm Des* 2001, 7, 567.
- Langer, T.; Krovat, E. M. *Curr Opin Drug Discov Devel* 2003, 6, 370.
- Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. *Curr Med Chem* 2004, 11, 71.
- Hou, T.; Xu, X. *Curr Pharm Des* 2004, 10, 1011.
- Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. *Drug Discov Today* 2004, 9, 27.
- Balakin, K. V.; Kozintsev, A. V.; Kiselyov, A. S.; Savchuk, N. P. *Curr Drug Discov Technol* 2006, 3, 49.
- Kurogi, Y.; Güner, O. F. *Curr Med Chem* 2001, 8, 1035.
- Güner, O. F.; Clement, O.; Kurogi, Y. *Curr Med Chem* 2004, 11, 2991.
- Clement, O. O.; Mehl, A. T. In *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line (IUL Biotechnology Series): 2000; pp. 69–83.
- Li, H.; Sutter, J.; Hoffmann, R. In *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line (IUL Biotechnology Series): 2000; pp. 171–189.
- Laggner, C.; Schieferer, C.; Fiechtner, B.; Poles, G.; Hoffmann, R. D.; Glossmann, H.; Langer, T.; Moebius, F. F. *J Med Chem* 2005, 48, 4754.
- Olsen, L.; Jost, S.; Adolph, H.-W.; Pettersson, I.; Hemmingsen, L.; Jorgensen, F. S. *Bioorg Med Chem* 2006, 14, 2627.
- Schuster, D.; Maurer, E. M.; Laggner, C.; Nashev, L. G.; Wilckens, T.; Langer, T.; Odermatt, A. *J Med Chem* 2006, 49, 3454.
- Bernard, D.; Coop, A.; MacKerell, A. D. *J Am Chem Soc* 2003, 125, 3101.
- Vedani, A.; Zbinden, P.; Snyder, J. P. *J Recept Res* 1993, 13, 163.
- Seifert, M. H. J. *J Chem Inf Model* 2005, 45, 449.
- Schuller, A.; Fechner, U.; Renner, S.; Franke, L.; Weber, L.; Schneider, G. *Comb Chem High Throughput Screen* 2006, 9, 359.
- Renner, S.; Ludwig, V.; Boden, O.; Scheffer, U.; Gobel, M.; Schneider, G. *Chembiochem* 2005, 6, 1119.
- Renner, S.; Schneider, G. *J Med Chem* 2004, 47, 4653.
- Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. *J Comput Aided Mol Des* 2003, 17, 687.
- Perola, E.; Charifson, P. S. *J Med Chem* 2004, 47, 2499.
- Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. *J Comput Aided Mol Des* 2002, 16, 653.
- Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. *J Chem Inf Model* 2005, 45, 461.
- Jones, G.; Willett, P.; Glen, R. C. *J Comput Aided Mol Des* 1995, 9, 532.
- Richmond, N. J.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. *J Comput Aided Mol Des* 2006, 20, 567.
- Shepphird, J.; Clark, R. *J Comput Aided Mol Des* 2006, 20, 763.
- Cottrell, S. J.; Gillet, V. J.; Taylor, R.; Wilton, D. J. *J Comput Aided Mol Des* 2004, 18, 665.

44. Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. *J Med Chem* 1986, 29, 899.
45. Agrafiotis, D. K.; Xu, H. *Proc Natl Acad Sci USA* 2002, 99, 15869.
46. Xu, H.; Izrailev, S.; Agrafiotis, D. K. *J Chem Inf Comput Sci* 2003, 43, 1186.
47. Agrafiotis, D. K. *J Comput Chem* 2003, 24, 1215.
48. Nathan Reed, post in discussion forum titled "Average Normal Vector" on devmaster.net, <http://www.devmaster.net/forums/showthread.php?t=6741>, accessed May 3, 2007.
49. Wang, S.; Milne, G. W.; Yan, X.; Posey, I. J.; Nicklaus, M. C.; Graham, L.; Rice, W. G. *J Med Chem* 1996, 39, 2047.
50. Martin, J.; Andrews, P. *J Comput Aided Mol Des* 1987, 1, 53.
51. Froimowitz, M. *NIDA Res Monogr* 1993, 134, 178.
52. Filizola, M.; Villar, H. O.; Loew, G. H. *J Comput Aided Mol Des* 2001, 15, 297.
53. Kaczor, A.; Matosiuk, D. *Curr Med Chem* 2002, 9, 1567.
54. Grundt, P.; Williams, I. A.; Lewis, J. W.; Husbands, S. M. *J Med Chem* 2004, 47, 5069.
55. Orsini, M. J.; Nesmelova, I.; Young, H. C.; Hargittai, B.; Beavers, M. P.; Liu, J.; Connolly, P. J.; Middleton, S. A.; Mayo, K. H. *J Biol Chem* 2005, 280, 8134.
56. Messer, W. S., Jr. Opioid systems. Available at: <http://www.neurosci.pharm.utoledo.edu/MBC3320/opioids.htm>, accessed May 3, 2007.
57. Portoghese, P. S.; Nagase, H.; Takemori, A. E. *J Med Chem* 1988, 31, 1344.
58. Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. *Biopolymers* 1996, 38, 221.
59. Huang, P.; Kim, S.; Loew, G. *J Comput Aided Mol Des* 1997, 11, 21.
60. Liao, S.; Alfaro-Lopez, J.; Shenderovich, M. D.; Hosohata, K.; Lin, J.; Li, X.; Stropova, D.; Davis, P.; Jernigan, K. A.; Porreca, F.; Yamamura, H. I.; Hruby, V. J. *J Med Chem* 1998, 41, 4767.
61. Liu, D. X.; Tang, Y.; Jiang, H. L.; Chen, K. X.; Ji, R. Y. *Zhongguo Yao Li Xue Bao* 1998, 19, 445.
62. Shenderovich, M. D.; Liao, S.; Qian, X.; Hruby, V. J. *Biopolymers* 2000, 53, 565.
63. Munro, T. A.; Rizzacasa, M. A.; Roth, B. L.; Toth, B. A.; Yan, F. *J Med Chem* 2005, 48, 345.
64. Wu, Y. C.; Hsieh, J. Y.; Lin, H. C.; Hwang, C. C. *J Mol Model* 2006, 13, 171.
65. Pearce, H. L.; Safa, A. R.; Bach, N. J.; Winter, M. A.; Cirtain, M. C.; Beck, W. T. *Proc Natl Acad Sci USA* 1989, 86, 5128.
66. Pearce, H. L.; Winter, M. A.; Beck, W. T. *Adv Enzyme Regul* 1990, 30, 357.
67. Lovekamp, T.; Cooper, P. S.; Hardison, J.; Bryant, S. D.; Guerrini, R.; Balboni, G.; Salvadori, S.; Lazarus, L. H. *Brain Res* 2001, 902, 131.
68. Varma, S.; Hou, Z. Available at: <http://www.accelrys.com/reference/cases/studies/pharmacophore-patterns.html>, accessed May 3, 2007.
69. Garrigues, A.; Loiseau, N.; Delaforge, M.; Ferte, J.; Garrigos, M.; Andre, F.; Orłowski, S. *Mol Pharmacol* 62, 1288.
70. Pajeva, I. K.; Globisch, C.; Wiese, M. *J Med Chem* 2004, 47, 2523.
71. Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. *J Chem Inf Comput Sci* 2004, 44, 1497.
72. Cianchetta, G. R.; Singleton, W.; Zhang, M.; Wildgoose, M.; Giesing, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. *J Med Chem* 2005, 48, 2927.
73. Kubinyi, H. In *Handbook of Cheminformatics*; J. Gasteiger, Ed.; Wiley-VCH: Weinheim, 2003.
74. Pandit, D.; So, S.-S.; Sun, H. *J Chem Inf Model* 2006, 46, 1236.
75. Meagher, K. L.; Lerner, M. G.; Carlson, H. A. *J Med Chem* 2006, 49, 3478.
76. Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumder, A.; Sunder, S.; Chen, J.; Milne, G. W.; Pommier, Y. *J Med Chem* 1997, 40, 920.
77. Mustata, G. I.; Brigo, A.; Briggs, J. M. *Bioorg Med Chem Lett* 2004, 14, 1447.
78. Brigo, A.; Mustata, G. I.; Briggs, J. M.; Moro, S. *Med Chem* 2005, 1, 263.
79. Barreca, M. L.; Ferro, S.; Rao, A.; De Luca, L.; Zappala, M.; Monforte, A.-M.; Debyser, Z.; Witvrouw, M.; Chimirri, A. *J Med Chem* 2005, 48, 7084.
80. Deng, J.; Lee, K. W.; Sanchez, T.; Cui, M.; Neamati, N.; Briggs, J. M. *J Med Chem* 2005, 48, 1496.
81. Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. *J Med Chem* 2002, 45, 1737.
82. Wang, S.; Sakamuri, S.; Enyedy, I. J.; Kozikowski, A. P.; Deschaux, O.; Bandyopadhyay, B. C.; Tella, S. R.; Zaman, W. A.; Johnson, K. M. *J Med Chem* 2000, 43, 351.
83. Zhu, F.; Agrafiotis, D. K. *J Comput Chem* 2007, 28, 1234.
84. Agrafiotis, D. K.; Gibbs, A.; Zhu, F.; Izrailev, S.; Martin, E. *J Chem Inf Model* 2007, 47, 1067.