Scoring Ensembles of Docked Protein:Ligand Interactions for Virtual Lead Optimization

Janet L. Paulsen and Amy C. Anderson*

Department of Pharmaceutical Sciences, University of Connecticut, 69 North Eagleville Road, Storrs, Connecticut 06269

Received August 17, 2009

Ensembles of protein structures to simulate protein flexibility are widely used throughout several applications including virtual lead optimization where they have been shown to improve ligand ranking. Yet, there is no established convention for weighting individual scores generated from ensemble members. To investigate the best method for weighting ensemble scores for proper ligand ranking, a series of dihydrofolate reductase inhibitors was docked to ensembles of *Candida albicans* dihydrofolate reductase (CaDHFR) structures created from a molecular dynamics (MD) simulation. From a single MD simulation, two ensemble collections were generated, one of which was subjected to a minimization procedure to create a group of structures of equal probability. As expected, ligand ranking accuracy was significantly improved when Boltzmann weighting was applied to the energies of the ensemble without structural minimization (60%), relative to that achieved with averaging (36%). However, accuracy was further improved (72%) by averaging docking scores across a minimized ensemble. To examine whether this accuracy results from structural variation in the single trajectory versus the possibility that error is minimized by averaging, a third collection of receptor structures was created in which each member was taken from an independent molecular dynamics simulation after minimization. Comparison of the docking accuracy results from the single trajectory (72%) to this third collection (61%) showed decreased accuracy, suggesting that ligands are more accurately oriented and assessed when docked to the minimized ensemble from a single MD trajectory, an effect that is more than simply error minimization. Averaging docking scores over a minimized ensemble of another target, influenza A neuraminidase, yielded a ligand ranking accuracy of 83%, representing a 24% improvement over other methods tested.

INTRODUCTION

Ensembles of protein structures are valuable representations of conformational flexibility that find utility in drug design, protein folding studies, enzyme mechanism, and protein redesign. Comparing ensembles, as is often practiced within these applications, necessitates the consolidation of appropriately weighted individual scores or energies of the ensemble members to arrive at a single evaluation for the entire ensemble. A priori, it may appear that a Boltzmann weighting scheme in which the ensemble follows a probability distribution function would be most appropriate. However, the literature describes several popular methods for evaluating ensembles including averaging,[1−4] Boltzmann weighting,[5] creating a composite grid or united description of the ensemble,[6,7] and decision trees or algorithms to determine the most essential member(s).[8−10] While much research on ensembles has been conducted, there is no single convention for evaluation, and to our knowledge the basis for successful methods is rarely investigated.

While ensembles are valuable to a variety of applications, we are specifically interested in virtual lead optimization, in which designed analogues are accurately docked and ranked, thus prioritizing improved compounds for synthesis. Accurate ranking is paramount to successful virtual lead optimization, which often suffers from the inability to reproduce the absolute empirical rank order. Both ligand and receptor flexibility have been shown to substantially influence ranking accuracy. Many docking programs have algorithms for handling ligand flexibility; however, receptor flexibility is only marginally considered.[11] In previous work, we[1,3,4,12] and others[2,5−10,13−15] have established that the use of receptor ensembles created using NMR, molecular dynamics simulations, or multiple experimentally determined models aids in representing a flexible receptor. In addition, we have shown that docking to an ensemble leads to increased ranking accuracy for virtual lead optimization[3] and that the ensembles can be pruned a priori to include only the critical members that maintain a conserved binding core.[4]

Here, we evaluate the most efficient treatment of the ensemble member scores as they directly affect ranking accuracy during lead optimization. First, we tested the a priori idea that a Boltzmann weighting scheme would be most appropriate by creating an ensemble of structures that follows a probability distribution function from *Candida albicans* dihydrofolate reductase (CaDHFR) and docking a series of related ligands with known affinities. Ranking accuracy was determined by using a neighbor binning technique that clusters ligands according to their potency. Higher ranking accuracy reflects a larger number of occurrences in which both the empirical and the in silico scores place a ligand into its correct group. Comparing the effects of applying Boltzmann weighting or averaging the energy of the resultant complexes on ranking accuracy shows that Boltzmann weighting prevailed

* Corresponding author phone: (860) 486-6145; fax: (860) 486-6857; e-mail: amy.anderson@uconn.edu.

10.1021/ci9003078 © 2009 American Chemical Society
Published on Web 12/02/2009

as expected (60% versus 36%). Interestingly, however, results from docking to a second collection of structures that were energy minimized prior to use and therefore are all equally probable show that the most accurate ligand ranking results for CaDHFR (72% of ligands are properly ranked) are obtained by averaging empirical docking scores across an ensemble. To assess if this observation is the result of minimizing error by averaging (increasing signal-to-noise) or an intrinsic property of the minimized ensemble, ligands were docked to a collection of structures created from snapshots at 25 ps of 21 different molecular dynamics simulations. The ranking accuracy for this ensemble was decreased to 61%, suggesting that the properties of the ensemble based on the single trajectory are important to the improved ranking accuracy. A structural analysis of the ensembles illustrates the importance of each conformation of both the protein and the ligand. To further validate these methods for ensemble treatment, we create and evaluate an ensemble of a different target, influenza A neuraminidase (NA). The ligand ranking accuracy obtained for this target is 83% when docking scores are averaged over the minimized ensemble, representing a 24% improvement over alternative methods.

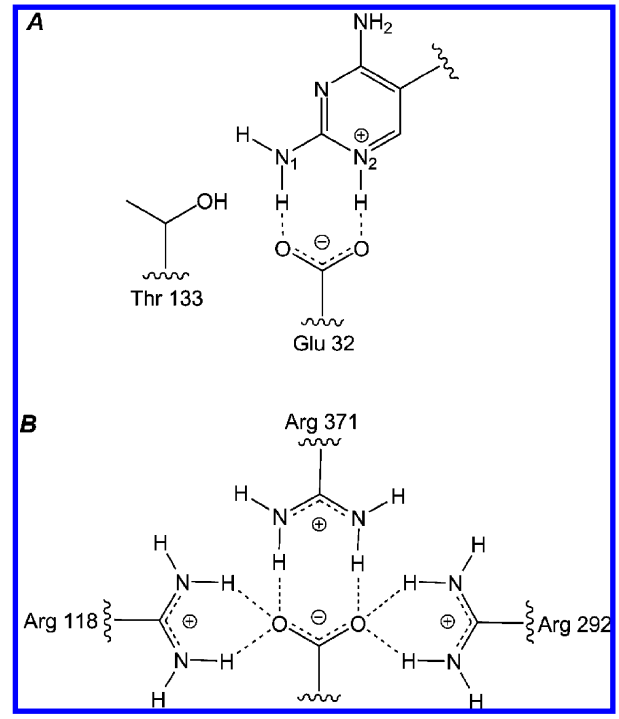

Figure 1. Conserved binding core of (A) CaDHFR with a 2,4-diaminopyrimidine and (B) influenza NA with a carboxylate moiety.

MATERIALS AND METHODS

Receptor Preparation. A structure of the CaDHFR docking receptor was obtained from the Protein Data Bank (PDB) as a crystallographic ternary complex with NADPH and an antifolate (PDB ID: 1AOE[16]). Because two molecules were in the asymmetric unit, one molecule was deleted along with all modeled water molecules. Using the biopolymer structure preparation tool available within Sybyl,[17] hydrogens were added, charges assigned according to AMBER 7 FF99,[18] atoms were typed for AMBER 7 FF99, and missing side chains were added to the model. Occupancies were also checked to ensure that the dominant conformer was represented in the structure. A model of influenza A neuraminidase bound to an antiviral agent (PDB ID: 1NNC[19]) was also prepared following the procedure outlined above. This model showed several glycosylated residues that were left intact in the prepared model; none were near the active site.

Ensemble Generation. A molecular dynamics (MD) simulation was used to create the necessary ensembles. A 3.5 Å shell around a properly oriented ligand of a prepared model was defined as flexible, and any residue with at least one atom in this shell was allowed to move during the MD simulation. In addition, residues 58−66 (CaDHFR) forming a loop that flanks the active site were also allowed to be flexible during the simulation. To preserve the binding core of CaDHFR and influenza A NA, key atoms were constrained during the MD simulation. For CaDHFR, the carboxyl oxygens of Glu 32 were restricted to a distance between 1.6−2.75 Å of the N_1 and N_2 hydrogens of the ligand (see Figure 1A). The conserved binding core of NA is created by an arginine triad that forms a hydrogen-bond network with a carboxylate group on the ligand. Distances between atoms that form key hydrogen bonds as shown in Figure 1B were also restricted to 1.6−2.75 Å.

The simulations were performed using Sybyl 8.0 following a canonical ensemble (NTV) at 300 K using the AMBER 7 F99 force field. Snapshots were taken at 500 fs intervals from 20 ps to 30 or 40 ps with a time step of 1 fs creating ensembles of 21 or 42 members, respectively. The first 20 ps of the simulation was discarded to allow convergence prior to ensemble collection (see Supporting Information Figure 1). Solvent was implicit. The velocities of the atoms were scaled at each step so that the kinetic energy of the system corresponds to the specified temperature. Three collections of protein structures were created. In the first collection, receptor structures were used as is (without minimization), in the second collection each ensemble member was minimized using AMBER 7 FF99 parameters, and the third collection was created from 21 independent MDs captured at 25 ps followed by minimization using AMBER 7 FF99 parameters. While the second and third collections are not ensembles in the statistical mechanical sense, we will continue to refer to these collections of protein structures as ensembles to reflect their unity as a collection. Formal charges were assigned and the active site cleared of ligands for all ensemble collections.

Ligand Preparation. The synthesis and characterization of a propargyl-linked library of DHFR inhibitors has previously been reported.[20−24] CaDHFR enzyme inhibition for this library has also been determined.[25] Ligands in this library were created in silico using Sybyl 8.0, assigned charges using Gasteiger−Marsili protocols, and minimized using the Tripos[17] force-field accompanying Sybyl. Minimization proceeded until the energy change between successive cycles reached zero (generally 1000 cycles). Sixty-seven ligands with IC_{50} values from 17 nM to 11.9 μM were used for this investigation. For racemic ligands, both enantiomers were drawn and individually docked.

Neuraminidase ligands and values for activity against influenza A neuraminidase were taken from the literature.[26−28] Ligand overlap between sources suggests a similar experimental procedure for enzyme inhibition, allowing for cross-comparison. Twenty-nine NA ligands with IC_{50} values from

2 μM to 5.5 mM (see Supporting InformationTable 1) were drawn, charges added, and minimized akin to the ligand preparation procedure for CaDHFR.

Docking. All docking was accomplished with Surflex−Dock[29] as implemented by Sybyl 8.0. For both CaDHFR and influenza A NA, all prepared ligands were docked to an ensemble of receptors. Top-scoring complexes were also assessed for compliance with the geometry of the respective binding core shown in Figure 1.

Scoring Metrics. Both empirical and force-field-based scoring functions were used to rank ligands in silico. Surflex−Dock uses an empirical-based scoring function. It is a continuous differentiable nonlinear function with terms for hydrophobic and polar complementarity, as well as entropic and solvation effects. Surflex−Dock's scoring function has been described rigorously elsewhere.[30−32] In brief, each atom of the receptor and ligand is assigned as polar or nonpolar; for each protein:ligand atom pair, the distance between the van der Waals surfaces is calculated. If this distance exceeds 2 Å or if another atom transects the atom pair, the pair is discarded. The training set used to derive the Surflex−Dock scoring function consisted of 34 protein:ligand complexes with known binding affinities.[31] Scores obtained using the Surflex−Dock scoring function are related to binding affinities by $-\log(K_d)$; therefore, higher scores indicate stronger affinity.

Ligand ranking was also performed using force-field-based scoring by calculating the energies of the complex. All energies were calculated using the Tripos force-field and include terms for bond stretching, bending, out of plane bending, torsion, van der Waals, electrostatics, and an energy penalty associated with deviations from the distance range constraint.[17]

Ensemble Treatment. Once docking scores for each ligand against each ensemble member were obtained, the ensemble of scores was treated with one of three methods to obtain one score per ligand.

(1) A Boltzmann distribution function (eq 1) was applied to the scores.

$$\frac{N_i}{\sum N_i} = \frac{\exp^{-E_i/kt}}{\sum \exp^{-E_i/kt}} \quad (1)$$

E is the change in energy from the lowest energy state, k is the Boltzmann constant, and t is the temperature of the system. The fractional occupancy of the ith state is $N_i/\sum N_i$. This probability was applied to the energy of the complex for each ensemble member to arrive at a single scoring metric for the ensemble.

(2) Scores were averaged across the ensemble so that each ensemble member received equal weight.

(3) The best scores for each protein:ligand complex were extracted from the ensemble and subsequently used in ranking. In the case of energy scores (force-field-based approach), the lowest energy is considered the best score. Using Surflex−Dock, the highest docking scores are considered optimal.

Evaluation of Ranking Accuracy. Ligand ranking accuracy was assessed using a neighbor binning technique. Ligands were ranked according to empirically determined enzyme inhibition activity and divided into bins or clusters based on their affinities (see Supporting InformationTables 3 and 4). The bin definition is kept constant for the purposes of comparison. Ligands were also ranked according to their in silico scores after ensemble treatment. If the in silico ranking placed a ligand into the same bin as the empirical ranking, that ligand received a score of one, otherwise it received a scored of zero. If a ligand was placed in a bin neighboring its correct bin, it also received a scored of one, allowing bins to have soft boundaries. The accuracy of the ranking was measured by determining the percentage in silico scores placed ligands into their correct bin or a neighboring bin as determined by the empirical ranking. For the binning scores to be statistically relevant, they must be greater than random. Therefore, the binning scores were compared to the binning score obtained by random ligand placement to determine if this condition was true.[3,25]

RESULTS

Despite the growing use of ensembles in docking, there is little consensus regarding the most efficient and accurate treatment of the scores resulting from the individual members. Clearly, accurately calculating an overall score for the ensemble is critical because the overall score establishes the rank order of the complex in the context of other scored complexes. To investigate the best method for calculating an ensemble-based score for the process of virtual lead optimization, we calculated and ranked ensemble-based scores for ligands with known affinity and then used the resulting ranking accuracy as the primary metric. We establish the superiority of applying the Boltzmann distribution function to an ensemble that follows a probability distribution function (i.e., has not been minimized), then show that averaging scores over a minimized ensemble leads to improved accuracy. Also, we performed experiments to determine if the improvement observed using minimized ensembles was the result of increasing the signal-to-noise ratio or another intrinsic property of the ensemble. Last, in an effort to refine and understand ensembles further, we compared force-field-based versus empirical scoring, the use of an ensemble versus a single member, and the effects of ensemble size. With this target and group of ligands, we again reinforce the value of the use of an ensemble. Interestingly, the most accurate ranking results were obtained by equally weighting all members of a minimized ensemble and by using an empirical scoring scheme.

Candida albicans dihydrofolate reductase (CaDHFR) was chosen as the initial target receptor for this study because it maintains a conserved ligand binding core, allowing docking orientations to be quickly evaluated for compliance. In addition, we have amassed experimental binding values (inhibition concentration) for a series of analogues based on the propargyl-linked antifolate lead.[25] A crystal structure of CaDHFR was subjected to a molecular dynamics simulation to create an ensemble of structures. Two collections of structures were initially created: one collection was minimized, while the other was not. Each ligand was docked to each member of both ensembles, and each complex was scored using a force-field and/or empirical scoring approach. The resultant scores from the ensembles were treated by a variety of methods including application of the Boltzmann distribution function, averaging, evaluation by the lowest energy complex, and evaluation by the highest Surflex−Dock

Table 1. Effect of Ensemble Treatments on Ligand Ranking Accuracy for CaDHFR

ensemble type/ensemble treatment for CaDHFR	binning score	NMC >1%[a]
30[b] ps no_min ensemble/Boltzmann to energy of complex	60%	1.3
30 ps no_min ensemble/avg energy of complex	36%	21
30 ps min ensemble/avg SD[c]	72%	21
25 ps min snapshots from 21 random seed MDs/avg SD	61%	21
40[d] ps min ensemble/avg SD	73%	42
30 ps min ensemble/avg energy of complex	60%	21
40 ps min ensemble/avg energy of complex	60%	42
30 ps min ensemble/lowest energy complex	58%	1
30 ps min ensemble/highest SD	63%	1
40 ps min ensemble/highest SD	63%	1
crystal structure 1AOE/SD	66%	1
random ligand placement	17%	NA

[a] Number of ensemble members contributing greater than 1% to binning score. [b] Ensemble members were collected between 20 and 30 ps. [c] Surflex−Dock scores. [d] Ensemble members were collected between 20 and 40 ps.

scores. Ranking was assessed using a binning method whereby ligands are divided into bins by their empirical affinities. The aggregate binning score reflects the percentage of ligands that rank in the same bin as determined in silico or empirically (Table 1).

Accurate Ligand Ranking Using Ensembles. It is well established that an ensemble will follow the Boltzmann distribution function if the ensemble adheres to the probability distribution function. To establish this a priori result, the energies of the resultant docked complexes using the ensemble without minimization were weighted by the Boltzmann function and also averaged across the ensemble. The Boltzmann weighting yields more accurate ranking (60%) versus averaging across the ensemble (36%) (see Table 1). Clearly, not all members of the ensemble are equally probable, and the Boltzmann weighting function accurately establishes the proper weighting.

In an effort to improve upon this established 60% accuracy of the Boltzmann weighting as applied to an ensemble without minimization, a second minimized ensemble was evaluated. Because this collection of structures was minimized, all receptor structures should be equally probable, suggesting that averaging across the ensemble is the most accurate evaluation. The resultant docked complexes were scored using Surflex−Dock, an empirical scoring function, and averaged across the ensemble, resulting in a 72% ligand ranking accuracy.

Error Minimization versus Flexibility. To further investigate whether the improved accuracy using the minimized ensemble and the empirical scoring function was the result of the intrinsic flexibility imparted to the ensemble by following a single trajectory or performing additional measurements, thus increasing the signal-to-noise ratio, a third experiment was performed. Twenty-one molecular dynamic simulations beginning with random seeds were performed (see Material and Methods). A snapshot at 25 ps was collected and minimized to create a third collection of structures independent of trajectory. Ranking accuracy for the 25 ps snapshots of the 21 random seed MDs was 61% as compared to the 72% ranking accuracy of the 30 ps ensemble along a single trajectory (see Table 1). These results suggest that the minimized ensemble along the single trajectory allows for more accurate ligand placement and assessment of interactions, leading to more accurate ligand ranking.

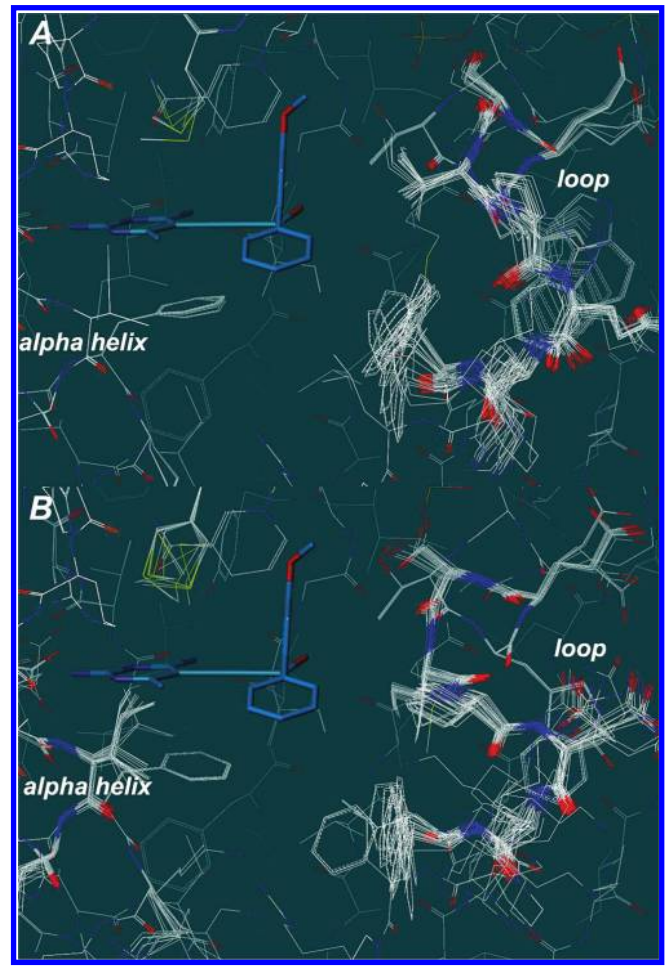

Figure 2. Ensembles of (A) 30 ps minimized ensemble of CaDHFR ensemble collected over a single trajectory and (B) 25 ps snapshots collected from 21 MD random seed simulations.

Further examination of the ensemble also suggests a structural basis for this observation. The range of Cα movement measured in the loop region that flanks the active site was found to be as much as 0.8 Å greater in the ensemble following a single trajectory versus the collection from 21 random seed MDs (see Figure 2). Also, the Cα displacement from the starting crystallographic model of the loop region was found to be as much as 2 Å greater in the ensemble following a single trajectory. While there appears to be more flexibility in the α helix flanking the active site from the 21 random seed MDs, it is the loop flexibility that appears to play a vital role in ligand ranking accuracy.

Force-Field versus Empirical Scoring. Using the collection of minimized structures, results show that employing empirical scoring using Surflex−Dock leads to significantly more accurate ligand ranking. The accuracies (binning scores) for the 30 ps ensemble where Surflex−Dock scores that were averaged across the ensemble versus the 30 ps ensemble where the energy of the complex was averaged across the ensemble were 72% and 60%, respectively (see Table 2).

Effect of Ensemble versus Single Structure on Ranking Accuracy. To determine the effect using an ensemble has on ranking accuracy, minimized single structure

Table 2. Effect of Ensemble Treatments on Ligand Ranking Accuracy for NA

ensemble type/ensemble treatment for NA	binning score	NMC >1%[a]
30[b] ps ensemble/avg SD[c]	83%	21
30 ps ensemble/avg energy	59%	21
30 ps ensemble/highest SD	59%	1
random ligand placement	24%	NA

[a] Number of ensemble members contributing greater than 1% to binning score. [b] Ensemble members were collected between 20 and 30 ps. [c] Surflex−Dock scores.

and ensemble rankings were compared for this model of CaDHFR. The complex with the lowest energy on a per ligand basis was extracted from the ensemble. Ranking based on these lowest energy complexes exhibited 58% accuracy. A slight improvement in accuracy (2%) was observed when the energies of the complex were averaged across the entire ensemble and subsequently used for ranking. However, a significant improvement in ranking accuracy was observed when the Surflex−Dock scores were averaged across the ensemble (72% accuracy). This value was compared to rankings using the highest Surflex−Dock scores extracted from the ensemble and the Surflex−Dock scores obtained by docking to the crystal structure PDB ID: 1AOE, which gave accuracies of 63% and 66%, respectively. These results suggest that docking and ranking to an ensemble is more accurate than using a single structure for docking and ranking.

Effect of Ensemble Size on Ranking Accuracy. To determine if the accuracy of any scoring or weighting metric was limited by the size of the ensemble, the ensemble size was doubled and compared to the results reported above. The results for the comparison of the 30 ps ensemble with 21 members and the 40 ps ensemble with 42 members are found in Table 1. Despite doubling the ensemble size, no ensemble treatment as applied to either scoring metric led to improved ligand ranking accuracy.

Validation of Results. Docking to another target, influenza A neuraminidase, validated the scoring and ensemble treatment results for CaDHFR, that empirical scores (Surflex−Dock) averaged across a minimized ensemble provide the most accurate ligand ranking. Like DHFR, NA has a conserved binding core and is a commonly used docking test case.[33] Scoring and ligand ranking accuracies for influenza A neuraminidase are found in Table 2.

First, to verify that empirical scoring yields better results relative to a force-field-based approach, each scoring method was averaged across the NA ensemble. Empirical scoring (Surflex−Dock) was 24% more accurate than using the average energies of the complex (83% versus 59%). Second, we verified that an equal weighting of scores across the ensemble was the best method for scoring the ensemble. Ranking using the highest Surflex−Dock scores was compared to ranking by averaging Surflex−Dock scores, yielding accuracies of 59% and 83%, respectively. By replicating our findings in another system, we concluded that averaging empirical scores across the ensemble provides the most accurate ligand ranking.

Structural Evidence. Thus far, we have introduced evidence showing that ensembles improve ranking accuracy relative to a single member and specifically that averaging

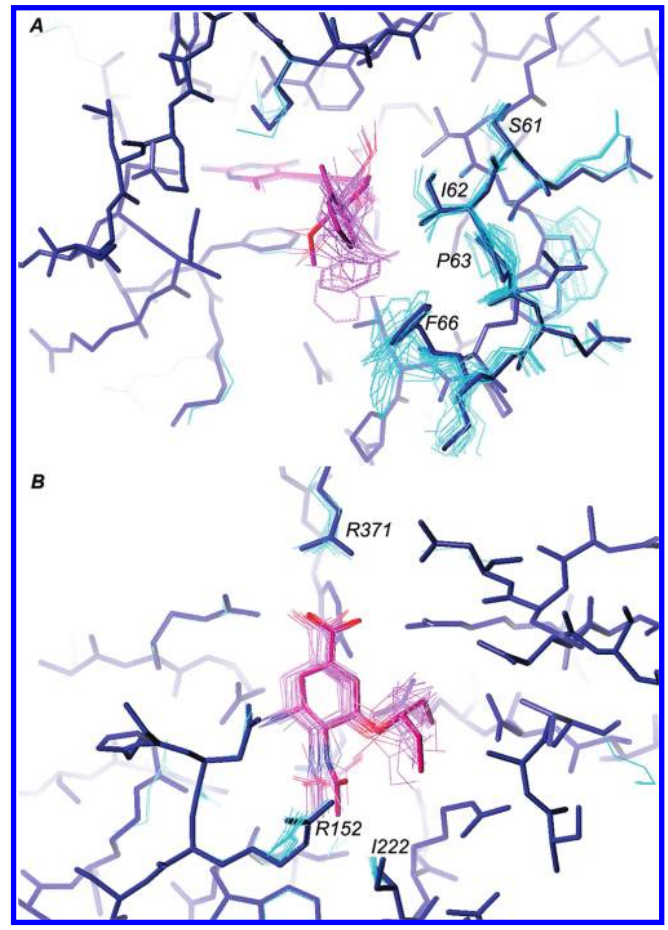

Figure 3. Ensembles of (A) CaDHFR bound to UCP11s1F2Me and (B) neuraminidase bound to oseltamivir (active metabolite). The ensemble members shown in sticks are the lowest energy complex.

the Surflex−Dock scores across the ensemble provides more accurate ranking than using the lowest energy member of the ensemble. As a consequence of these results, it is possible to identify several ligands that are improperly ranked using the lowest energy member relative to averaging across the ensemble. As examples, the two ligands UCP11s1F2Me (5-{3-[3-methoxy-4-(2-methylphenyl)phenyl]but-1-yn-1-yl}-6-methylpyrimidine-2,4-diamine) and oseltamivir were ranked too low if only the lowest energy complex was considered; averaging across the ensemble effectively moved these ligands up in ranking into their correct bins. Specifically, to receive a score of 1 and contribute positively to the accuracy count, UCP11s1F2Me should be ranked within the range 1−25. The compound ranked 17th using averaging but ranked 49th using only the lowest energy complex. For oseltamivir, the compound should be ranked within the range 1−10 to be counted positively; with averaging it ranked eighth and using the lowest energy complex it ranked 18th.

To determine the structural basis of the effect of the ensemble, all conformations of each member of these ensembles bound to these two ligands were compared to the conformation of the lowest energy complex. For the CaDHFR ensemble, the conformations of most residues in the active site superimpose. However, residues Ser 61, Ile 62, Pro 63, and Phe 66 in a loop adjacent to the active site (refer to Figure 2) are an important exception and exhibit many different conformations throughout the ensemble (Figure 3A). The ligand UCP11s1F2Me also adopts several conforma-

tions. The conformations of the lowest energy ensemble members and the lowest energy conformation of the ligand (shown in sticks in Figure 3) are certainly complementary. However, there are several other conformations of the ligand that can only be accommodated by a shift in the conformation of the loop. The results in Table 1 emphasize that the equal weighting scheme (averaging) is superior to one that favors only the lower energy conformations.

For neuraminidase, the conformational ensemble is less obvious but is dominated by a few residues including Arg 152, Ile 222, and Arg 371. Again, the lowest energy conformations of these flexible residues complement the ligand ensemble. However, the ensemble allows the ligand conformational freedom that would not be realized by docking to any single structure. The ensemble allows the ligand to effectively vibrate in the binding site and sample more conformational space than the crystallographic models allow (see Figure 3B). In crystallographic models, the ligand is relatively fixed (PDB ID: 1NNC B-factors <18) due to the 12 hydrogen bonds observed between the ligand and protein. These alternate ligand conformations maintain the extensive protein:ligand hydrogen-bond network, and results from Table 2 show that these conformations should be equally weighted.

DISCUSSION

There is now significant evidence showing that the use of an ensemble of target structures during docking improves enrichment and ligand ranking.[1,3,10,34−37] There are many approaches to determining the appropriate weight for the individual ensemble member scores. To compare Boltzmann weighting and averaging, we analyzed methods for scoring ensembles of docked complexes and the resulting impact on ligand ranking. Several key findings will enhance the pursuit of virtual lead optimization: (1) as expected, a Boltzmann weighting provides the most accurate result when considering a traditional ensemble, without structural minimization; however, averaging scores across a minimized ensemble is more accurate, (2) an empirical scoring function such as Surflex−Dock yields increased accuracy relative to force-field-based scoring functions, (3) a confirmation that ensembles yield more accurate ranking relative to a single member, and (4) the type of ensemble created and the weighting of that ensemble is critical. Averaging the scores across the minimized ensemble based on a single trajectory yields the best results, and increasing the size of the ensemble has little effect on ligand ranking accuracy.

For these new case studies, we again show that docking to an ensemble provides the most accurate ligand ranking for virtual lead optimization. We also show that averaging scores across all minimized ensemble members provides the most accurate ligand ranking. The underlying reason that averaging is more accurate relative to Boltzmann weighting is likely that minimized ensembles approximate a microcanonical ensemble rather than a canonical or traditional ensemble. Thus, more receptor states are contributing to the aggregate score in a microcanonical ensemble, and receptor flexibility is more appropriately simulated. For a canonical ensemble (constant number of particles, constant volume, and temperature), as the number of states approaches infinity, the population follows a Boltzmann distribution. The significant improvement observed in ligand ranking relative to averaging by applying the Boltzmann distribution function to an ensemble without minimization suggests that this ensemble is canonical. Each ensemble member of a canonical ensemble can be thought of as a microcanonical ensemble comprised of structures with energies within an infinitesimal interval, making them all equivalent and equally probable (equiprobability postulate).[38−43] The energies of ensemble members derived from a minimized ensemble are within a relatively narrow energy range, and thus are equally probable according to the equiprobability postulate.

Several factors discussed by Jain[30] may account for the better performance of empirical scoring functions such as Surflex−Dock. The empirical scoring function tolerates atom−atom interpenetrations, further emphasizing the importance of receptor flexibility to accurate docking. Also, the Surflex−Dock scoring function contains terms that account for solvation and the lost entropy of the ligand upon binding; these contributions are not considered in force-field scoring functions.

Examining the number of ensemble members contributing to different treatments (see Table 1) also shows that all treatments other than averaging significantly winnow the number of members contributing more than 1% to the score. For the functions using the lowest energy complex or highest Surflex−Dock scores, only one member contributes to the aggregate scores. These methods essentially reduce a rich data set and correspondingly reduce accuracy.

In this study, the virtual ensemble definition was restricted to just 21 states; however, in the physical realm the protein and ligands are not limited to a specific number of possible states. However, increasing the ensemble members to 42 did not improve accuracy regardless of the ensemble treatment. Using a signal-to-noise ratio where the signal is the number of measurements (ensemble members) and the noise is proportional to the square root of the number of ensemble members shows that an increase in ranking accuracy will always be proportional to the square root of the number of ensemble members. Nearly a 5-fold increase in accuracy can be calculated when 21 members are used versus just one. However, when 42 members are used, there is slightly more than a 6-fold increase in accuracy. Accuracy is not significantly increased beyond a certain number of ensemble members; this limit is practically reached when 21 equally contributing members are considered.

Throughout the course of this investigation, we have quantitatively evaluated the use of ensembles for accurate ligand ranking. The ensemble type, scoring function, and the method for weighting those scores significantly impact ligand ranking accuracy, with averaging empirical scores across a minimized ensemble providing the most accurate results. Structural evidence supports that the ensemble samples conformational space not accessible to individual members but vital to ligand ranking accuracy. Treating each member of the minimized ensemble as equally probable emulates a microcanonical ensemble, a well-established concept borrowed from statistical mechanics.

ACKNOWLEDGMENT

This work was supported by the NIH (GM067542 to A.C.A.).

Supporting Information Available: Tables of neuraminidase inhibitors, rmsd of 100 ps molecular dynamics simulation, and the binning definitions for CaDHFR and influenza A NA inhibitors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Popov, V. M.; Yee, W. A.; Anderson, A. C. Towards in silico lead optimization: scores from ensembles of protein/ligand conformations reliably correlate with biological activity. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 375–387.

(2) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Bockmann, R. A. Predicting free energy changes using structural ensembles. *Nat. Methods* **2009**, *6*, 3–4.

(3) Bolstad, E. S.; Anderson, A. C. In pursuit of virtual lead optimization: the role of the receptor structure and ensembles in accurate docking. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 566–580.

(4) Bolstad, E. S.; Anderson, A. C. In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 62–74.

(5) Fennen, J.; Torda, A. E.; van Gunsteren, W. F. Structure refinement with molecular dynamics and a Boltzmann-weighted ensemble. *J. Biomol. NMR* **1995**, *6*, 163–170.

(6) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424–440.

(7) Kim, J.; Park, J. G.; Chong, Y. FlexE ensemble docking approach to virtual screening for CDK2 inhibitors. *Mol. Simul.* **2007**, *33*, 667–676.

(8) Yoon, S.; Welsh, W. J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.

(9) Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J. Med. Chem.* **2008**, *51*, 7469–7477.

(10) Huang, S. Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399–421.

(11) Morris, G. M.; Lim-Wilby, M. Molecular docking. *Methods Mol. Biol.* **2008**, *443*, 365–382.

(12) Lilien, R. H.; Stevens, B. W.; Anderson, A. C.; Donald, B. R. A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *J. Comput. Biol.* **2005**, *12*, 740–761.

(13) Lerner, M. G.; Bowman, A. L.; Carlson, H. A. Incorporating dynamics in E. coli dihydrofolate reductase enhances structure-based drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 2358–2365.

(14) Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. Method for including the dynamic fluctuations of a protein in computer-aided drug design. *J. Phys. Chem. A* **1999**, *103*, 10213–10219.

(15) Amaro, R. E.; Cheng, X.; Ivanov, I.; Xu, D.; McCammon, J. A. Characterizing loop dynamics and ligand recognition in human- and avian-type influenza neuraminidases via generalized born molecular dynamics and end-point free energy calculations. *J. Am. Chem. Soc.* **2009**, *131*, 4702–4709.

(16) Whitlow, M.; Howard, A. J.; Stewart, D.; Hardman, K. D.; Kuyper, L. F.; Baccanari, D. P.; Fling, M. E.; Tansik, R. L. X-ray crystallographic studies of *Candida albicans* dihydrofolate reductase. High resolution structures of the holoenzyme and an inhibited ternary complex. *J. Biol. Chem.* **1997**, *272*, 30289–30298.

(17) *SYBYL*, version 8.0; Tripos Inc.: St. Louis, MO, 2007.

(18) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER, version 7*; University of California: San Francisco, CA, 2002.

(19) Varghese, J. N.; Epa, V. C.; Colman, P. M. Three-dimensional structure of the complex of 4-guanidino-Neu5Ac2en and influenza virus neuraminidase. *Protein Sci.* **1995**, *4*, 1081–1087.

(20) Liu, J.; Bolstad, D. B.; Smith, A. E.; Priestley, N. D.; Wright, D. L.; Anderson, A. C. Structure-guided development of efficacious antifungal agents targeting *Candida glabrata* dihydrofolate reductase. *Chem. Biol.* **2008**, *15*, 990–996.

(21) Liu, J.; Bolstad, D. B.; Smith, A. E.; Priestley, N. D.; Wright, D. L.; Anderson, A. C. Probing the active site of *Candida glabrata* dihydrofolate reductase with high resolution crystal structures and the synthesis of new inhibitors. *Chem. Biol. Drug Des.* **2009**, *73*, 62–74.

(22) Beierlein, J. M.; Frey, K. M.; Bolstad, D. B.; Pelphrey, P. M.; Joska, T. M.; Smith, A. E.; Priestley, N. D.; Wright, D. L.; Anderson, A. C. Synthetic and crystallographic studies of a new inhibitor series targeting *Bacillus anthracis* dihydrofolate reductase. *J. Med. Chem.* **2008**, *51*, 7532–7540.

(23) Bolstad, D. B.; Bolstad, E. S.; Frey, K. M.; Wright, D. L.; Anderson, A. C. Structure-based approach to the development of potent and selective inhibitors of dihydrofolate reductase from *Cryptosporidium*. *J. Med. Chem.* **2008**, *51*, 6839–6852.

(24) Pelphrey, P. M.; Popov, V. M.; Joska, T. M.; Beierlein, J. M.; Bolstad, E. S.; Fillingham, Y. A.; Wright, D. L.; Anderson, A. C. Highly efficient ligands for dihydrofolate reductase from *Cryptosporidium hominis* and *Toxoplasma gondii* inspired by structural analysis. *J. Med. Chem.* **2007**, *50*, 940–950.

(25) Paulsen, J. L.; Liu, J.; Bolstad, D. B.; Smith, A. E.; Priestley, N. D.; Wright, D. L.; Anderson, A. C. In vitro biological activity and structural analysis of 2,4-diamino-5-(2′-arylpropargyl)pyrimidine inhibitors of *Candida albicans*. *Bioorg. Med. Chem.* **2009**, *17*, 4866–4872.

(26) Chand, P.; Babu, Y. S.; Bantia, S.; Chu, N.; Cole, L. B.; Kotian, P. L.; Laver, W. G.; Montgomery, J. A.; Pathak, V. P.; Petty, S. L.; Shrout, D. P.; Walsh, D. A.; Walsh, G. M. Design and synthesis of benzoic acid derivatives as influenza neuraminidase inhibitors using structure-based drug design. *J. Med. Chem.* **1997**, *40*, 4030–4052.

(27) Taylor, N. R.; Cleasby, A.; Singh, O.; Skarzynski, T.; Wonacott, A. J.; Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Bethell, R.; Colman, P.; Varghese, J. Dihydropyrancarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *J. Med. Chem.* **1998**, *41*, 798–807.

(28) Smith, P. W.; Robinson, J. E.; Evans, D. N.; Sollis, S. L.; Howes, P. D.; Trivedi, N.; Bethell, R. C. Sialidase inhibitors related to zanamivir: synthesis and biological evaluation of 4H-pyran 6-ether and ketone. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 601–604.

(29) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(30) Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407–420.

(31) Jain, A. N. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427–440.

(32) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(33) David, L.; Nielsen, P. A.; Hedstrom, M.; Norden, B. Scope and limitation of ligand docking: Methods, scoring functions and protein targets. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 275–306.

(34) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–627.

(35) Park, M. S.; Dessal, A. L.; Smrcka, A. V.; Stern, H. A. Evaluating docking methods for prediction of binding affinities of small molecules to the G protein betagamma subunits. *J. Chem. Inf. Model.* **2009**, *49*, 437–443.

(36) Polgar, T.; Keseru, G. M. Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and beta-secretase. *J. Chem. Inf. Model.* **2006**, *46*, 1795–1805.

(37) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161–1182.

(38) van Lith-van Dis, J. H. Stir in Stillness: a study in the foundations of equilibrium statistical mechanics. Ph.D. Thesis, Universiteit Utrecht, Veenendaal, The Netherlands, 2001.

(39) Touchette, H. The large deviation approach to statistical mechanics. *Phys. Rep.* **2009**, *478*, 1–69.

(40) Yates, J. T., Jr.; Johnson, J. K. *Molecular Physical Chemistry for Engineers*; University Science Books: Sausalito, CA, 2007; pp 183−283.

(41) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; John Wiley & Sons, Ltd.: West Sussex, UK, 2004; pp 69−103.

(42) Pathria, R. K. *Statistical Mechanics*; Butterworth-Heinemann: Oxford, UK, 2001; pp 9−101.

(43) Phillies, G. D. J. *Elementary Lectures in Statistical Mechanics*; Springer-Verlag: New York, NY, 2000; pp 11−38.

CI9003078