

Automatic Selection of Near-Native Protein-Ligand Conformations using a Hierarchical Clustering and Volunteer Computing

Trlce Estrada
University of Delaware
Dept. of Computer & Inf.
Sciences
Newark, DE, 19716
estrada@udel.edu

Roger Armen
University of Michigan Ann
Arbor
Dept. of Chemistry
Ann Arbor, MI , 48109
armenrs@umich.edu

Michela Taufer
University of Delaware
Dept. of Computer & Inf.
Sciences
Newark, DE, 19716
taufer@udel.edu

ABSTRACT

Docking simulations are commonly used to understand drug binding and require the search of a large space of protein-ligand conformations. Cloud and volunteer computing enable computationally expensive docking simulations at a rate never seen before but at the same time require scientists to deal with larger datasets. When analysing these datasets, a common practice is to reduce the resulting number of candidates up to 10 to 100 conformations based on energy values and then leave the scientists with the tedious task of subjectively selecting a possible near-native ligand. Scientists normally perform this task manually by using visual tools. Not only the manual process still depends on inaccurate energy scoring but also can be highly error-prone.

The contributions of this paper are twofold: First, we address the problem of extensively searching large spaces of protein-ligand docking conformations, supported by the volunteer computing project Docking@Home (D@H). Second, we address the problem of accurately, and automatically, selecting near-native ligand conformations from the large number of D@H results by using a probabilistic hierarchical clustering based on ligand geometry. Our method holds up even when we test for a search that is not biased by starting from near-native ligand conformations and clearly outperforms energy-based scoring methods.

1. INTRODUCTION

The design of new pharmaceutical drugs relies on finding small molecules, called ligands, that dock into proteins and play an essential role in turning protein functions on or off. Studying protein-ligand interactions in the wet lab is extremely expensive and time demanding especially for high-throughput experimental structure determination by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Computer simulations are used to accelerate this

process and to reduce costs. The computational search for putative drugs (i.e., ligands that dock well in a protein) is a search under uncertainty in a very large space of potential docking conformations; this space is shaped by the protein, the ligand, the computational methods, and the degrees of freedom to be explored [11].

In order to explore a large space, scientists can rely on cutting edge distributed technologies, such as cloud and volunteer computing. These technologies can perform computationally expensive protein-ligand simulations at a rate never seen before. At the same time, this capability leads to larger simulation datasets, resulting in new challenges for scientists who have to analyze these data. In particular, in docking simulations this results in the analysis of very large set of ligand conformations docked in a protein. In addition to the size of the data, scientists have to deal with the challenge of selecting well docked ligands under uncertainty. Protein-ligand docking complexes are normally scored based on approximated energy values. Unfortunately, these energy estimations can be inaccurate; in other words minimum energy conformations do not always correspond to the correct near-native conformation. Thus, the selection of the correct near-native ligand conformation out of a large ensemble of conformations is a selection process under uncertainty.

When dealing with the analysis of large ligand datasets, a common practice is to reduce the number of candidates up to 10 to 100 conformations based on energy values and then leave the scientists with the tedious task of subjectively selecting a possible near-native ligand. Scientists normally perform this task manually by using visual tools such as VMD [10] or Chimera [5]. Not only the manual process still depends on inaccurate energy scoring but also can be highly error-prone. To the best of our knowledge, most advanced methods of handling this task are not fully automated and there is always a need for improved methodology and automation of this process.

The contributions of this paper are twofold: First, we address the problem of extensively searching large spaces of protein-ligand docking conformations, supported by the volunteer computing project Docking@Home (D@H). Second, we address the problem of accurately, and automatically, selecting near-native ligand conformations from the large number of D@H results. In this paper we:

- Use the volunteer computing D@H project to collect extensive simulation results with two different docking algorithms (each with different levels of accuracy for the solvent representation) and two different approaches to generate initial ligand conformations.
- Present a clustering methodology that enables an accurate and efficient analysis of the large dataset even in the presence of data uncertainty. Our method uses a probabilistic hierarchical clustering that efficiently organizes ligand structures in a variable number of sets based on their geometry.
- Use our method to identify the set with less uncertainty from the large data set collected with D@H and we select a single ligand structure that potentially better represents a near-native candidate conformation.
- Empirically prove that our method is insensitive to different proteins, docking algorithms, and starting conditions and in average it provides an accurate near-native solution in 85% of the cases considered in this work.

The rest of this paper is organized as follows: Section 2 presents how D@H explores the large space of ligand conformations. Section 3 introduces the problem of accurately selecting near-native conformations; Section 4 describes our probabilistic hierarchical clustering and how to use it to analyze large protein-ligand docking datasets; Section 5 presents our results along with a comparison of a more traditional method for selection of near-native conformations; Section 6 discusses related work and Section 7 concludes the paper

2. EXPLORING THE LARGE SPACE OF LIGAND CONFORMATIONS WITH D@H

Docking@Home (D@H) is a volunteer computing project that aims to build a distributed computational environment to assist scientists in understanding the atomic details of protein-ligand interactions and accurately choosing near-native ligand structures. D@H uses thousands of volunteered computers to simulate the behavior of small molecules (called ligands) when docking into a protein to control its functions. The D@H framework relies on the BOINC [1] (Berkeley Open Infrastructure for Network Computing) middleware to deal with the generation, distribution, and execution of jobs as well as the collection of results across the Internet. More in particular, Docking@Home distributes jobs consisting of a ligand and a protein to the volunteer machines (also called D@H clients). The docking simulation is performed on the D@H client, which at the end of the computation, returns the ligand conformation when docked into the protein. For ligand conformation we mean the three dimensional position of the ligand atoms and their binding. Currently D@H is supported by 12,000 volunteers and 30,000 hosts; D@H has resulted in the collection of over 2TBytes of data in six months and about 30,000 docking results per day. Collected results are stored in a repository and analyzed in a phase in which we select from million of candidates a very reduced set of near-native ligand conformations based on the likelihood of the docking algorithm convergence.

D@H models a protein-ligand complex as a composition of a flexible ligand and a rigid protein structure (on a three

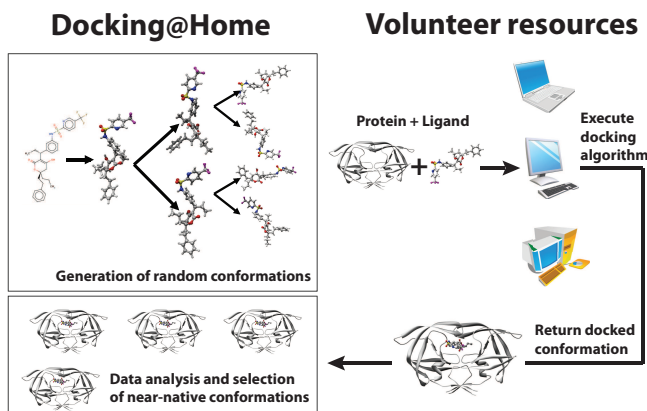


Figure 1: Docking@Home

dimensional lattice of regularly spaced points surrounding and centered on the active site of the protein, where each point on the grid stores the potential energy of a 'probe' atom's interaction with the molecule. A D@H job consists of a sequence of independent trials. For each trial, either a randomly generated conformation or a user defined conformation for a ligand is used as initial conformation. Random conformations are generated starting from the ligand crystal structure with random initial velocities on each ligand atom. Then the initial random conformation is randomly rotated to produce a set of different orientations that are placed into the active site of the protein. Once the ligand is docked into the protein site, an MD simulation consisting of a gradual heating phase of 4000 1-femtosecond (1-fs) steps from 300K to 700K, followed by a cooling phase of 10000 1-fs steps back to 300K, is performed. To facilitate the penetration of ligands into protein sites and allow larger conformational changes, van der Waals (vdW) and electrostatic potentials with soft-core repulsions are utilized. A soft-core repulsion reduces the potential barrier at vanishing interatomic distances to a finite limit allowing ligands to pass between conformational minima with a relatively small potential barrier that would normally be very large and impossible to overcome with an unmodified standard potential.

The solvent, usually water, in which the protein-ligand complexes reside, has a fundamental influence on protein and ligand interactions in any docking simulation. The solvent acts on the structures, screening the electrostatic interactions among atoms in the molecular structure. D@H uses two docking algorithms with different solvent representations:

- Method 1: a implicit representation of water using a distance-dependent dielectric coefficient (low if the atoms are close and progressively larger as the inter-atom distance increases)
- Method 2: a more physically accurate implicit representation of water using a Generalized Born model [12, 8].

The method based on the Generalized Born model is a more compute and memory intensive method. At the same time it provides a more physically accurate description of the poten-

tial energy of a ligand where part of the ligand conformation is exposed to solvent. In many situations where a large portion of the ligand is solvent exposed, the Generalized Born model should help significantly in providing better ligand conformations (e.g. when one orientation of a given ligand leaves a large bulky hydrophobic group exposed to solvent, this is penalized, where exposing a hydrophilic group like a hydroxyl OH group to solvent is much more favorable). Recent work assessing the accuracy of Method 1, and several implementations of Method 2 have demonstrated that the particular implementation used in this work has a much poorer performance for discriminating near-native geometry (Rahaman, Armen, Estrada, Taufer and Brooks unpublished data). This observation concurs with the poor accuracy of this method observed in this work, but newer properly tuned implementations of the Generalized Born method are able to outperform Method 1. However, in the current paper, even given the poor accuracy of Method 2, we demonstrate that our probabilistic hierarchical clustering is able to significantly improve the discrimination of near-native conformations, even given the inherent uncertainty of the scoring function for Method 2.

D@H targets three different proteins: trypsin, HIV, and p38-alpha. These proteins were selected because characterized by different degrees of flexibility during the docking process. *Trypsin* [7] is a relatively rigid protease that breaks down other proteins in the digestive system. Recent studies suggest that inhibitors of trypsin can have potential application in breast cancer treatment. It has been observed that trypsin-like proteases activate Protease-Activated Receptor-2 (PAR2), a protein in the tumor cell membrane. While activated, PAR2 causes the degradation of extracellular matrix (ECM) resulting into the spread of the tumor cell from one place to the other (metastasis). Drugs can act as inhibitors by de-activating the trypsin-like protease and are therefore potential agents capable of stopping the spread of breast cancer. *HIV protease (HIV PR)* [2] is a relatively flexible protein in the HIV virus that is essential for its replication in human cells. During the process of building a new HIV virus inside the human cell, HIV PR cleaves some newly synthesized viral protein in the post translational processing of the viral genome. The cleaved pieces are required to build a mature HIV virus. HIV PR is a well known therapeutic target for the treatment of HIV infection and preventing the development of AIDS in infected patients. A drug that can bind tightly to the active site of HIV PR will significantly inhibit the enzymatic activity of a large population of individual HIV PR molecules and significantly reduce the process of viral replication in infected cells. These drugs are called protease inhibitors. Several protease inhibitors like saquinavir, ritonavir, indinavir, and nelfinavir are available for the treatment of HIV infection. *p38-alpha* is the most flexible protein among the three proteins considered. P38alpha is also known as SAPK2a and MAPK14. It is involved in the regulation of cellular stress responses as well as the control of proliferation and survival of many cell types. Several promising compounds that inhibit p38 alpha are being investigated as potential therapies for arthritic and inflammatory diseases [18]. The protein-ligand complexes used in these docking studies are from the Ligand-Protein-Database (LPDB) [15].

3. SELECTING NEAR-NATIVE CONFORMATIONS UNDER UNCERTAINTY

Protein-ligand docking uses scoring functions for two separate tasks: the first step is the discrimination of ligand binding geometry (identification of near-native conformations), and the second step is a comparison of different ligands (different chemical species) to predict which ligands bind strongest to the protein. D@H is an engine for the first step, and the scoring for the second step can be performed as a post-processing step. This paper does not focus on the second step, but is entirely focused on the first step.

While dealing with the scoring, we initially relied on the traditional scoring approach based on energy values: we selected those ligands with lower energy as the more likely near-native conformations. We immediately identified the deficiencies of this approach in terms of accuracy. Figure 3 shows an example for 100,000 ligand conformations (every point in the figure is a ligand conformation) obtained with Docking@Home for the complex 1ajx. Here, the ligand conformations are scored in terms of their potential energy (x-axis) and their Root-Mean-Square-Deviation (RMSD) with respect to the known crystal structure (y-axis). The RMSD is measured in Angstroms (\AA) and is calculated by the root square of the average squared difference of all non-hydrogen ligand atoms in the simulated ligand conformation and the ligand atoms in the crystal structure. The figure shows three regions of relevance: (1) The area of conformations with minimum energy, which is the vertical rectangle that goes from -26 to -22 kcal/mol¹. A ligand conformation with minimum energy does not always has a near-native conformation. Conformations in this area would be selected by a method that only accounts for the energy and chances are that those candidates are not near-native conformations. In the figure we can see two local minima areas between 3-4 \AA and 8-9 \AA . (2) The area of conformations with minimum deviation (RMSD). The RMSD is calculated with respect to the crystal structure as explained above. This area is denoted by the horizontal rectangle that goes from 0 to 1 \AA . Ideally, the global minimum of a scoring function with high accuracy would be in this area. However, the global minimum is not always found (this is the case in Figure 3). For the discovery of new drugs, the deviation dimension (y-axis) is unknown and cannot be used to select candidate ligand conformations. In this paper we assume that this restriction always holds and we use the RMSD only for validation purposes. (3) The area of conformations with minimum energy and minimum deviation, which is the intersection of the other two areas described before. Ideally this area should be densely populated to increase the opportunity of selecting good candidate ligand conformations. As the figure shows, this may not happen, increasing the level of uncertainty and making harder the selection of near-native ligand candidates.

We observed the problem across docking results generated with the two different methods for the three proteins and the several ligands considered. Figure 3 shows an example of this phenomenon. 1w83 is the p38alpha kinase in complex with a small molecule inhibitor (ligand). Both subfigures, 3.a and 3.b, show the graphical comparison of the 1w83 ligand only

¹The range of minimum energy is complex dependent

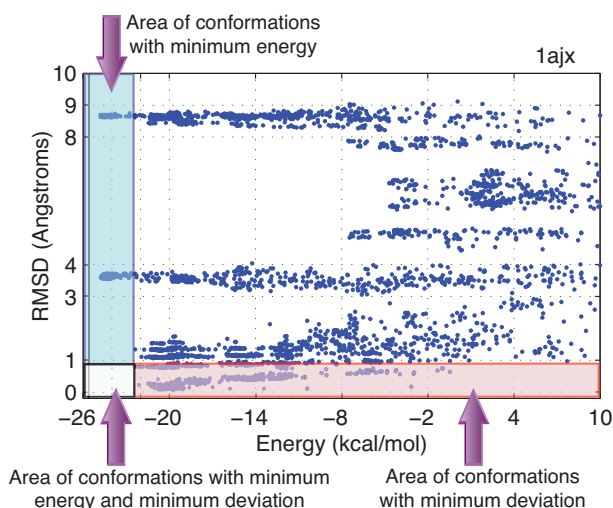


Figure 2: Selecting ligand conformations under uncertainty

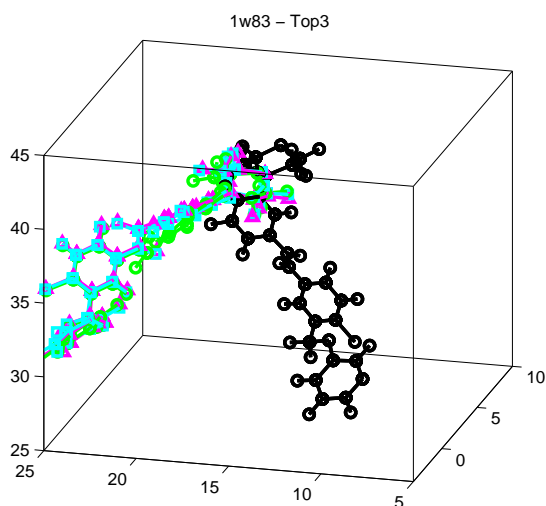
in the crystal structure (black ligand) versus the top 3 ligand conformations scoring the minimum energy over the whole set of D@H samples for this complex (grey ligands). Figure 3.a corresponds to the ligand conformations produced by Method 1. This is an extreme case where the scoring function assigns the lowest energy to a set of converging conformations that dock with a significantly different orientation than the crystal structure. Figure 3.b corresponds to the ligand conformations produced by Method 2 for the same complex. This figure shows that the minimum-energy scored structures do not converge to a single solution despite the large number of D@H samples. At the same time, these three results are substantially different among each other and none of them is accurate enough to be called a near-native conformation.

We exclude that the scoring uncertainty problem is related to an insufficient or inefficient sampling of the docking space: D@H is indeed capable to extensively sample the space under consideration. On the other hand, the modeling of the energies is still inaccurate even when using the method based on the Generalized Born implicit solvent model. Thus, not necessarily D@H near-native conformations score the lowest energy but, at the same time, they may have been numerously sampled with D@H and just need to be identified.

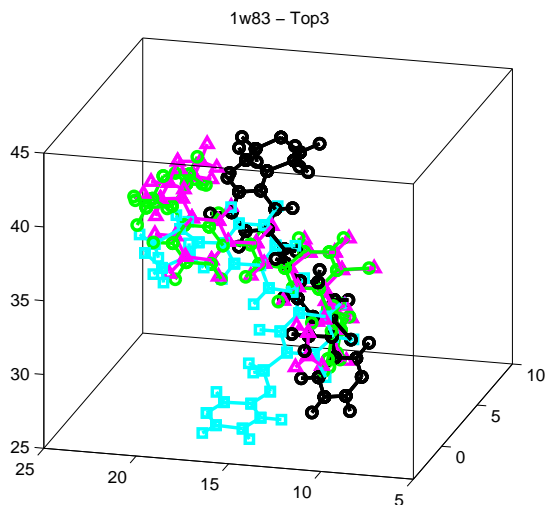
4. AN ALGORITHM FOR PROBABILISTIC HIERARCHICAL CLUSTERING

Results in Section 3 raise an important question. Given the inaccuracy of the docking algorithms and millions of collected conformations, how can the scientists select those that are more likely to occur in nature, considering that the energy is not always a reliable scoring metric?

Clustering algorithms can be used to successfully narrow down data in post-processing phases. They can also be used to identify results that are representative solutions in simulations. Clustering algorithms can be partitional or hierarchical. A partitional clustering divides the data according to a distance metric. A major limitation for this family of



(a) Method 1



(b) Method 2

Figure 3: Comparison of ligand structures selected by energy only for 1w83 - crystal structure (black ligand) vs. Top 3 scoring minimum energy (light-colored ligands)

algorithms, when used with data such as the protein-ligand conformations collected in D@H, is that the number of clusters has to be known a-priori. In D@H, this number is not known and an accurate and efficient estimate of this number is not possible neither when the simulation is in progress nor when the simulation is completed and all the results are collected. On the other hand, hierarchical clustering algorithms do not need previous knowledge on the final number of clusters and can be classified in agglomerative or divisive. A hierarchical agglomerative clustering starts with every element as an individual group and then merges similar elements into bigger clusters. A hierarchical divisive algorithm starts with one single cluster with all the data and then divides the data according to a specified distance metric. The

challenge for hierarchical clustering consists in determining the best clustering depth, i.e., knowing when to stop the merging or division of clusters. Another important requirement when using any clustering method is the ability to automatically organize data without having any previously classified, tagged, or annotated data and without explicit human intervention.

Clearly, none of the clustering techniques above serve the scope of efficiently and accurately clustering data sets such as the D@H results, when considered individually. Therefore, rather than using a single clustering technique, we propose to combine two techniques to benefit from their strengths. In particular, we propose to use a probabilistic hierarchical framework that combines (1) the capability to deal with data uncertainty by using a fuzzy c-means partitioning clustering with (2) the capability of identifying the number of needed clusters at runtime by using a divisive hierarchical algorithm for which the cluster hierarchy-depth is probabilistically determined based on result variability. Rather than using the energies, we use the geometrical conformations of the ligands as input to our clustering and the RMSD among the D@H resulting ligands as our distance metric. Note that here we refer to the RMSD as a metric to compare resulting ligands among them and we do not refer to the crystal structure that is unknown for us during the scoring process. We also assume that D@H provides us with the sufficient number of samples, and thus the docking simulations converge toward near-native solutions. Our resulting probabilistic hierarchical framework is able to perform an effective unsupervised clustering of the large D@H datasets even in the presence of uncertainty.

More in particular, the fuzzy c-means (FCM) [4] allows for a non-disjoint partitioning of the data providing a way to deal with uncertainty. In traditional clustering, every element belongs to just one cluster; on the contrary, in fuzzy clustering each element has a score, or degree of belonging, to each of the clusters. Elements belong to each cluster with different degrees depending on their distance to the center (also called centroid) of that cluster. More formally, for any dataset D and for any element in the dataset $d_i \in D$, there is a scoring vector \vec{s}_i giving the probability of d_i being an element of each cluster: $\vec{s}_i = s_{i,1}, s_{i,2}, \dots, s_{i,k}$, where k is the predefined number of clusters and the sum of \vec{s}_i is 1. FCM is an iterative clustering: it starts by selecting at random k elements (or initial centroids) from the dataset D . The second step is calculating the scoring vector \vec{s}_i for each element d_i , where the degree of belonging is a normalized inverse of the distance from the cluster centroid C_k as shown in Equation 1, where $distance(x, y)$ is a user-defined distance function and can be customized for the type of data being analyzed. For each cluster, a new centroid C'_k is calculated as the mean of all points, weighted by their degree of belonging to the cluster (as shown in Equation 2). This process iterates until the centroids stabilize, i.e., until there is no change in the centroids.

$$s_{i,k} = \frac{1}{\sum_j \left(\frac{distance(C_k, d_i)}{distance(C_j, d_i)} \right)^2} \quad (1)$$

$$C'_k = \frac{\sum_{i=1}^n s_{i,k}^2 d_i}{\sum_{i=1}^n s_{i,k}^2} \quad (2)$$

In our framework, FCM works in concert with a divisive hierarchical clustering algorithm. The divisive hierarchical clustering starts with the data set D_m ($m \geq 0$) and uses the FCM algorithm to divide the set into two subsets ($k=2$), one of which is defined as the complement of the other (D_{m+1} and $D_m - D_{m+1}$). Redundant elements (those that are not strongly biased to one cluster or the other) are temporarily removed from the two main partitions. Our probabilistic hierarchical framework selects the partition (D_{m+1}) with a probability directly proportional to its size and inversely proportional to its internal variance. This partition is used to further subdivide. The division process continues until the means of the two partitions (D_{m+1} and $D_m - D_{m+1}$) are equal to each other with a statistical significance of 0.05. To determine if the two means of both clusters are equal, we use the Welch's t-test [19] of D_{m+1} and $D_m - D_{m+1}$, as shown in Equation 3, where C_{m+1} is the centroid of D_{m+1} , $C_{m'+1}$ is the centroid of $D_m - D_{m+1}$ and $\sigma_{m+1}, \sigma_{m'+1}$ are their standard deviations respectively. Once we calculated the t-test we find its p-value from a Student's t-distribution [17]. We save the centroids and continue dividing the current partition until the p-value is less than 0.05 and the number of elements in D_{m+1} is larger than a threshold defined by the accuracy (e.g., 1Å). At every step, a hierarchy of centroids is kept and it is used to summarize the data space.

$$t = \frac{C_{m+1} - C_{m'+1}}{\sqrt{\frac{\sigma_{m+1}}{|D_{m+1}|} + \frac{\sigma_{m'+1}}{|D_m - D_{m+1}|}}} \quad (3)$$

Our probabilistic hierarchical framework can be used to (1) automatically organize data into disjoint sets, and (2) find a global, most likely single solution from multiple independent results. Figure 4 presents an example for which the hierarchical algorithm is graphically shown as a tree structure (dendrogram). For this particular figure, centroids for $D_0, D_0 - D_1, D_1, D_1 - D_2, D_2, D_2 - D_3$, and D_3 are saved and can be used to analyze and summarize the different dimensions of the dataset. Also, the last cluster is by definition the most compact one (i.e., the larger cluster with smaller internal variance). Thus, this last cluster (D_3) represents the most reliable consensus obtained from the data. Consequently, the centroid of D_3 can be used as the most likely solution of the whole data.

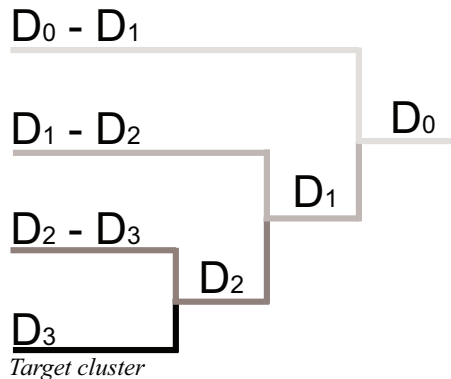


Figure 4: Hierarchical clustering represented as a dendrogram

5. MEASURING ACCURACY OF DOCKING PREDICTIONS

To test our probabilistic hierarchical framework, we ran docking trials for each of the 23 protein-ligand complexes for HIV protease (i.e., 1ajv, 1ajx, 1d4h, 1d4i, 1d4j, 1ebw, 1ebz, 1g2k, 1g35, 1gno, 1hbv, 1hih, 1hps, 1hpx, 1hsg, 1htf, 1hvi, 1hvj, 1hvk, 1liq, 1m0b, 1ohr, and 1t7k), 21 ligands docking into the Trypsin protein (i.e., 1c2d, 1c5p, 1c5q, 1ce5, 1g36, 1ghz, 1gi4, 1gi6, 1gj6, 1kli, 1klj, 1kl1, 1klm, 1k1n, 1ppc, 1pph, 1qb6, 1tpp, 1xug, 2bza, 3ptb), and 12 ligands docking into the P38alpha kinase (i.e., 1a9u, 1bl6, 1bl7, 1di9, 1kv1, 1kv2, 1ouk, 1ouy, 1oz1, 1w83, 1w84, 1yqj). For each of these complexes we ran 2 million trials.

In a first set of tests to evaluate if our probabilistic hierarchical clustering is robust and can capture near-native conformations independently from the docking method used, we considered both the two docking methods described in Section 3 (Method 1 and Method 2) and randomly generated ligands as initial conformations (see Figures 5 and 6). In a second set of tests, to assess whether the initial conformations used in the docking method play any active role in biasing the accuracy of our clustering-based selection, we used Method 1 and user-defined ligands whose conformations are $>5\text{\AA}$ from the correct crystal structure (see Figure 7). Note that a conformation with $>5\text{\AA}$ from the correspondent crystal structures is considered a bad docked conformation.

We used our probabilistic hierarchical clustering to find the most likely near-native ligand conformations for each complex. For each complex, the input to our framework was the set of 100,000 ligand conformations. The distance metric used to cluster each ligand was the RMSD of its atom coordinates versus all the other ligands already in the cluster. If a simulation converges, then the largest cluster with lower internal variance (denoted as a *target cluster*) is likely the cluster that contains more near-native conformations. In our experiments, the ligand conformation with highest degree of belonging (*centroid*) to the *target cluster* is selected as our predicted near-native conformation (see Equation 1). In the rest of this paper, we refer to this conformation as the clustering candidate of a given complex.

The entire process of clustering and selection of the clustering candidate was performed without using the crystal structures available for the complexes in LPDB [15]. The crystal structures played an important role only in the validation phase of our framework when, for each complex, we calculated the RMSD of the clustering candidate with respect to its crystal structure. We also consider 100 D@H conformations selected based on their lowest energy versus the same crystal structure, emulating in this way the naive approach of scientists when not supported by our framework. A conformation can be considered a near-native conformation if its RMSD is smaller than or equal to two \AA ; however, conformations with RMSD between two and three \AA are still considered results of interest. In the case of the naive approach we consider that we capture a near-native conformation if the arithmetic median is below or equal to two \AA . The use of the median is preferred as the accuracy metric over the mean because less affected by extreme values [9]. Figures 5.a, b, and c present the two validation

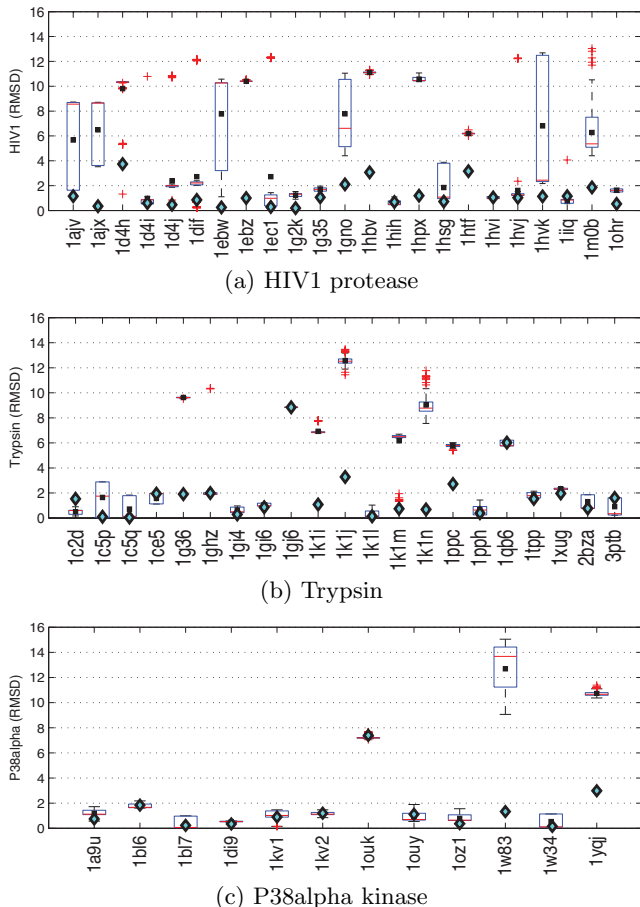
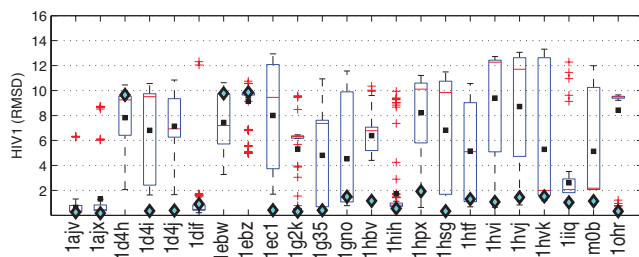


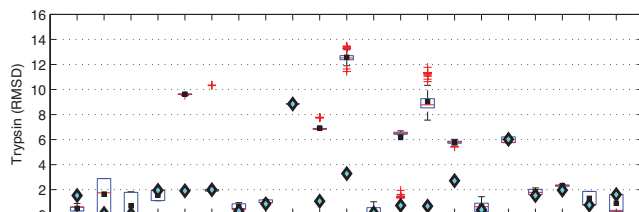
Figure 5: Docking Method 1. RMSD comparison of 100 ligand conformations selected based on minimum energy (box plot) vs. 1 ligand selected by hierarchical clustering (diamond)

comparisons for the three proteins of interest for Method 1. Figures 6.a, b, and c present the two comparisons for the same three proteins for Method 2, and Figures 7.a, b and c present the comparisons for the same three proteins for Method 1 starting from a user-defined conformation that was at least 5\AA away from the crystal structure. Figures 5.a, 6.a and Figures 7.a refer to the HIV1 protease, Figures 5.b, 6.b and 7.b refer to Trypsin, and Figures 5.c, 6.c and Figures 7.c refer to P38alpha kinase. On the x-axes we show the different complexes and on the y-axes we show their RMSD, the lower the better. Diamonds represent the RMSD of the clustering candidate w.r.t. the crystal structure. The box plot graphics represent the the RMSD of the 100 conformations selected based on energy. The box plot data graphics consist of seven different pieces of information. The whiskers on the bottom extend from the 10th percentile (bottom decile) to the top 90th percentile (top decile). Outliers are placed at the end of the top decile whiskers. The top, bottom, and line through the middle of the box correspond to the 75th percentile (top), 25th percentile (bottom), and 50th percentile (middle). A square indicates the arithmetic mean.

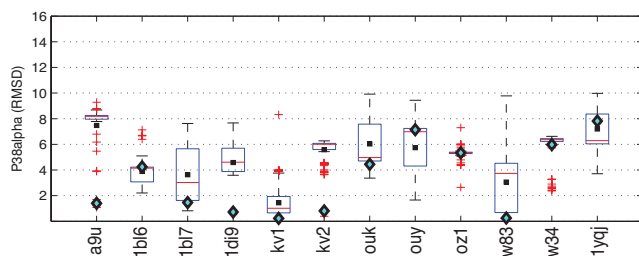
For Method 1 and the HIV1 protease, the naive approach



(a) HIV1 protease



(b) Trypsin



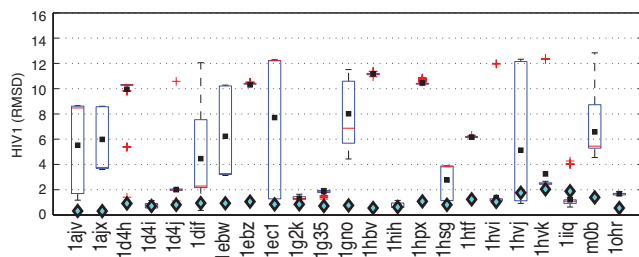
(c) P38alpha kinase

Figure 6: Docking Method 2. RMSD comparison of 100 ligand conformations selected based on minimum energy (box plot) vs. 1 ligand selected by hierarchical clustering (diamond)

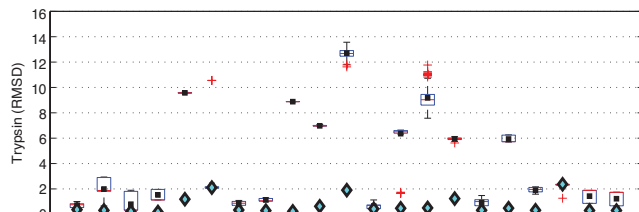
Table 1: Comparison of number of hits per docking method and protein

Docking Method	Protein	Min. Energy Selection	Clustering Selection
Method 1	HIV1	10 (43%)	19 (82%)
Method 2	HIV1	8 (34%)	20 (86%)
Method 1&2	HIV1	-	23 (100%)
Method 1	Trypsin	12 (57%)	17 (80%)
Method 2	Trypsin	11 (52%)	16 (76%)
Method 1&2	Trypsin	-	17 (80%)
Method 1	P38alpha	9 (75%)	10 (83%)
Method 2	P38alpha	1 (1%)	6 (50%)
Method 1&2	P38alpha	-	10 (83%)
Method 1	All	31 (55%)	46 (82%)
Method 2	All	20 (35%)	42 (75%)
Method 1&2	All	-	50 (89%)

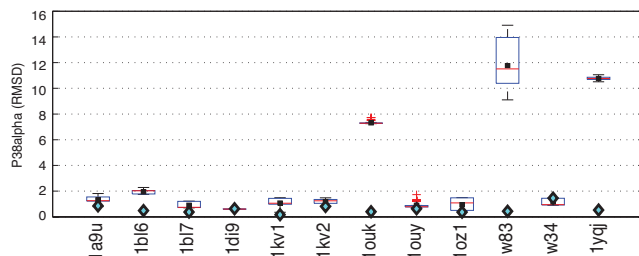
is able to identify only 10 of the 23 near-native conformations (hit rate of 43%) while our clustering method captures 19 of the 23 near-native conformations (hit rate of 82%).



(a) HIV1 protease



(b) Trypsin



(c) P38alpha kinase

Figure 7: Docking Method 1 using starting conformations >5Å from LPDB crystal structure. RMSD comparison of 100 ligand conformations selected based on minimum energy (box plot) vs. 1 ligand selected by hierarchical clustering (diamond)

For Trypsin, the naive approach identifies only 12 of the 21 near-native conformations (hit rate of 57%) and our clustering method captures 17 of the 21 near-native conformations (hit rate of 80%). For the P38alpha kinase, the naive approach identifies 9 of the 12 near-native conformations (hit rate of 75%) and our clustering method captures 10 of the 12 near-native conformation (hit rate of 83%). When we consider the more compute-intensive Method 2, for the HIV protease, the naive approach is able to identify only 8 of the 23 near-native conformations (hit rate of 34%) and our clustering method captures 20 of the 23 near-native conformations (hit rate of 86%). For Trypsin, the naive approach identifies only 11 of the 21 near-native conformations (hit rate of 52%) and our clustering method captures 16 of the 21 near-native conformations (hit rate of 76%). For the P38alpha kinase the naive approach identifies 1 of the 12 near-native conformations (hit rate of 0.8%) and our approach captures 6 of the 12 near-native conformation (hit rate of 50%). Table 1 summarizes the hit rates for the two docking methods. As shown in the table, overall our framework outperforms the naive approach for all the complexes and for each method. With our clustering method we can see that none of the two

docking methods clearly outperform the other. The combination of the two docking methods (Method 1 and Method 2) can further strengthen the accuracy of our predictions only for the HIV protease for which we observed a hit rate of 100%. For the other two proteins we observed the same hit rates. In the table, we do not compare the energy based selection when combining the two docking methods, since they are composed of quite different approximations of the potential energy.

Table 2 summarizes the hit rates for the docking Method 1 when the docking process starts from a user-defined starting conformation at least 5Å away from the conformation in the LPDB crystal structure. For the HIV1 protease, the naive approach based on energy is able to identify only 10 of the 23 near-native conformations (hit rate of 43%) while our clustering method captures all the 23 near-native conformations (hit rate of 100%). For the Trypsin protein, the naive approach based on energy is able to identify only 12 of the 21 near-native conformations (hit rate of 57%) while our clustering method captures 20 of the 21 near-native conformations (hit rate of 95%). For the P38alpha kinase, the naive approach identifies 9 of the 12 near-native conformations (hit rate of 75%) and our approach captures 12 of the 12 near-native conformation (hit rate of 100%). When considering the three sets of complexes together the energy-based selection found near-native conformations only 55% of the times while our clustering method identified near-native ligand conformation in 98% of the cases. As shown in Table 2, our framework still outperforms the approach based on energy for the two set of complexes and shows similar trends as for Method 1 when using random ligand starting conformations that are generated from MD starting from the crystallographic ligand conformation. Moreover the data in Table 2 demonstrates that the final conclusions of this study are robust and not biased from starting the conformational search from near-native initial ligand conformations.

Table 2: Comparison of number of hits for Method 1 starting from a user-defined conformation >5Å from the crystal structure

Docking Method	Protein	Min. Energy Selection	Clustering Selection
Method 1	HIV1	10 (43%)	23 (100%)
Method 1	Trypsin	12 (57%)	20 (95%)
Method 1	P38alpha	9 (75%)	12 (100%)
Method 1	All	31 (55%)	55 (98%)

If we reconsider the same complex (1w83) presented in Figure 3 and this time we use our clustering method for the selection of the candidate conformation, we observe that we are able to find a near-native conformation for both docking methods (Figure 8.a for Method 1 and Figure 8.b for Method 2). The black ligand, representing the ligand in the LPDB crystal structure, practically overlaps with the grey ligand, which in this case represents the candidate ligand conformation selected by our probabilistic hierarchical clustering. Contrary to the energy-based ligand selection, our probabilistic clustering is able to accurately identify the near-

native ligand conformation independently from the docking method used.

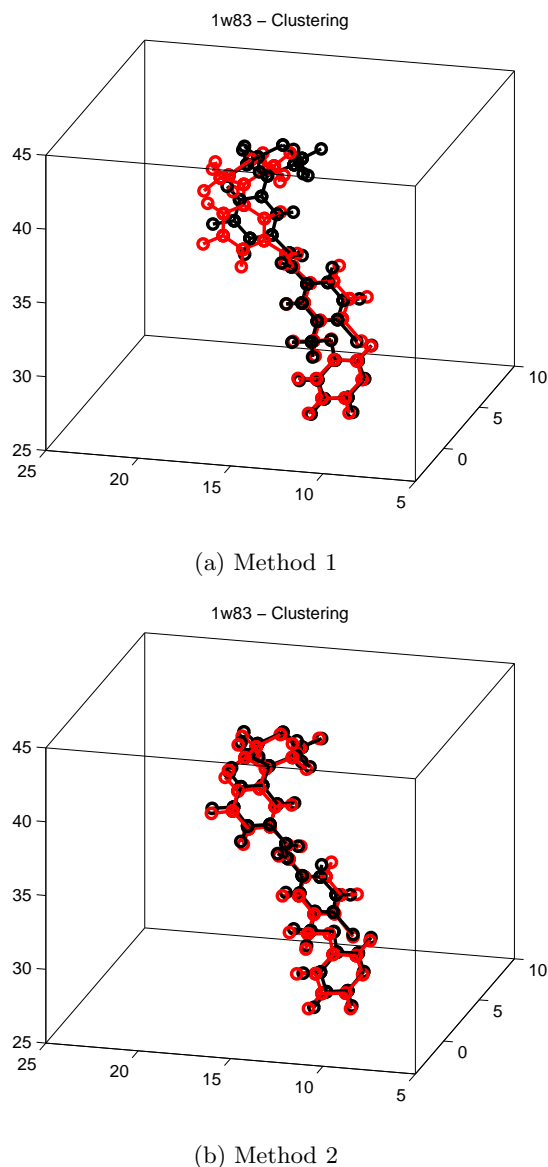


Figure 8: Comparison of ligand structures selected by hierarchical clustering for 1w83 - crystal structure (black) vs. clustering-selected conformation (lighter color)

To further illustrate the behavior of our clustering method and its capability to identify a variable number of clusters dynamically, Figure 9 shows the resulting clustering for nine of the complexes using Method 2 (three for the HIV, i.e., 1d4i, 1dif, and 1ebw; three for Trypsin, i.e., 1k1m, 1c2d, and 3ptb; and three for the P38alpha, i.e., 1a9u, 1oz1, and 1ouy). For each protein we present a complex for which our method clearly outperforms the naive approach (left column), a complex for which our method has similar accuracy as the naive approach (central column), and a complex for which the naive approach has better accuracy (right col-

umn). We use Method 2, since for Method 1 our framework is always either better or equal in accuracy than the naive approach (see Figure 5). This is not always the case for Method 2 (see Figure 6). After the clustering is completed, we map each conformation in each cluster to its energy (x-axes) and its RMSD with respect to the crystal structure (y-axes). Different colors show the different clusters (as denoted in the legend). The deepest cluster in the hierarchy is the *target cluster* containing the clustering candidate shown as an horizontal solid line. The best conformation selected based on its minimum energy is shown with a dashed horizontal line. As shown in Figure 9, the depth of the cluster hierarchy (number of clusters) is variable, ranging from two for e.g., 1dif, to four for e.g., 1d4i, and depends on energy landscape of the complex [6]. The maximum number of clusters found in the D@H datasets is six. No human intervention is required to define the clustering depth.

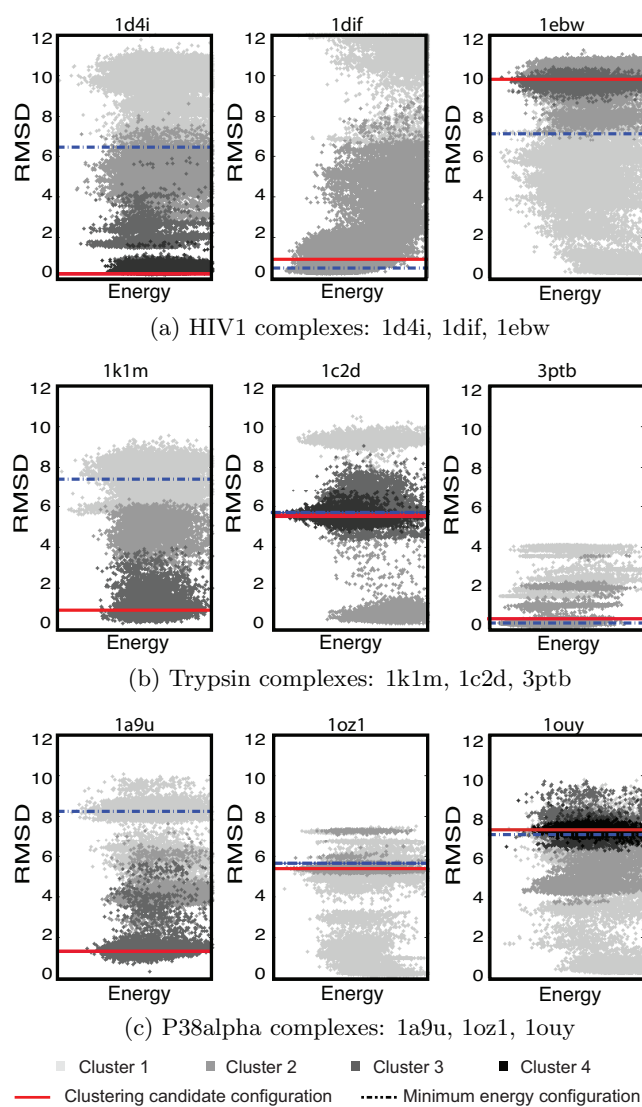


Figure 9: Examples of clusters with variable depth and variable accuracy found with our framework for a D@H dataset

Because of its high hit rates, the proposed framework can be used unmodified to select near-native ligand conformations in the protein-ligand docking simulations when crystal structures are unknown. Moreover, the framework can be used for the selection of near-native conformations in other relevant simulations such as protein folding and protein prediction.

6. RELATED WORK

Previous work successfully used different techniques to explore the search space of docking conformations. A common approach not based on clustering is the simulation sampling refinement. The refinement approach finds at runtime which conformations are likely to be near-native and then performs an extensive sampling around the predicted conformations. Important work in this direction includes Yang et. al. [16] and Liang et. al. [13].

Yang et. al. [16] underestimate a set of local energy minima and uses this model to drive further sampling. Yang’s algorithm explores the free energy surface spanned by encounter complexes and finds docking predictions with an accuracy within 5Å from the crystal structure. Liang et. al. [13] propose a method to refine the accuracy of predicted protein-ligand docking conformations based on several scoring functions and a computationally efficient algorithm for conformation refinement. By using this method, the authors were able to improve the accuracy of predicted conformations by 0.5Å compared to other methods. Both [16] and [13] improve the accuracy of docking methods and increase the probability of selecting near-native conformations but do not provide automatic selection of near native conformations.

Clustering methods have also been used to find a ‘reduced’ set of possibly near-native docking conformations. This group of conformations is manually analyzed by experts who decide which conformations are good candidates and which are not. To our knowledge none of these clustering methods are fully automatic and require some degree of human intervention. Clustering approaches include work of Lorenzen et. al. [14], Bouvier et. al. [3], and Chang et. al. [6].

Lorenzen et. al. [14] select near-native docking conformations based on a clustering approach considering that a bigger cluster is more likely to have better candidate conformations. As in our work, the selection based on cluster size outperforms the ranking based on the energy value. The clustering is driven by manually-defined thresholds and can find docking conformations with an accuracy of about 5Å. Bouvier et. al. [3] uses a Kohonen self-organizing map (SOM) that is trained in a preliminary phase using drug-protein contact descriptors. As in this paper, Bouvier’s work describes the possibility of overcoming the inherent problems of scoring functions by using a statistical analysis of different properties of the docked conformations. Chang et. al. [6] performed a simple cluster analysis of docking simulations and uses the size of the clusters to estimate the vibrational entropy of the resulting conformations. The conformation frequency provides information on the energy landscape of binding. A high frequency is a measure of favorable entropy in the binding process.

The methods based on clustering presented above use single

clustering algorithms to group docking conformations and are not fully automatic. When based on thresholds, as in the case of [14, 6], the methods require additional tuning to achieve optimal decisions; different tuning parameters produce different results depending on the set of complexes. When based on training datasets, as in the case of [3], the methods require extensive validation to prove their robustness. This extensive validation is missing in [3]. In contrast, our framework does not require tuning of parameters and can be applied unmodified to a new set of unseen complexes.

7. CONCLUSION

Cutting edge distributed technologies, such as cloud and volunteer computing, provide scientists with an efficient and scalable way to perform computationally-expensive docking simulations at a rate never seen before. In this paper we use Docking@Home, a volunteer computing project, to run this type of simulations. More in particular D@H uses 30,000 volunteered computers to simulate the behavior of small molecules (called ligands) when docking into a protein to control its functions.

Supported by the D@H capability we searched the large space of protein-ligand docking conformations for three major proteins and 56 ligands. When using only energy-based scoring methods, only in 35% (worst scenario) and 55% (best scenario) of the cases we were able to identify a near-native ligand conformation. We significantly improved this scoring accuracy by using our novel method that enables an automatic analysis of protein-ligand docking results even in the presence of data uncertainty. Our method is based on a probabilistic hierarchical clustering that efficiently organizes data in a variable number of sets based on their geometry. Each set has a single representative solution; thus, scientific conclusions can be achieved in a short turnaround time by analyzing only a reduced set of representative solutions. By using our method and a randomly generated ligand as our starting conformation, we were able to identify near-native ligand conformations in 75% (worst scenario) and 89% (best scenario) of the cases. When starting from an unbiased conformation such as a ligand whose conformation is $>5\text{\AA}$ from the correct crystal structure, our method still outperforms the energy-based selection: we identified near-native ligand conformations in 98% of the cases for the three proteins considered while the energy-based method found such a conformation in only 55% of the cases for the same datasets.

Acknowledgment

This work was supported by the NSF, grant #0941318 ‘CDI-Type I: Bridging the Gap Between Next-Generation High Performance Hybrid Computers and Physics Based Computational Models for Quantitative Description of Molecular Recognition’, and grant #0922657 ‘MRI: Acquisition of a Facility for Computational Approaches to Molecular-Scale Problems’, and by the U.S. Army, grant#ARO 54723-CS ‘Computer-Aided Design of Drugs on Emerging Hybrid High Performance Computers’, and by the CONACyT fellowship #171595. The authors thank Charles L. Brooks III for his valuable advice and the Docking@Home volunteers for providing us with essential resources.

8. REFERENCES

- [1] D. P. Anderson. Boinc: A system for public-resource computing and storage. In *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing.*, pages 4–10, November 2004.
- [2] K. Bckbro, S. Lwgren, K. Osterlund, J. Atepo, T. Unge, J. Hultn, N.M. Bonham, W. Schaal, A. Karl, and A. Hallberg. Unexpected binding mode of a cyclic sulfamide hiv-1 protease inhibitor. *Journal of Medical Chemistry*, 40:898–902, 1997.
- [3] G. Bouvier, N. Evrard-Todeschi, J. P. Girault, and G. Bertho. Automatic clustering of docking poses in virtual screening process using self-organising map. *Bioinformatics Advance Access*, 2009.
- [4] R. L. Cannon, J. V. Dave, and J. C. Bezdek. Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:248–255, 1986.
- [5] S. Ceri and R. Manthey. Chimera: A model and language for active dood systems. In *In Proceedings of the 2nd East-West Database Workshop, Workshops in Computing*, pages 3–16. Springer, 1994.
- [6] M. W. Chang, R. K. Belew, K. S. Carroll, A. J. Olson, and D. S. Goodsell. Empirical entropic contributions in computational docking: Evaluation in aps reductase complexes. *Journal of Computational Chemistry*, 29:1753–1761, 2008.
- [7] F. Dullweber, M.T. Stubbs, D. Musil, J. Strzebecher, and G. Klebe. Factorising ligand affinity: a combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *Journal of Molecular Biology*, 313:593–614, 2001.
- [8] M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks III. Performance comparison of generalized born and poisson methods in the calculation of electrostatic solvation energies for protein structures. *Journal of Computational Chemistry*, 25:265–84, 2004.
- [9] P. C. D. Hawkins, G. L. Warren, A. G. Skillman, and A. Nicholls. How to do an evaluation: pitfalls and traps. *J. of Computer Aided Molecular Design*, 22:179–190, 2008.
- [10] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
- [11] A. N. Jain. Bias, reporting, and sharing: computational evaluations of docking methods. *Journal of Computer Aided Molecular Design*, 22:201–212, 2008.
- [12] M. S. Lee, M. Feig, F. R. Salsbury Jr., and C. L. Brooks III. New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *Journal of Computational Chemistry*, 24:1348–56, 2003.
- [13] S. Liang, G. Wang, and Y. Zhou. Refining near-native protein-protein docking decoys by local resampling and energy minimization. *PROTEINS: Structure, Function, and Bioinformatics*, pages 309–316, 2008.
- [14] S. Lorenzen and Y. Zhang. Identification of near-native structures by clustering protein docking conformations. *PROTEINS: Structure, Function, and Bioinformatics*, 68:187–194, 2007.
- [15] LPDB - protein-ligand database. <http://lpdb.scripps.edu/>.
- [16] Y. Shen, I. C. Paschalidis, P. Vakili, and S. Vajda. Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLOS Computational Biology*, 4, 2008.
- [17] W. S. Student Gosset. The probable error of a mean. *Biometrika*, 6:1–25, 1908.
- [18] Z. Wang, B.J. Canagarajah, J.C. Boehm, S. Kassis, M.H. Cobb, P.R. Young, S. Abdel-Meguid, J.L. Adams, and E.J. Goldsmith. Structural basis of inhibitor selectivity in map kinases. *Structure*, 6:1117–28, 1998.
- [19] B. L. Welch. The generalization of ‘student’s’ problem. *Biometrika*, 34:28–35, 1947.