JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

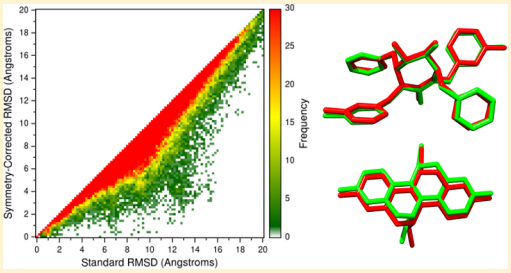
Open Access on 01/11/2015

Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design

William J. Allen[†] and Robert C. Rizzo*[,†,‡,§]

[†]Department of Applied Mathematics & Statistics, Stony Brook University, Stony Brook, New York 11794, United States
[‡]Institute of Chemical Biology & Drug Discovery, Stony Brook University, Stony Brook, New York 11794, United States
[§]Laufer Center for Physical & Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States

ABSTRACT: False negative docking outcomes for highly symmetric molecules are a barrier to the accurate evaluation of docking programs, scoring functions, and protocols. This work describes an implementation of a symmetry-corrected root-mean-square deviation (RMSD) method into the program DOCK based on the Hungarian algorithm for solving the minimum assignment problem, which dynamically assigns atom correspondence in molecules with symmetry. The algorithm adds only a trivial amount of computation time to the RMSD calculations and is shown to increase the reported overall docking success rate by approximately 5% when tested over 1043 receptor−ligand systems. For some families of protein systems the results are even more dramatic, with success rate increases up to 16.7%. Several additional applications of the method are also presented including as a pairwise similarity metric to compare molecules during de novo design, as a scoring function to rank-order virtual screening results, and for the analysis of trajectories from molecular dynamics simulation. The new method, including source code, is available to registered users of DOCK6 (http://dock.compbio.ucsf.edu).



INTRODUCTION

Docking is a well-established computational technique often used at the early stages of structure-based drug discovery.[1−3] In general, the purpose of a docking experiment is to predict the optimal binding geometry (pose) of a *ligand* (typically a small organic molecule) with respect to a *receptor* (typically a protein drug target), and to provide an estimate of the interaction affinity relative to other ligands.[4] DOCK is one such program designed for binding pose prediction and, when used in a virtual screening capacity to rank-order large databases of candidate molecules, is a powerful approach for lead discovery. Numerous DOCK successes have been reported;[1,2] recent examples include identification of leads targeting CTX-M β-lactamase,[5] β_2-adrenergic receptor,[6] XPA protein,[7] and the riboswitch of mRNA,[8] among others. Examples of successful virtual screens from our laboratory employing the most recent version of DOCK (version 6)[9,10] include the drug targets HIVgp41[11] and fatty acid binding protein.[12] Other related programs, including AutoDock,[13] AutoDock Vina,[14] GOLD,[15,16] Glide,[17,18] Surflex,[19] and FlexX,[20] are widely used for similar goals and have documented analogous successes (reviewed in Yuriev et al.[4]).

Root-mean-square deviation (RMSD), a measure of intermolecular differences in position and conformation, has broad applications in computational biology,[21−23] including testing the efficacy of docking programs.[10,24] For example, in a typical docking validation experiment, a cocrystallized ligand is removed from its binding site, then multiple candidate binding poses are predicted using a specific protocol. The positive-control experiment is successful if the heavy-atom RMSD of the *top-ranked pose* is within a certain tolerance from the original ligand position; a typical cutoff for success is 2.0 Å. However, if a molecule contains symmetric functional groups or is itself symmetric, inflated RMSD values and a false negative result can occur.[25] As an example, docking 1,2-dichlorobenzene to T4 lysozyme (PDB code 2OTY)[26] can yield chemically equivalent binding poses, yet different RMSD values (Figure 1). In one scenario, the pose RMSD is less than 2.0 Å from the crystallographic reference and is considered a docking success. In the second scenario, the pose is inverted over its axis of symmetry, yielding an RMSD greater than 2.0 Å, which is reported as a docking failure. DOCK6 was recently reported to successfully redock between 63.8% (over 780 systems)[24] and 66.3% (over

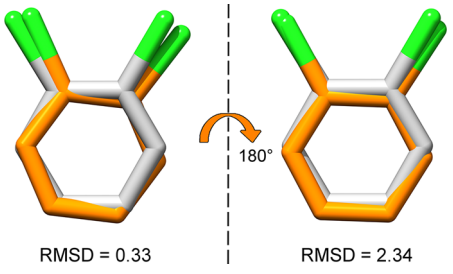


Figure 1. Illustration of symmetry problem from two docking outcomes for 1,2-dichlorobenzene. Crystallographic reference pose, shown in gray, is identical in both panels. Docked poses, shown in orange, are inverted approximately 180° about the axis of symmetry. RMSD values are in Å.

Received: September 12, 2013
Published: January 11, 2014

147 systems)[10] of known crystallographic ligand−receptor complexes in a series of flexible ligand docking experiments using a standard energy-based scoring function. A subset of the failures from these experiments could be attributed to molecular symmetry in the ligand. When symmetry is accounted for, an increase in success of approximately 4−8% is observed, depending on the experiment.[10] These trends demonstrate that neglecting to account for symmetry in ligands can negatively impact the accurate evaluation of new docking programs, protocols, search methods, or scoring functions.

Despite the importance of symmetry in docking, there remains a surprising lack of published details regarding specific functions used for symmetry correction, as was recently noted by Brozell et al.[10] A notable exception is the publication describing AutoDock Vina,[14] which reports the exact procedure used to compute what the authors refer to as the *minimum-distance* RMSD. However, while functionally simple to implement, minimum-distance RMSD in some cases may not enforce a one-to-once atom correspondence. Under these circumstances, some atoms could be used multiple times for the RMSD calculation while others could be neglected, which may not be desirable (see Theoretical Methods). As an alternative approach, our group has implemented a symmetry-correction strategy into the DOCK6 program[10] based on the method described by Kuhn[27] and later by Munkres[28] known as the Hungarian algorithm.[29] A similar strategy was recently reported by Helmich et al.[22] to find the best overlap in alignment between two conformations of the same molecule, and by Ellingson et al.[23] to efficiently maximize the number of atom superpositions in an overlay of protein binding sites.

In this report, we demonstrate that our implementation of the Hungarian algorithm effectively corrects for symmetry in molecular docking which provides the means to assess the impact of various algorithmic or protocol changes as the DOCK program continues to evolve. Further, we discuss an extension of the algorithm for adding chemical diversity to a pool of molecules that are assembled in a de novo design version of DOCK currently under development, enabling improved sampling of diverse chemotypes. Additional applications presented include rank-ordering compounds relative to a known reference to aid virtual screening and computing symmetry-corrected RMSDs to analyze molecular dynamics simulations. The Hungarian method is available in the most current release of DOCK6.

THEORETICAL METHODS

Standard RMSD. The standard pairwise RMSD function, as it is most commonly used in structural biology, quantifies the difference between two poses of the same molecule. In this context, differences between poses can include rigid geometric rotations and translations within the six standard degrees of freedom (three rotational and three translational), plus any changes in internal degrees of freedom (such as dihedral rotations). The standard RMSD function ($\mathrm{RMSD}_{\mathrm{std}}$) is defined as follows:

$$\mathrm{RMSD}_{\mathrm{std}}(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| \mathbf{a}_i - \mathbf{b}_i \|^2} \tag{1}$$

where molecule poses A and B are sets comprising individual atoms: $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_N\}$ and $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_N\}$. Atoms $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ can be interpreted as Cartesian coordinates or positional vectors: for example, $\boldsymbol{a}_i = \{x_{a,i}, y_{a,i}, z_{a,i}\}$. The total number of heavy atoms in the molecule is N.

Historically, this function is dependent on a predefined one-to-one correspondence between atoms in two poses (A and B) of the same molecule. For practical purposes, this is typically achieved by matching atom indices or evaluating the sets of atoms in the same order. In the case of highly symmetric molecules, however, it would be advantageous to impose a dynamic atom-matching algorithm prior to the standard RMSD calculation that is not subject to the constraints of arbitrary atom numbers or orders, such that physically indistinguishable atoms would be matched.

Minimum-Distance RMSD. In an attempt to overcome this limitation, Trott and Olson[14] recently described a modified function termed minimum-distance RMSD ($\mathrm{RMSD}_{\mathrm{min}}$), computed as follows:

$$\mathrm{RMSD}_{\mathrm{min}}(A, B) = \max\{\mathrm{RMSD'}_{\mathrm{min}}(A, B), \mathrm{RMSD'}_{\mathrm{min}}(B, A)\} \tag{2}$$

where

$$\mathrm{RMSD}'_{\mathrm{min}}(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\min_j \| \mathbf{a}_i - \mathbf{b}_j \|)^2} \tag{3}$$

In this method, atoms $\boldsymbol{a}_i$ from reference pose A are iteratively compared to all atoms $\boldsymbol{b}_j$ from pose B that are of the same element type. The minimum distance identified defines the atom correspondence, which is subsequently used in the root-mean-square calculation. The same procedure is repeated using pose B as the reference, and the maximum of the two calculations is reported as the $\mathrm{RMSD}_{\mathrm{min}}$. However, as this function does not enforce a one-to-one correspondence, some atoms may be used more than once in the calculation while others that are far removed in coordinate space have the potential to be neglected. Therefore, while this method is effective, it can under-represent the true RMSD.

In order to account for all atoms in each pose A and B, the minimum-distance RMSD could be modified to enforce a one-to-one correspondence through removal of atom pairs after they have been matched via following the steps: (1) for each i, compute the minimum distance: $\min_j \|\boldsymbol{a}_i - \boldsymbol{b}_j\|$, (2) remove $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ from atom sets A and B, respectively, and (3) repeat steps 1−2 until no atoms remain to be matched. Once the complete atom correspondence is determined, the RMSD would be computed similar to eq 3. However, while this modified approach would ensure that no atoms are neglected from the final RMSD calculation, it is dependent on the order in which atoms are evaluated and may give a nonoptimal solution.

Optimal-Correspondence RMSD. As an alternative approach, the *optimal* one-to-one correspondence between sets of atoms can be solved using the Hungarian algorithm,[27,28] a method for solving the minimum assignment problem.[29] In this approach, every atom from pose A will be matched to exactly one atom from pose B in such a way that minimizes the sum of distances between all atom pairs, and no outlying atoms will be neglected from the RMSD calculation. The optimal-correspondence RMSD ($\mathrm{RMSD}_{\mathrm{cor}}$), termed here the symmetry-corrected RMSD, can be defined as follows:

$$\mathrm{RMSD}_{\mathrm{cor}}(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\mathrm{cor}_j \| \mathbf{a}_i - \mathbf{b}_j \|)^2} \tag{4}$$

where atom $\boldsymbol{a}_i$ is matched uniquely to an atom $\boldsymbol{b}_j$ that was predetermined by solving the optimal correspondence $\mathrm{cor}_j\|\boldsymbol{a}_i - \boldsymbol{b}_j\|$.

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

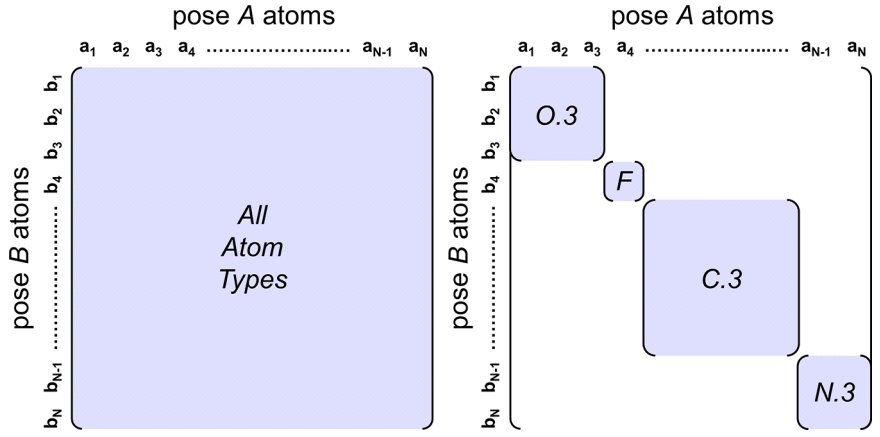


Figure 2. In this example, a theoretical molecule that contains only four Sybyl atom types (*O.3*, *F*, *C.3*, and *N.3*) is described. The search space was reduced from all possible atom comparisons (left) to comparisons between only atoms of the same atom type (right). The amount of calculation required, represented by blue shading, is significantly decreased by dividing the molecule into subsets, or submatrices.

Briefly described, the Hungarian algorithm solves this correspondence in four steps: (1) form a cost matrix M where each matrix element m_{ij} is the squared norm of atom positional vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$. Subtract the lowest value in each row from all elements in the same row and the lowest value in each column from all elements in the same column. (2) Find the minimum number of horizontal and vertical lines that can cover all zeroes in the matrix. If the number of lines is equal to the size of the matrix in one dimension, a set of independent zero-values exists, and the positions of these zeroes represent the optimal-correspondence solution. Otherwise, proceed to the next step. (3) Determine the smallest matrix element that is not covered with a line. Subtract that value from all uncovered matrix elements and add that same value to all matrix elements that are covered by two lines. (4) Repeat steps 2−3 until the matrix is solved as described in step 2. Each iteration of steps 2−3 constitutes a *cycle* of the algorithm. A more thorough description and proof are documented by Munkres.[28]

COMPUTATIONAL DETAILS

Implementation and Testing. A very fast, $O(n^4)$ version of the Hungarian algorithm[29] was implemented into the source code of DOCK6. Disjoint distance matrices (termed submatrices) are assembled for each atom type (using Tripos Sybyl conventions)[30] in a given molecule, thus only atoms of the same type can correspond to one another in the RMSD calculation (see Figure 2). All tests were performed on a Dell PowerEdge C6100 Linux cluster consisting of Intel x5660 2.8 GHz hex-core Nehalem-based processors.

Docking for Pose Reproduction. Pose-reproduction docking experiments were performed over a large test set using the flexible ligand protocol described by Mukherjee et al.[24] The test set used is an expanded version of the SB2010 test set ($N = 780$)[24] recently increased to include 1043 receptor−ligand complexes, available for download as SB2012 at www.rizzolab.org. Key aspects of the test set preparation and flexible ligand docking protocol include the following: Ligands were protonated within the program MOE,[31] then semiempirical AM1-BCC charges[32,33] were added with *antechamber*, part of the AmberTools package.[34] Receptors were protonated and assigned force field parameters with *tleap* (AmberTools), then the receptor structures were subjected to a short energy minimization using *sander* (Amber11)[34] with strong restraints on the heavy atoms.

Following standard DOCK6 practices,[9,35,36] the receptor surface was generated with DMS,[37] spheres were generated using the utility SPHGEN,[38] and a box was defined that surrounded the spheres plus an 8.0 Å margin in all directions. A docking grid[39] was generated within the boundaries of the box at 0.3 Å resolution with 6−9 van der Waals exponents and a distance-dependent dielectric of $\varepsilon = 4r$. During docking, a maximum of 1000 anchor orientations were attempted beginning from each anchor with at least 5 atoms. Ligand positions were minimized during growth with a simplex minimizer for a maximum of 500 iterations or until the change in energy score between steps was less than 1.0 kcal/mol. A maximum of 5000 fully grown conformers were kept for each ligand, following best-first clustering with a 2.0 Å standard RMSD threshold. The resulting ensembles of candidate poses (88 853 total poses; average 85.2 poses per system) were evaluated using both the standard (eq 1) and symmetry-corrected (eq 4) RMSD algorithms described above. All RMSD values reported were measured between the docked pose and the crystallographic position of the same ligand. RMSD values are computed using heavy-atoms only − hydrogen atom positions were not considered.

Chemical Diversity in De Novo Design. As an alternative application of the Hungarian algorithm, we have adapted the code to enable clustering and pruning between molecules of differing chemical composition in order to aid in de novo design. Depending on several factors, the number of candidate molecules during de novo assembly can quickly expand exponentially, thus rigorous pruning methods are required to keep the search space tractable. To evaluate the impact of the new pruning heuristic, experiments to reconstruct known molecules and to construct diverse ensembles of molecules were performed using a de novo version of DOCK currently under development (further discussed in Balius et al.[40]).

Prior to the de novo experiments, two different groups of molecule fragment libraries were prepared. First, all ligands from the SB2012 test set with exactly seven rotatable bonds ($N = 103$) were independently disassembled on those bonds into 103 small, *restricted* fragment libraries. Second, a subset of the ZINC database[41] consisting of approximately 1×10^6 representative drug-like molecules was used to form a larger, *generic* fragment library. New molecules were then assembled de novo in the binding sites of each receptor from the SB2012 subset using either (1) the smaller, restricted fragment libraries associated with each system or (2) the larger, generic library from ZINC

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

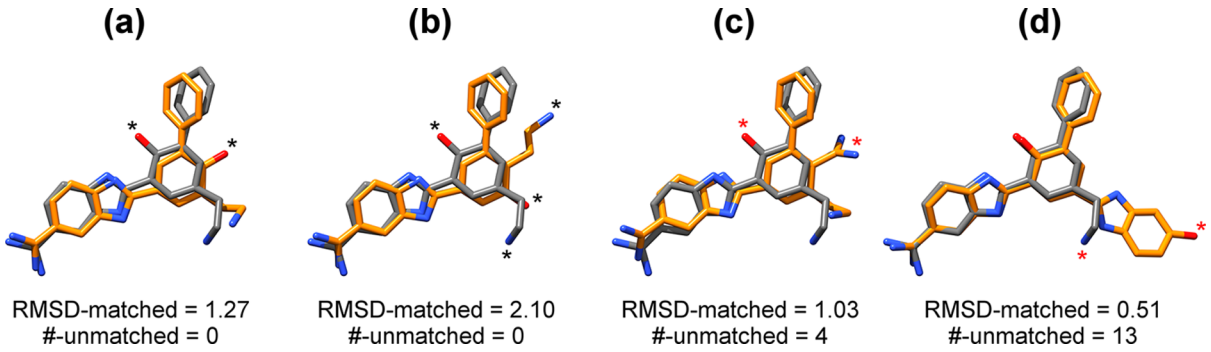


Figure 3. Comparison between a gray reference molecule (same in all four panels) and four example molecules constructed by de novo design with: (a) slight differences in functional group positions, (b) larger differences in functional group positions, (c) slight differences in functional group composition, and (d) larger differences in functional group composition. Black asterisks indicate changes in position; red asterisks indicate changes in composition. RMSD-matched values are reported for common atoms in Å.

seeded with fragments from the restricted libraries. Following each layer of assembly, molecule ensembles were rank-ordered by the value of their DOCK grid energy score,[39] then pruned using a best-first clustering method and the adaptation of the Hungarian algorithm. The criteria for pruning were as follows: All possible atom matches (using Sybyl conventions) between the two molecules were identified, the optimal-correspondence of those matches was determined, and the symmetry-corrected RMSD for matched atoms was computed (hereafter referred to as *RMSD-matched*). In addition, the atoms that remained unmatched in either molecule were enumerated (hereafter referred to as *#-unmatched*). Two molecules were deemed to be sufficiently similar in Cartesian space *and* chemical space if the RMSD-matched term was less than a user-defined cutoff (typically 2.0 Å), and the #-unmatched term was less than a second user-defined cutoff (typically 0−10 atoms). Under these conditions, the worse-scoring molecule would be pruned. In practice, as the threshold for number of unmatched atoms increases, molecules would be pruned more frequently. For clarity, this heuristic is illustrated in Figure 3. Tanimoto coefficients[42] were computed using a fingerprinting method implemented into DOCK6, inspired by the MOLPRINT 2D algorithm described by Bender et al.[43,44]

RESULTS AND DISCUSSION

Timing and Efficiency. In total, 88 853 docked poses derived from the SB2012 test set were compared against 1043 crystallographic reference conformations using the standard and symmetry-corrected RMSD functions. The present implementation of the Hungarian algorithm was made significantly more efficient through a process where submatrices are prepared containing only atoms of the same Sybyl atom type for matching (see Figure 2). In terms of the chemistry of matching atoms, it follows that only atoms of an identical topology should be matched to correctly account for molecular symmetry. In order to verify that the Hungarian algorithm could successfully solve each minimum assignment problem in the data set, we measured the number of algorithm cycles required for each calculation. The 88 853 poses were divided into 675 056 submatrices based on atom type (average 7.60 submatrices, or atom types, per molecule) and the minimum assignment was determined. The results are plotted in Figure 4.

The maximum number of algorithm cycles required to solve the most difficult matrix was 265 cycles, which occurred in a matrix of 33 × 33 elements. In addition, the largest matrix evaluated in the data set was 35 × 35 elements, derived from a

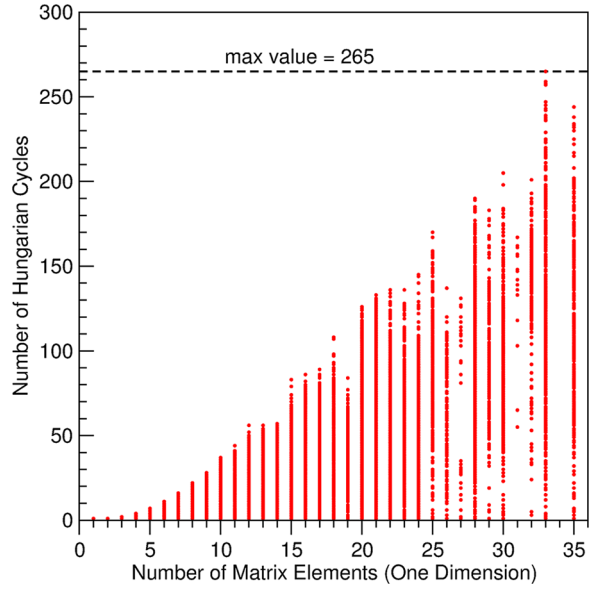


Figure 4. Scatter plot of the number of cycles of the Hungarian algorithm (*y*-axis) required to solve the minimum assignment problem for matrices of a given size (*x*-axis). Each red point represents one submatrix (N = 675 056). The maximum number of cycles required was 265.

molecule with 138 total atoms, of which 35 atoms were the same Sybyl atom type (sp3 carbon, PDB code 1KQY).[45] On average, 5.59 cycles were required to solve each submatrix. The complexity of solving the minimum assignment problem, as measured by the number of algorithm cycles required, increases with increasing number of matrix elements. Conceptually this is consistent with increasing opportunity for alternative atom correspondence among larger ligands. Figure 4 also demonstrates that all distance matrices from a very large data set were solvable, exhibiting the robustness of the algorithm. Finally, it is noteworthy that rescoring all 88 853 poses (675 056 total submatrices) required 77 981 ms total of computer time. This value factors to an average of 0.88 ms per pose or 0.12 ms per submatrix to solve the minimum assignment problem. Submillisecond time scales for evaluating the symmetry-corrected RMSD are negligibly small when compared to an average on-the-fly flexible ligand docking experiment in the program DOCK6, which under the present conditions required approximately 4.9 min per ligand. This millisecond time scale becomes an important consideration when, as discussed below, the algorithm is adapted to enable pairwise comparisons for de novo design.

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

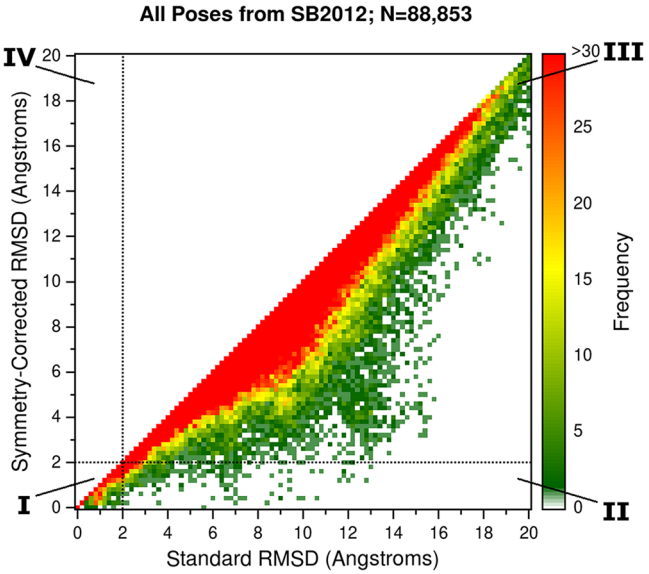


Figure 5. Molecule pose evaluation using standard and symmetry-corrected RMSD. Populations are binned in 0.2 Å increments, and are colored from green (low population) to red (high population) as shown on the right-hand side bar. Dashed lines at $x = 2$ Å and $y = 2$ Å delineate four quadrants: I = true positives, II = false negatives, III = true negatives, IV = false positives.

Standard RMSD versus Symmetry-Corrected RMSD. By design of the algorithm, the symmetry-corrected RMSD can only be less than or equal to the standard RMSD. That is, the atom correspondence will not change from the standard correspondence if any change increases the sum of distances between all atom pairs. A heat map of RMSD comparisons for all 88 853 poses is shown in Figure 5. Data here are grouped in 0.2 Å bins, red indicates areas of high population, green indicates areas of low population, and white areas are unpopulated. The heat map can be divided into four quadrants: Poses in quadrant I have both a low standard and symmetry-corrected RMSD (<2.0 Å), thus would be considered successfully docked poses by either metric. Among the results, these are considered to be the *true positives*. Poses in quadrant II have high standard RMSD values (≥2.0 Å), which would ordinarily be considered a docking failure, but they have low symmetry-corrected RMSD values (<2.0 Å). Thus, they are *false-negatives*, or poses that are *rescued* from failure by the symmetry-correction algorithm. Poses in quadrant III have a high RMSD by both metrics (≥2.0 Å), and are considered to be the *true negatives*. Finally, quadrant IV (false positives) cannot contain any poses and should always remain empty.

From Figure 5, it is evident that no poses are assigned a symmetry-corrected RMSD greater than the standard RMSD, consistent with expectations. The vast majority of poses fall near to the diagonal, as indicated by the thick red band across the heat map. However, while the average difference in RMSD between the two metrics ($RMSD_{std} - RMSD_{cor}$) is only 0.96 Å, 11 914 (13.4%) poses differ by more than 2.0 Å, and 5490 (6.2%) poses differ by more than 3.0 Å. The poses that differ by greater than 3.0 Å represent the most interesting cases of highly symmetric molecules that the Hungarian algorithm was designed to uncover, as discussed in the next section. The vast majority of poses in Figure 5, encompassed within the breadth of the red band, are likely indicative of commonly occurring, smaller corrections to symmetric functional groups. For example, as a point of comparison, in a hypothetical molecule of 25 heavy atoms, a 180° rotation of a single phenyl group would result in relatively small RMSD of 0.96 Å, while the 180° rotation of two phenyl groups would result in a larger RMSD of 1.36 Å.

Table 1. Quadrant Populations for Poses Docked Using the SB2012 Test Set

	IV: False Positives		III: True Negatives	
$RMSD_{cor} \geq 2.0$ Å	all poses[a]	0 (0%)	all poses	86 728 (97.6%)
	top pose[b]	0 (0%)	top pose	285 (27.3%)
	I: True Positives		II: False Negatives (Rescues)	
$RMSD_{cor} < 2.0$ Å	all poses	1272 (1.4%)	all poses	853 (1.0%)
	top pose	706 (67.7%)	top pose	52 (5.0%)
	$RMSD_{std} < 2.0$ Å		$RMSD_{std} \geq 2.0$ Å	

[a]Out of 88 853. [b]Out of 1043.

Rescued Poses in Standard Docking. To facilitate discussions of differences between standard and symmetry-corrected RMSD, it is convenient to analyze docking outcomes grouped by the four quadrants as shown in Table 1. The results here are further subdivided to specifically look at outcomes wherein only the top-scoring pose was retained for each system ($N = 1043$), or all poses that were generated were retained ($N = 88\,853$). It is important to note that the poses analyzed here follow a clustering protocol such that only a diverse set of poses from each ligand docked to its target will be retained. These poses are diverse in terms of both conformation and orientation within the binding pocket. Therefore, given the constraints of diversity, typically only one pose would be expected to be the correct pose, or within 2.0 Å from the crystallographic reference. In addition, it is expected that a small subset of poses, those that overlay within 2.0 Å from the reference but inverted over an axis of symmetry, will be rescued by the symmetry-correction algorithm. All remaining poses should be perceived as incorrect, or not within 2.0 Å from the reference. From the perspective of the DOCK6 scoring function, the correct poses and the rescued poses are likely to have indistinguishable energy scores, and it is a statistical coin-flip that ultimately determines which one is the top-ranked pose.

Among the data presented in Table 1, these expectations appear to hold true. Specifically, approximately one pose per system, or 1272 (1.4%), is identified as a true positive (quadrant I). The slight discrepancy between true positive poses ($N = 1272$) and number of systems ($N = 1043$) can be attributed to the DOCK6 sampling algorithm, which may generate two poses that are each less than 2.0 Å from the crystallographic reference ligand, but greater than 2.0 Å from each other, in which case both poses would be kept. In addition, 853 poses (1.0%) are identified as false negatives and rescued by the symmetry-correction algorithm (quadrant II), which is consistent with the hypothesis that only a small subset of poses can overlay with the crystallographic reference, but be inverted over some axis of symmetry. Finally, the vast majority, or 86 728 (97.6%), of all poses are true negatives (quadrant III). This demonstrates that the ensembles of candidate poses generated by DOCK6 fully explore the boundaries of the binding pockets and vary largely in conformation and orientation from the crystallographic ligands. As expected, no poses are considered false positives given that symmetry-corrected RMSD can never yield a higher value than standard RMSD (quadrant IV).

The need to account for symmetry becomes more important when only the top-scored pose for each system ($N = 1043$) is examined, as is typical for most applications of docking and

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

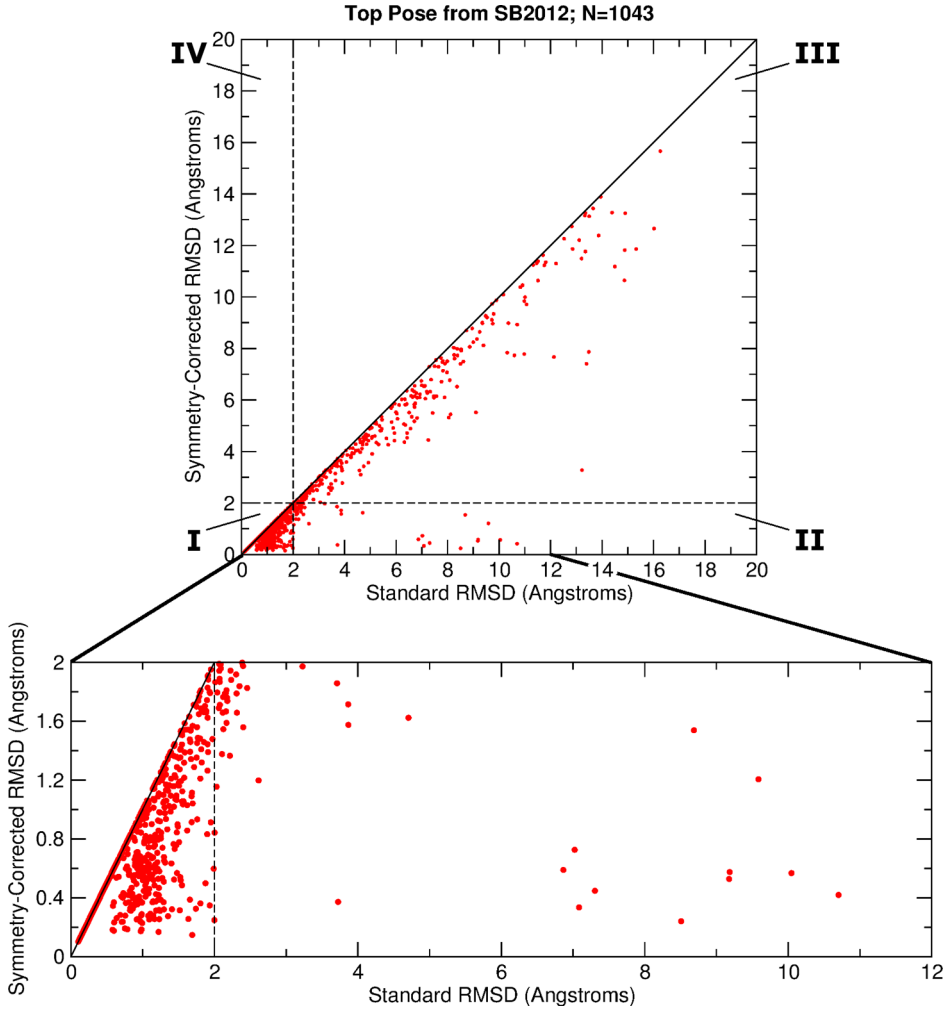


Figure 6. Scatter plot of standard RMSD (x-axis) vs symmetry-corrected RMSD (y-axis). Each red point represents the top-scored pose from systems in the SB2012 test set. The dashed lines plotted by $x = 2$ Å and $y = 2$ Å delineate the four quadrants. The solid line indicates the diagonal. The bottom insert is a magnification of quadrants I and II.

virtual screening. While the top pose in Table 1 is ranked successfully by both RMSD metrics 67.7% of the time, which is consistent with our laboratory's earlier statistics,[10,24] the number of poses rescued through symmetry correction increases from 1.0% to 5.0%. Therefore, the true pose-reproduction success rate using the flexible ligand docking protocols presented here becomes 72.7% (67.7% + 5.0%) across all 1043 systems in the SB2012 test set. A scatter plot comparing standard and symmetry-corrected RMSD values for the top pose from each system is shown in Figure 6. Again, most of the top-scored poses fall near to the diagonal. Focusing on quadrants I and II (Figure 6, insert), specific examples of top-scored poses begin to emerge as individual red points. The 20 most interesting cases (out of 52 in quadrant II) with large standard RMSD and small symmetry-corrected RMSD values are shown in Figure 7.

Each panel of Figure 7 contains the crystallographic reference pose (in red) overlaid with the top-scoring docking outcome (in green). Some of the poses, especially in the top three rows of Figure 7, are easily identifiable as highly symmetric molecules. The symmetry-correction algorithm has in some cases improved the reported RMSD by as much as 10 Å, a very significant change. Also in Figure 7, there are examples of molecules (e.g., 1SG0 and 1SRI) that were not inverted over a major axis of symmetry, but instead the RMSD was improved above the success threshold when accounting for a ring flip (as described previously). Importantly, a breakdown by protein families from the test set reveals that although the overall increase in success rate is 5.0%, more dramatic success rate improvements ranging up to 16.7% for specific families are observed (Table 2). Particularly encouraging is the 13.3% success rate improvement across the large HIV protease test set ($N = 60$) containing highly symmetric cyclic-urea based ligands, which provides a challenging test case for docking. Of the nine protein families in Table 2 for which symmetry-correction appears to offer no benefit (0% increase in success), five families already yield a standard RMSD success rate of greater than 50%, including the highly accurate results for HIV reverse transcriptase at 95.2% and neuraminidase at 93.0%. In these latter two cases, symmetry-correction would be expected to have only minimal impact. The remaining four families in Table 2 with standard RMSD success rates of less than 50% are for systems for which docking has been particularly challenging (see Figure 7 in Mukherjee et al.[24]) and symmetry correction here provides no improvement.

Chemical Diversity in De Novo Design. As described in Computational Details, we have adapted the Hungarian algorithm to enable pruning of molecules during de novo growth to avoid combinatorial explosion. The ultimate goal is a

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

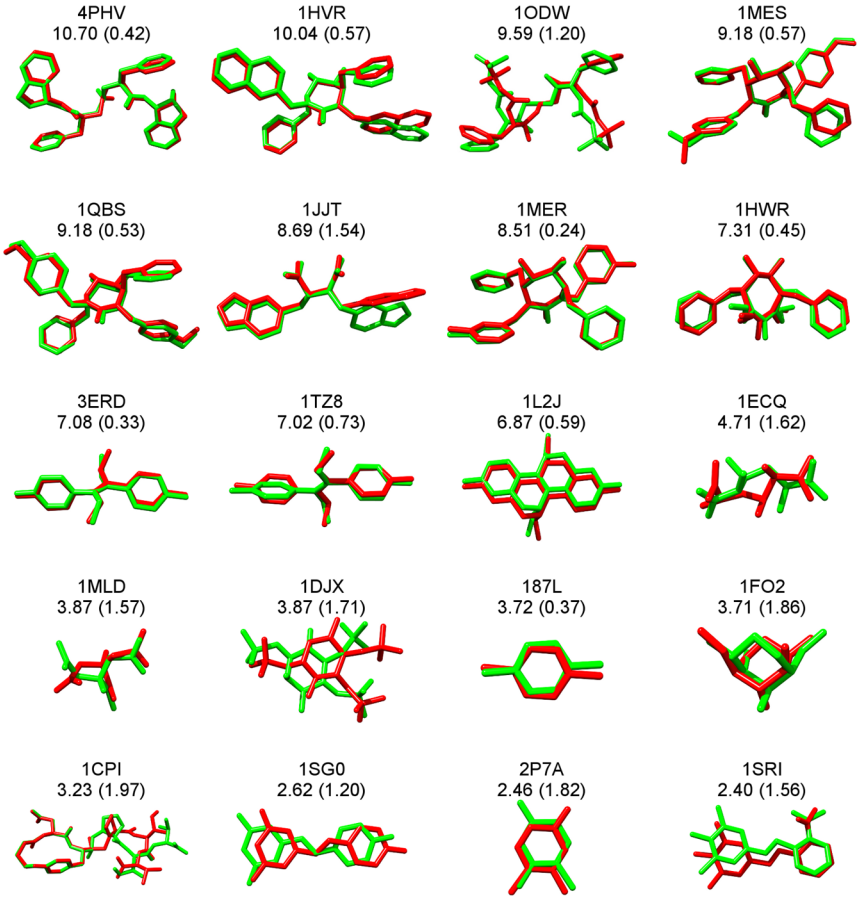


Figure 7. Examples of 20 molecules from a total of 52 rescues among the top-scored poses. The PDB code of the system is shown, followed by the standard RMSD, and in parentheses the symmetry-corrected RMSD. Crystallographic references are in red, candidate poses are in green. Poses are arranged by the magnitude of the standard RMSD. Values are in Å. Protein atoms are hidden for clarity.

Table 2. True Positive, False Negative, and Corrected Success Rates among the Top Pose from Families in SB2012

family	size	true positives	false negatives (rescues)	corrected success
thymidylate synthase	12	7 (58.3%)	2 (16.7%)	9 (75.0%)
thermolysin	26	9 (34.6%)	4 (15.4%)	13 (50.0%)
T4 lysozyme	13	9 (69.2%)	2 (15.4%)	11 (84.6%)
HIV protease	60	28 (46.7%)	8 (13.3%)	36 (60.0%)
tyrosine phosphatase	20	15 (75.0%)	2 (10.0%)	17 (85.0%)
estrogen receptor	45	40 (88.9%)	4 (8.9%)	44 (97.8%)
matrix metalloproteinase	14	5 (35.7%)	1 (7.1%)	6 (42.9%)
acetylcholinesterase	19	10 (52.6%)	1 (5.3%)	11 (57.9%)
all	*1043*	*706 (67.7%)*	*52 (5.0%)*	*758 (72.7%)*
factor Xa	41	37 (90.2%)	1 (2.4%)	38 (92.7%)
trypsin	46	27 (58.7%)	1 (2.2%)	28 (60.9%)
neuraminidase	43	40 (93.0%)	0 (0.0%)	40 (93.0%)
thrombin	37	23 (62.2%)	0 (0.0%)	23 (62.2%)
carbonic anhydrase	29	6 (20.7%)	0 (0.0%)	6 (20.7%)
β-trypsin	29	22 (75.9%)	0 (0.0%)	22 (75.9%)
HIV reverse transcriptase	21	20 (95.2%)	0 (0.0%)	20 (95.2%)
HMG-CoA reductase	20	15 (75.0%)	0 (0.0%)	15 (75.0%)
phospholipase A2	15	2 (13.3%)	0 (0.0%)	2 (13.3%)
ribonuclease A	14	7 (50.0%)	0 (0.0%)	7 (50.0%)
egg lysozyme	14	5 (35.7%)	0 (0.0%)	5 (35.7%)

final ensemble of molecules with high diversity and constructed within a reasonable amount of time, that can then be used to guide organic synthesis or to identify related purchasable compounds for experimental testing. In terms of evaluation and validation of our de novo routines, we hypothesize that Hungarian-based pruning will increase efficiency in experiments designed to (1) reconstruct known parent molecules from restricted fragment libraries in their respective binding sites (i.e.,

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

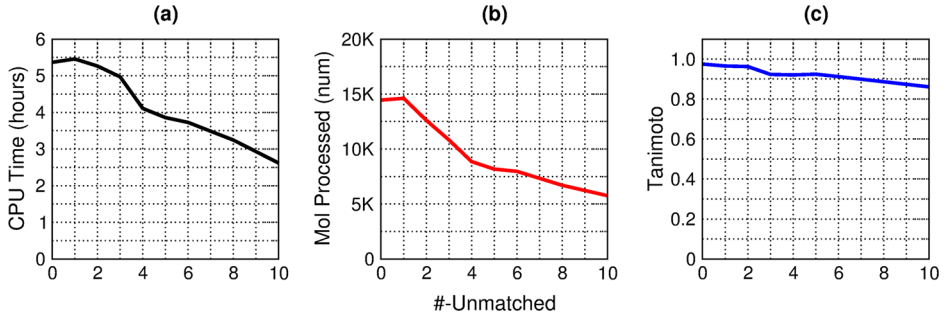


Figure 8. (a) CPU time required to complete the de novo experiment, (b) total number of molecules considered during the de novo experiment, and (c) best Tanimoto coefficient to the parent compound plotted with respect to the user-defined #-unmatched threshold. Results are averaged over all systems with 7 rotatable bonds from the SB2012 test set (N = 103).

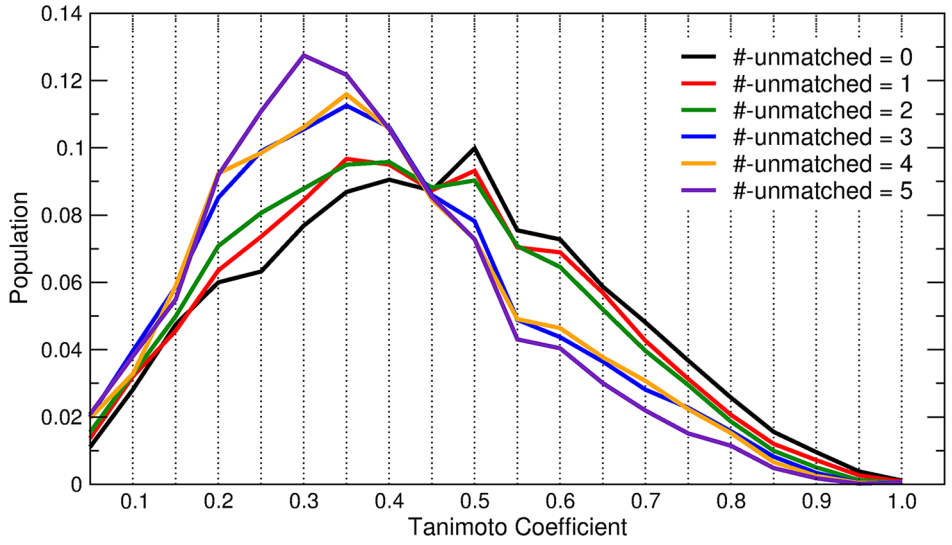


Figure 9. Histograms of Tanimoto coefficients for all pairwise comparisons within de novo molecule ensembles. Results are binned in 0.05 coefficient increments over all systems with 7 rotatable bonds from the SB2012 test set (N = 103).

in less CPU time and with fewer total molecules sampled) and (2) yield increased diversity in final molecule ensembles constructed from generic fragment libraries.

As shown in Figure 8, the first set of experiments designed to promote ligand growth was evaluated on the basis of CPU-time, number of molecules processed, and Tanimoto coefficient computed with respect to each of 103 parent compounds. Using the new Hungarian algorithm-based pruning heuristic, a sharp decrease in CPU time is observed (Figure 8a) from approximately 5.4 → 2.6 h as the number of unmatched atoms (#-unmatched) threshold increases from 0 to 10. Similarly, the total number of molecules processed shows a marked decrease from approximately 15K → 5K molecules over the same threshold (Figure 8b). Both of these indicators demonstrate that the de novo experiments are finishing more quickly and more efficiently. Importantly, the average best Tanimoto coefficient observed during the experiment decreases only gradually from 0.97 → 0.89 (Figure 8c). Thus, the parent molecules or analogs that are very similar to the parent molecules are being rebuilt with high frequency in approximately half the time, as was the goal of the experiment.

In the second set of experiments, we performed de novo growth using the larger seeded generic libraries (see Computational Details) with a variable threshold of 0−5 for #-unmatched (Figure 9). The focus here was to test our hypothesis that increasing the #-unmatched threshold in the heuristic will increase diversity among the final molecule ensemble. Following de novo growth of new compounds, the diversity among members in each ensemble was determined by computing all the possible pairwise Tanimoto coefficients between all molecules constructed. For example, an ensemble of 50 molecules would have 50 × 50 pairs, minus 50 self−self-comparisons, for a total of 2450 Tanimoto values. For a given #-unmatched threshold, Tanimoto coefficients from different systems were compiled, then histogrammed into discrete bins as shown in in Figure 9.

Figure 9 shows six final populations of molecules that each characterize different distributions of Tanimoto coefficients. As the distribution shifts to the right, molecules within the ensembles are more similar to one another (closer to 1.0) and thus have less diversity. As the distribution shifts to the left, molecules within the ensembles are less similar to one another (closer to 0.0) and thus have greater diversity. Consistent with our expectations, the peak shifted furthest to the right (corresponding to the least diversity) occurs when the #-unmatched term is set to 0 (Figure 9, black line). And, the de novo ensemble shifted furthest to the left (corresponding to the most diversity) occurs when the #-unmatched term is set to 5 (Figure 9, purple line). These data indicate that, during de novo growth, the Hungarian algorithm is successfully identifying chemically similar molecules that overlap sufficiently in Cartesian space, then pruning the population to make room for more diversity.

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

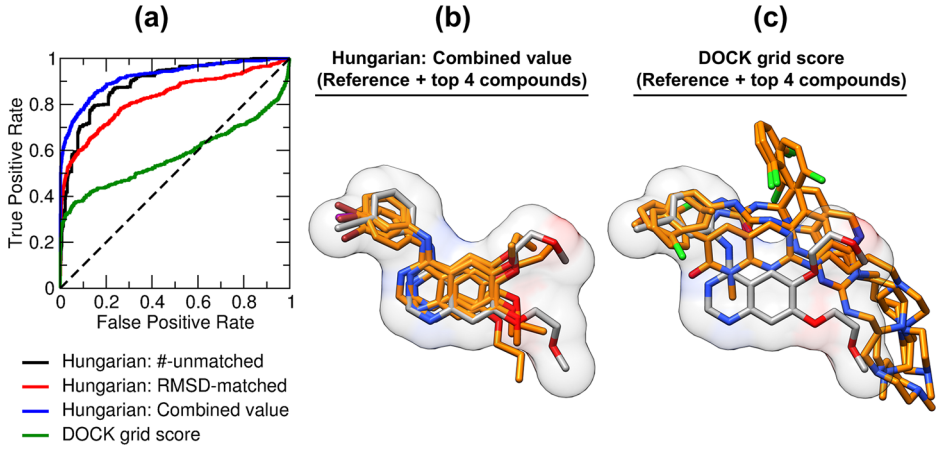


Figure 10. (a) ROC curves from docking 475 actives and 15 990 decoys to EGFR, then ranking the results by one of four metrics. The diagonal, indicated by a dashed line, represents random enrichment. (b, c) Crystallographic pose of reference ligand erlotinib shown as gray sticks and transparent surface, identical in both panels, relative to the top four ligands from the (b) combined value, shown as orange sticks, or the top four ligands from the (c) DOCK grid score, shown as orange sticks.

Extended Applications. In addition to the applications described above, two extended applications of the Hungarian algorithm are envisioned: (1) as a scoring function to rank-order docked ligands from virtual screening based on similarity to a known reference and (2) as a tool to analyze ligand behavior during a molecular dynamics (MD) simulation.

Scoring Function. To illustrate the first extended application, we examined docked results from an enrichment study previously performed in our laboratory, in which Balius et al.[40] docked 475 known active and 15 990 decoy compounds to the target epidermal growth factor receptor (EGFR, PDB 1M17).[46] The adapted version of the Hungarian algorithm for comparing dissimilar molecules was used to rerank that docked ensemble by comparing each docked pose to the cognate inhibitor (i.e., reference) 4-anilinoquinazoline (erlotinib) in its crystallographic binding pose with EGFR. The total number of unmatched atoms (#-unmatched) and the RMSD between the matched atoms (RMSD-matched) were returned for each ligand. We hypothesized that many of the known actives would have significant spatial and chemical overlap with the erlotinib reference, and thus would be ranked early in the list using the Hungarian approach, thereby providing good enrichment. Enrichment was evaluated using four unique ranking methods: (1) first by #-unmatched, and in case of a tie, then by RMSD-matched, (2) first by RMSD-matched, and in case of a tie, then by #-unmatched, (3) by a combination of the #-unmatched and RMSD-matched as shown in eq 5, and (4) the standard DOCK grid score.

$$\text{combined value} = C_1\left(\frac{\text{\#-ref atoms} - \text{\#-unmatched}}{\text{\#-ref atoms}}\right) + C_2(\text{RMSD matched}) \quad (5)$$

For the ranking method described in eq 5, #-ref atoms is the number of non-hydrogen atoms in the reference ligand, and C_1 and C_2 are constants of −5 and 1, respectively, chosen such that the two terms are weighted approximately equally, and that greater negative values represent more favorable scores. Figure 10 presents the receiver operating characteristic (ROC) curves derived from each of the four ranked lists along with structural overlays for top-scoring compounds from two of the four methods. While visualization of the ROC curves can provide a qualitative assessment of enrichment, results are also quantified below using area under the curve (AUC) computed across the entire database (overall enrichment), and the number of actives recovered after screening 1% of the database (early enrichment).

Quantitatively, the AUC data from Figure 10a reveal the following trend in terms of overall enrichment: theoretical maximum (1.00) > combined value (0.92) > #-unmatched (0.89) > RMSD-matched (0.83) > DOCK grid score (0.58) > theoretical random (0.50). All four methods provide substantially better enrichment relative to random, and the three Hungarian-based approaches provide enhanced enrichment (Figure 10 black, red, blue lines) relative to the standard DOCK grid score (green line), in accordance with our hypothesis. In terms of early enrichment, often considered to be a more useful metric for real-world applications of virtual screening, the number of actives recovered at 1% of the database screened are theoretical maximum (165) > combined value (158) > RMSD-matched (131) > DOCK grid score (113) > #-unmatched (85) > theoretical random (5). Again, all four methods provide significantly enhanced enrichment relative to random, and two of the three Hungarian methods provide enhanced early enrichment relative to the standard DOCK grid score. Upon visual examination, the four top-ranked compounds from the combined Hungarian method show much better Cartesian- and chemical-space overlap with the erlotinib reference (Figure 10b) compared to compounds obtained using the DOCK grid score, which are larger and have less well-overlaid groups (Figure 10c). Overall, while the effectiveness of the scoring function should be tested further on additional systems, these preliminary data indicate the method shows promise in enriching for compounds that are similar to a known reference molecule.

MD Analysis Tool. As an illustration of the second extended application, we explored the use of symmetry-corrected RMSD to characterize ligand dynamics as a function of time. We expected that typical sampling of a ligand during an MD simulation of a protein−ligand complex would lead to the occurrence of ring and other symmetric flips which would be discernible through comparisons of standard vs symmetry-corrected RMSD. Figure 11 plots results derived from two protein−ligand trajectories from MD simulations performed previously in our laboratory:[47] (1) T4 lysozyme in complex with

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

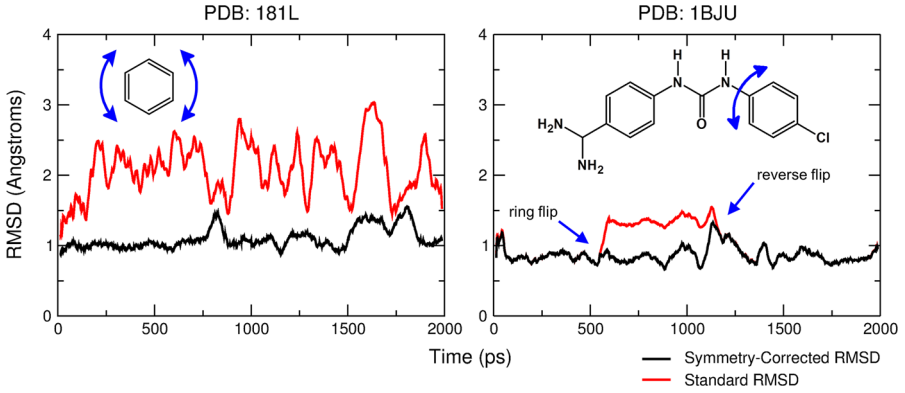


Figure 11. Symmetry-corrected (black lines) vs standard (red lines) heavy-atom RMSDs for two ligands during MD simulation. (Left) T4 lysozyme simulated in complex with benzene, PDB 181L. (Right) β-Trypsin simulated in complex with 1-(4-amidinophenyl)-3-(4-chlorophenyl)urea, PDB 1BJU. Lines are running averages with window size 50.

benzene (PDB 181L),[48] and (2) β-trypsin in complex with 1-(4-amidinophenyl)-3-(4-chlorophenyl)urea (PDB 1BJU).[49] Here, MD coordinate files saved every 1 ps were least-squares fit to the starting frame (using protein α-carbons as the reference), followed by the calculation of standard (Figure 11, red) and symmetry-corrected (Figure 11, black) RMSDs for the ligand. As a practical note, in the present example the ligand coordinates were extracted and then postprocessed using the Hungarian algorithm code in DOCK. In theory, however, there is no barrier to incorporating the code directly into an MD simulation or analysis package.

In the first example (Figure 11, left), the benzene ligand was observed to rotate both clockwise and counterclockwise about an axis normal to the plane of the ring (as indicated by blue arrows). Although no ring "flipping" was observed in this simulation, likely because of steric constraints imposed by the binding site, large variations in standard RMSD were measured as a direct result of the rotational motion (Figure 11, red). Importantly, the symmetry-corrected RMSD (Figure 11, black) effectively corrects for symmetry in the benzene, thus highlighting variations due to either translation or out-of-plane rotations, which in both instances are minimal. In the second example (Figure 11, right), the ligand remained fairly close to its original binding configuration during the MD simulation. However, at approximately 550 ps, the 4-chlorophenyl group flipped 180° about a rotatable bond (as indicated by blue arrows). Corresponding to this flip, a spike in the standard RMSD was observed, which was not reflected in the symmetry-corrected RMSD (Figure 11, red vs black). And, at approximately 1250 ps, the same ring flips 180° back to the original orientation, and the two RMSD measures reconverge. It should be noted that although this event was not easily detected visually, it was readily apparent in the plotted graphs. Overall, although more exhaustive tests should be done, the two extended applications presented in this manuscript suggest the Hungarian algorithm will have use in both virtual screening and ligand sampling analysis.

CONCLUSIONS

In this work, we describe implementation, validation, and application of the Hungarian algorithm within the DOCK6 program to compute symmetry-corrected RMSD. The method is fast, on the order of less than a millisecond per ligand (Figure 4), and therefore adds only a trivial amount of time to on-the-fly flexible ligand docking in DOCK6. More importantly, the negligible amount of time required of the algorithm allows its use in de novo design. In all, 88 853 docked poses from the SB2012 test set of 1043 receptor−ligand systems were evaluated using the new method. While the majority of poses showed a <1.0 Å improvement in RMSD, indicative of local corrections within symmetric functional groups, a smaller but significant set yielded >3.0 Å improvement, indicative of higher symmetry corrections (Figure 5). A closer examination of just the top-scoring pose (Table 1) from each system revealed 67.7% were true positives, or considered as successfully docked by either metric, and 5.0% were false negatives, or considered to be rescued by the symmetry-correction algorithm, for a total success rate of 72.7%.

Visualization of specific examples of rescued poses (Figures 6 and 7) reveals that the method is behaving as expected. Highly symmetric molecules with nearly perfect overlap to crystallographic references but inverted over the axis of symmetry are identified and rescued, and some molecules with local pockets of symmetry are also rescued. Analysis of results grouped by specific protein families reveals more dramatic increases in success rate up to 16.7%, including the challenging HIV protease system, for which an increase in success rate of 13.3% over 60 systems was observed (Table 2).

The Hungarian method was also adapted to be used as a heuristic for comparing spatial and chemical overlap between two dissimilar molecules to aid in pruning in a de novo design version of DOCK undergoing development. Preliminary tests demonstrate that by using the heuristic, sampling remains relatively unaffected, while time and efficiency of computation are significantly improved (Figure 8). In addition, diversity among molecules in the final growth ensembles is higher, which for these purposes was desirable (Figure 9). If preferred, it would be straightforward to modify the function in order to *decrease* chemical diversity in growing de novo populations.

Finally, two extended applications of the Hungarian algorithm were briefly introduced. First, the adapted version of the algorithm was used as a scoring function to rank-order compounds from a large database and thus identify those related to a known reference molecule (Figure 10). Compounds with high overlap using the Hungarian metric are both spatially and chemically similar to the reference compound. Second, the algorithm was used to compute symmetry-corrected RMSDs in the course of an MD simulation (Figure 11). While additional testing is desirable to more fully establish the utility of the algorithm for these extended applications, the work presented is a reasonable proof-of-concept demonstration.

dx.doi.org/10.1021/ci400534h | J. Chem. Inf. Model. 2014, 54, 518−529

The accurate determination of success and failure in docking calculations is essential for evaluation of new sampling routines, scoring functions, protocols, and other code developments. Prior to the current release, the program DOCK contained no method to efficiently correct for symmetry when computing RMSD for small molecule ligands. Now, an extensively validated and detailed method has been described. The source code for this Hungarian algorithm implementation is available to registered users of DOCK as part of the latest release at http://dock.compbio.ucsf.edu.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rizzorc@gmail.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors express gratitude to Trent E. Balius and Sudipto Mukherjee for helpful discussions and support during the early stages of this work, Jiangyang Liu for technical assistance, and Steven Skiena and Esther M. Arkin for insights into the Hungarian algorithm. This work was supported by the National Institutes of Health, grant numbers F32GM105400 (to W.J.A.) and R01GM083669 (to R.C.R.).

REFERENCES

(1) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078−1082.

(2) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862−865.

(3) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.

(4) Yuriev, E.; Ramsland, P. A. Latest Developments in Molecular Docking: 2010−2011 in Review. *J. Mol. Recognit.* **2013**, *26*, 215−239.

(5) Chen, Y.; Shoichet, B. K. Molecular Docking and Ligand Specificity in Fragment-Based Inhibitor Discovery. *Nat. Chem. Biol.* **2009**, *5*, 358−364.

(6) Kolb, P.; Rosenbaum, D. M.; Irwin, J. J.; Fung, J. J.; Kobilka, B. K.; Shoichet, B. K. Structure-Based Discovery of β2-Adrenergic Receptor Ligands. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6843−6848.

(7) Neher, T. M.; Shuck, S. C.; Liu, J.-Y.; Zhang, J.-T.; Turchi, J. J. Identification of Novel Small Molecule Inhibitors of the XPA Protein Using in Silico Based Screening. *ACS Chem. Biol.* **2010**, *5*, 953−965.

(8) Daldrop, P.; Reyes, F. E.; Robinson, D. A.; Hammond, C. M.; Lilley, D. M.; Batey, R. T.; Brenk, R. Novel Ligands for a Purine Riboswitch Discovered by RNA-Ligand Docking. *Chem. Biol.* **2011**, *18*, 324−335.

(9) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: Combining Techniques to Model RNA-Small Molecule Complexes. *RNA* **2009**, *15*, 1219−1230.

(10) Brozell, S. R.; Mukherjee, S.; Balius, T. E.; Roe, D. R.; Case, D. A.; Rizzo, R. C. Evaluation of DOCK 6 as a Pose Generation and Database Enrichment Tool. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 749−773.

(11) Holden, P. M.; Kaur, H.; Goyal, R.; Gochin, M.; Rizzo, R. C. Footprint-Based Identification of Viral Entry Inhibitors Targeting HIVgp41. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 3011−3016.

(12) Berger, W. T.; Ralph, B. P.; Kaczocha, M.; Sun, J.; Balius, T. E.; Rizzo, R. C.; Haj-Dahmane, S.; Ojima, I.; Deutsch, D. G. Targeting Fatty Acid Binding Protein (FABP) Anandamide Transporters—A Novel Strategy for Development of Anti-Inflammatory and Anti-Nociceptive Drugs. *PLoS One* **2012**, *7*, No. e50968.

(13) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock and AutoDockTools: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(14) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(15) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609−623.

(16) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(17) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(18) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(19) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking Benchmarks and Real-World Application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687−699.

(20) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(21) Bagaria, A.; Jaravine, V.; Huang, Y. J.; Montelione, G. T.; Guentert, P. Protein Structure Validation by Generalized Linear Model Root-Mean-Square Deviation Prediction. *Protein Sci.* **2012**, *21*, 229−238.

(22) Helmich, B.; Sierka, M. Similarity Recognition of Molecular Structures by Optimal Atomic Matching and Rotational Superposition. *J. Comput. Chem.* **2012**, *33*, 134−140.

(23) Ellingson, L.; Zhang, J. Protein Surface Matching by Combining Local and Global Geometric Information. *PLoS One* **2012**, *7*, No. e40540.

(24) Mukherjee, S.; Balius, T. E.; Rizzo, R. C. Docking Validation Resources: Protein Family and Ligand Flexibility Experiments. *J. Chem. Inf. Model.* **2010**, *50*, 1986−2000.

(25) Balius, T. E.; Mukherjee, S.; Rizzo, R. C. Implementation and Evaluation of a Docking-Rescoring Method Using Molecular Footprint Comparisons. *J. Comput. Chem.* **2011**, *32*, 2273−2289.

(26) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* **2007**, *371*, 1118−1134.

(27) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83−97.

(28) Munkres, J. Algorithms for the Assignment and Transportation Problems. *J. Soc. Indust. Appl. Math.* **1957**, *5*, 32−38.

(29) Ignizio, J. P.; Cavalier, T. M. *Linear Programming*; Prentice-Hall, Inc.: Upper Saddle River, NJ, 1994.

(30) *Tripos Mol2 File Format*; Tripos: St. Louis, MO, 2009.

(31) *MOE*; Chemical Computing Group: Montreal, Canada, 2009.

(32) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.

(33) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.

(34) Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AmberTools*; University of California at San Francisco: San Francisco, CA, 2010.

(35) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(36) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and Validation of a Modular, Extensible Docking Program: DOCK 5. *J. Comput.-Aided Mol. Des* **2006**, *20*, 601−619.
(37) *DMS*; University of California at San Francisco Computer Graphics Laboratory: San Francisco, CA, 2003, p http://www.cgl.ucsf.edu/Overview/software.html.
(38) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure. *J. Med. Chem.* **1988**, *31*, 722−729.
(39) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(40) Balius, T. E.; Allen, W. J.; Mukherjee, S.; Rizzo, R. C. Grid-Based Molecular Footprint Comparison Method for Docking and de novo Design: Application to HIVgp41. *J. Comput. Chem.* **2013**, *34*, 1226−1240.
(41) Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(42) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(43) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.
(44) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.
(45) Bokma, E.; Rozeboom, H. J.; Sibbald, M.; Dijkstra, B. W.; Beintema, J. J. Expression and Characterization of Active Site Mutants of Hevamine, a Chitinase from the Rubber Tree *Hevea brasiliensis*. *Eur. J. Biochem.* **2002**, *269*, 893−901.
(46) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the Epidermal Growth Factor Receptor Kinase Domain Alone and in Complex with a 4-Anilinoquinazoline Inhibitor. *J. Biol. Chem.* **2002**, *277*, 46265−46272.
(47) Holden, P. M.; Allen, W. J.; Gochin, M.; Rizzo, R. C. Strategies for Lead Discovery: Application of Footprint Similarity Targeting HIVgp41. *Bioorg. Med. Chem.* **2013**, DOI: 10.1016/j.bmc.2013.1010.1022.
(48) Morton, A.; Matthews, B. W. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry* **1995**, *34*, 8576−8588.
(49) Presnell, S. R.; Patil, G. S.; Mura, C.; Jude, K. M.; Conley, J. M.; Bertrand, J. A.; Kam, C.-M.; Powers, J. C.; Williams, L. D. Oxyanion-Mediated Inhibition of Serine Proteases. *Biochemistry* **1998**, *37*, 17068−17081.