Informatics-driven Protein-protein Docking

Irina Hashmi
Department of Computer Science,
George Mason University,
Fairfax, VA 22030
ihashmi@gmu.edu

Amarda Shehu*
Department of Computer Science,
Department of Bioengineering
School of Systems Biology
George Mason University,
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

Predicting the structure of protein assemblies is fundamental to our ability to understand the molecular basis of biological function. The basic protein-protein docking problem involving two protein units docking onto each-other remains challenging. One direction of research is exploring probabilistic search algorithms with high exploration capability, but these algorithms are limited by errors in current energy functions. A complementary direction is choosing to understand what constitutes true interaction interfaces. In this paper we present a method that combines the two directions and advances research into computationally-efficient yet high-accuracy docking. We present an informatics-driven probabilistic search algorithm for rigid protein-protein docking. The algorithm builds upon the powerful basin hopping framework, which we have shown in many settings in molecular modeling to have high exploration capability. Rather than operate de novo, the algorithm employs information on what constitutes a native interaction interface. A predictive machine learning model is built and trained a priori on known dimeric structures to learn features correlated with a true interface. The model is fast, accurate, and replaces expensive physics-based energy functions in scoring sampled configurations. A sophisticated energy function is used to refine only high-scoring configurations. The result is an ensemble of high-quality decoy configurations that we show here to approach the known native dimeric structure better than other state-of-the-art docking methods. We believe the proposed method advances computationally-efficient high-accuracy docking.

Keywords

rigid protein-protein docking; informatics-driven; probabilistic search; basin hopping; putative interaction interface; machine learning; decoy ensemble.

*Corresponding Author

ACM BCBW 2013 Washington, D.C.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; J.3 [**Computer Appplications**]: Life and Medical Sciences

General Terms

Algorithms

1. INTRODUCTION

It is said that structure is a carrier of function. This is particularly true of proteins, the ubiquitous workhorses found in almost every chemical pathway in the living cell. Proteins use their structures to dock onto one another and form supramolecular assemblies. Characterizing the three-dimensional structure of a supramolecular assembly is central to understanding the molecular basis of biological function in the healthy and diseased cell [50].

Computationally, the problem involves computing the structures of the protein units after docking and their spatial arrangement relative to one another in the assembly [30, 39]. This problem is particularly challenging due to a rather large search space of many dimensions. In the general setting, the configuration search space has as many as $N \times M + 6$ dimensions. Here, 6 parameters are reserved to specify the spatial arrangement (translation and rotation in 3D) of a selected moving unit onto the other base unit after docking, and N parameters are reserved to specify the internal structure of each of the M units. It is therefore not surprising that the problem is intractable [40].

Two key simplifications are made that are additionally found to describe most protein assemblies. The number of units is limited to two, capturing a large percentage of assemblies [14]. The resulting problem is referred to as protein-protein docking. Additionally, the structures of the participating units are considered to either not change during docking (thus, rigid) or undergo local fluctuations that can be modeled by a post-processing step [50]. The resulting problem is known as rigid protein-protein docking and is the subject of our investigation here.

Even rigid protein-protein docking remains a challenging problem. Though the search space is now only 6-dimensional, it is continuous. The different spatial arrangements that superimpose a region of the molecular surface of one protein unit onto that of another are beyond enumeration. Though the space can be discretized, accuracy and detail are sacrificed by discretization [17]. Probabilistic search algorithms are in principle capable of dealing with vast search spaces [56].

However, to be effective, these algorithms need to be guided towards relevant configurations. Predominantly, this is carried out by integrating an energy function that scores the contact interface of a (dimeric) configuration on essentially how close it is to the true interaction interface [17, 59]. The group of methods that implement this approach can be referred to as energy-driven, and advances by these methods are attributed to design of more sophisticated energy functions, design of more powerful probabilistic search algorithms with high exploration capability, or both.

Many methods are now available for protein-protein docking, such as Haddock [10], Zdock [6], ClusPro [8], PatchDock and SymmDock [11], SKE-DOCK [57], Combdock [24, 25], RosettaDock [38], GRAMM-X [58], and more [53]. Many of them are energy-based. Others delay consideration of a computationally demanding energy function, choosing instead to sample rigid-body transformations (representing spatial arrangements) that superimpose geometrically-complementary regions on the units' molecular surfaces. These methods are commonly referred to as geometry-based, but their accuracy is typically lower than that of energy-based methods if no energetic refinement is carried out afterwards.

Currently, no single method is sufficient to successfully predict native assemblies in every test case [40]. Often, only 30-58% of the correct interface is predicted in any given target [32]. In particular, the use of sophisticated energy functions while desirable, presents two challenges. First, such functions are computationally demanding and dominate CPU cycles. Second, all known functions contain errors and predominantly lead search algorithms towards non-native configurations [17, 31]. This is a common issue with molecular docking, observed in other settings, including de novo tertiary structure prediction [55, 56]. Though a useful energy function can correct and improve a structure, it alone cannot guide the search to find the low-energy near-native configurations. Some studies do show the importance of electrostatics, van der Waals, and desolvation potentials [7, 13, 28, 29] for scoring docked configurations. Others, such as Haddock, employ additional information about the actual native configuration obtained from NMR [10].

Given the current impasse, a complementary direction of research in structural characterization of protein assemblies is focusing on understanding what constitutes true or native interaction interfaces [9, 27]. Many methods currently exist that analyze known native structures of protein assemblies for summarizing interaction interfaces with few features [41]. For instance, native interfaces are generally found to contain residues of high evolutionary conservation [12, 35]. Docking methods have exploited knowledge of interaction interfaces in guiding the search towards native-like configurations [18, 26, 33, 34, 36, 62]. Recent work by us has extended geometry-based methods to directly sample rigid-body transformations that match geometrically-complementary and evolutionary-conserved surface regions [19, 20, 22]. However, evolutionary conservation is probably one of many features characterizing interaction interfaces. Research in finding such features is active [5, 27, 37].

In this paper we propose a method idDock, $\widehat{\text{i}}$nformatics-$\widehat{\text{d}}$riven $\widehat{\text{D}}$ocking, that integrates a machine learning model with a probabilistic search algorithm, effectively combining the two research directions summarized above. The search algorithm in idDock explores the bound configuration space using a fast supervised learning model rather than a computationally-demanding physics-based energy function to expediently score sampled configurations. The model is an entropy-based decision tree that is a priori trained on 138 known protein-protein interactions. Essentially, the model employs specified features about a contact surface to determine whether the surface is native-like or not. We extract information from the model to associate an integer score with a novel, unlabeled contact interface in a bound configuration sampled by the probabilistic search algorithm.

The search algorithm builds upon the basin hopping framework, which we have shown in diverse protein modeling settings to have high exploration capability [22, 43–46, 48]. The algorithm generates bound configurations consecutively, effectively conducting a biased random walk in the bound configuration space. Given a sampled configuration, the contact interface is perturbed to obtain a novel configuration. The configuration is discarded if the learned predictive model associates a high score with it (in keeping with how energy functions are used to score configurations, a low score is interpreted to indicate a more promising interface). Otherwise, a short energetic refinement is conducted to improve the quality of the configuration. A sophisticated physics-based energy function, FoldX [54], is employed for the refinement. A Metropolis criterion is then applied to determine whether the resulting configuration should be added to the growing trajectory.

The result of this process is an ensemble of consecutively-obtained high-quality decoy configurations that we show in this paper to approach the known native structure better than other state-of-the-art docking methods. Our comparison includes methods representative of various energy- or geometry-driven approaches, and methods that combine the two with evolutionary conservation information. We believe the proposed method and the integration of a predictive machine learning model in a probabilistic search algorithm opens the way towards computationally-efficient yet high-accuracy docking.

2. METHODS

The idDock method, detailed now in this section, wraps a predictive machine learning model within a probabilistic search algorithm. At a high-level, the search algorithm explores the bound configuration space by effectively sampling rigid-body transformations that superimpose geometrically-complementary regions on the molecular surfaces of the two units in a dimer. The algorithm itself is based on the basin hopping framework, hopping between consecutive minima of a sophisticated energy function. However, the energy function is not employed until a sampled configuration is deemed relevant by the predictive machine learning model. This presents an appealing approach to navigate a detailed energy surface while making good use of computational resources. Details now follow.

2.1 Probabilistic Search of Bound Configuration Space

The proposed search algorithm samples from the space of rigid-body transformations. One of the two units is considered the base (reference) unit. The other one is subjected to rigid-body transformations and docked onto the base unit. The algorithm starts with a transformation sampled at random, then conducts a biased random walk generating configurations consecutively, effectively modifying a current con-

figuration to obtain a new one. The algorithm addresses two important issues on 1) how to sample rigid-body transformations, and 2) how to modify a current configuration to obtain a new one using such transformations.

2.1.1 Rigid-body Transformations: Matching Geometrically-complementary Surface Regions

The molecular surfaces of both units are analyzed and represented through a collection of critical points. These points bear additional information on curvature and allow defining triangles (sets of three points) categorized as convex, concave, or saddle. Details can be found in related work on geometric hashing and geometry-driven docking [42, 61]. Two geometrically-complementary triangles are sampled, one from each unit, allowing to define a rigid-body transformation superimposing the triangle of the moving unit onto that of the base unit. This process can be repeated to sample rigid-body transformations superimposing geometrically-complementary regions/triangles on the units' molecular surfaces.

2.1.2 Generating Consecutive Bound Configurations

The above process is used to obtain an initial bound configuration C_0 for the search algorithm, which then proceeds to generate a series of configurations $\{C_1, C_2, \ldots\}$. A given, current configuration C_i in the trajectory is perturbed to obtain a new intermediate configuration $C_{i,\text{perturb}}$, which is then subjected to a short energetic minimization or refinement to obtain a candidate configuration for C_{i+1}.

This iterative application of perturbation followed by minimization is known as the basin hopping (BH) framework shown by us in various settings in protein modeling to have high exploration capability as long as the perturbation preserves adjacency between C_i and $C_{i,\text{perturb}}$, and the minimization does not consume computational resources [22, 43–46]. Both criteria are observed here, building upon a previous BH-based algorithm by us for protein-protein docking matching geometrically-complementary and regions that are evolutionary-conserved while making use of a simple physics-based energy function to rank a contact interface [22] (we note that our adaptation of BH here excludes evolutionary conservation).

A neighborhood of radius of dÅ is searched around each of the triangles superimposed to obtain C_i in order to find a new pair of geometrically-complementary triangles for superimposition. This effectively perturbs the contact interface in C_i to obtain $C_{i,\text{perturb}}$. The magnitude of d determines how different C_i and $C_{i,\text{perturb}}$ are. Our analysis in prior BH-based work in [22] indicates that $d = 5$Å is effective at preserving some adjacency, which has been shown to save BH from degenerating into a random restart algorithm and give it higher exploration capability [22, 43].

The traditional next step in BH is to project $C_{i,\text{perturb}}$ to a nearby minimum through an energetic minimization. In our prior investigation on BH for docking, we elected to make use of a simple physics-based energy function to save CPU cycles for the global search of the algorithm rather than the local refinement of each configuration. While the general approach was shown to be comparable with other docking methods, results were mixed [22]. Our investigation pointed to the need for a more sophisticated energy function. However, the computational demands of such a function present a practical issue. For this purpose, in this paper we elect to apply a sophisticated energy function only on perturbed configurations that are deemed promising to contain the native interface. This is accomplished by wrapping a predictive machine learning model in the algorithm.

The construction of the model is detailed below, but the search algorithm effectively uses it to label and then rank a perturbed configuration. The key idea is that once $C_{i,\text{perturb}}$ is obtained, it is not immediately subjected to minimization. Instead, the predictive model is used to make a prediction on whether the contact interface in the configuration is native or not. If the contact interface is predicted to have a label of 0 (thus, non-native) by the model, $C_{i,\text{perturb}}$ is discarded, and a new perturbed configuration is generated as above from C_i. If the prediction by the model is a label of 1 (thus, native), an actual score $1-5$ is associated with its contact interface (as detailed below). The configuration is discarded if the score is above some threshold (in keeping with how energy functions are typically used to score configurations, a low score indicates a more promising interface than a high score). Only a retained configuration is subjected to minimization.

We employ a short minimization protocol to lower interaction energy as measured through the FoldX force field [54]. The minimization proceeds for a fixed number of steps, at each step slightly modifying the rigid-body transformation in $C_{i,\text{perturb}}$ to lower its interaction energy as measured by FoldX (details on it are available in our prior investigation on BH in [22]). The energy terms in FoldX are weighted using experimental data obtained from protein engineering experiments. The terms include solvation energy, van der Waals, hydrogen bond potential, electrostatic energy, entropic and clash penalty. Details are available in [54].

FoldX is a sophisticated force field [15], but its computational demands are too high to be applied for refinement on each sampled configuration. Force fields like FoldX are most appropriate for employment when the search is in the vicinity of a native-like configuration. This is the reason idDock employs FoldX only to refine configurations deemed native-like by the predictive model. The result is that idDock is still able to hop between minima of a detailed physics-based energy function but in a computationally-efficient manner, as guided by the informatics model.

The result of this entire process is that given a configuration C_i representative of an energy minimum in the FoldX energy surface, a new configuration representative of another nearby minimum is obtained. This configuration is not automatically accepted to become C_{i+1} and advance the trajectory. Instead, a Metropolis criterion is used to guide the trajectory towards lower-energy minima. Essentially, a probability of acceptance $e^{-\delta E * \beta}$, is associated with the energetic change $\delta E = E(C_{i+1}) - E(C_i)$, where the scaling parameter β, set to 0.3, allows an energy increase of 2kcal/mol with probability 0.5. If the Metropolis criterion is met, the new minimum is added to the trajectory, thus advancing the search in the bound configuration space.

2.2 Machine Learning Model to Rank a Contact Interface

The predictive model wrapped inside the search algorithm is an entropy-based decision tree (DT), whose construction and training is detailed below. In summary, the model converts a contact interface into a feature vector which is then associated a binary 0/1 label. A label of 1 means the contact

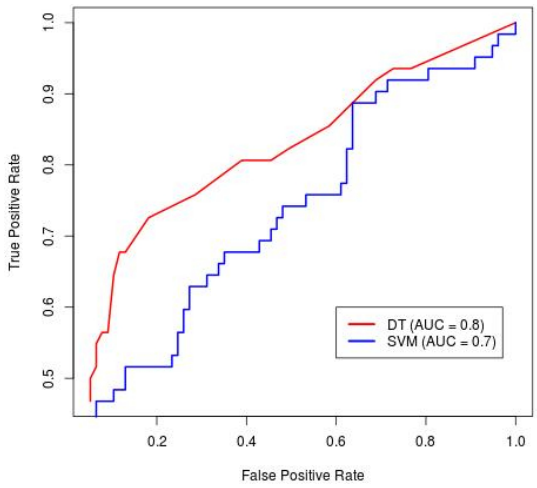

Figure 1: ROC comparison of DT to SVM.

interface is likely to be native. If not, the configuration is discarded, as described above. A 1-labeled configuration is further associated a $1-5$ score to restrict the minimization to more promising configurations. The score is constructed so that a low score means the interface is more likely to be native. The score is the depth of the decision tree at which a label of 1 is assigned to the interface. A lower score reflects the fact that the contact interface is found to be native-like quickly, with a fewer number of interface properties.

The model employed here is a J48 DT built through the WEKA package [16]. Our analysis compares this model to other standard classifiers (including support vector machines – SVMs) in a 10-fold validation classification setting on training data. The comparison includes various performance metrics, such as precision, recall, F-measure, and Receiver Operating Characteristic (ROC) curves. We summarize this comparison in Figure1 in terms of ROC curves, which plot the fraction of true positives versus the fraction of false positives as one continuously varies the decision threshold at which an instance is assigned a positive or negative label. An area under the ROC curve (AUC) of 1 represents perfect prediction; an area of 0.5 represents a random cointossing result. As shown in Figure 1, the AUC for DT is 0.8, whereas that for SVM is 0.7. Given the comparable performance and the additional property that DT is computationally faster at labeling an unlabeled contact interface and more intuitive at designing a score to rank a novel interface, we choose DT as the model by which to score perturbed configurations sampled by our probabilistic search algorithm in idDock. We now provide details on the actual DT construction.

From Contact Interface to a Feature Vector.

A contact interface is converted into a 7-dimensional vector. The first entry measures interface area calculated as in [62]: $\text{InterfaceArea}_{u_1+u_2} = 0.5 \cdot (\text{SASA}_{u1} + \text{SASA}_{u2} - \text{SASA}_{u1+u2})$, where SASA is the solvent accessible surface area measured through NACCESS [23], with u_i referring to unit i prior to docking and $u_1 + u_2$ referring to bound configuration. This formula is based on work in [62]. The second entry, also based on work in [62], is $\frac{\text{InterfaceArea}_{u_1+u_2}}{min(\text{SASA}_{u1}, \text{SASA}_{u2})}$. Entries $3-6$ measure compositions of amino-acid types in contact interface of a configuration. Four types are considered, such as hydrophobic, hydrophilic, acidic, and basic. The 7^{th} entry measures evolutionary conservation score of the contact interface by summing conservation scores obtained through iJET [12] over amino acids in the contact interface (iJET score of each amino acid ranges from 0.0 – least conserved to 1.0 – most conserved).

Given a set of bound configurations, their contact interfaces are determined as follows. A residue is said to be on the interface if its SASA decreases by $> 1\text{Å}^2$ upon complex formation (this definition is as in [62]). So-defined contact interfaces are converted to a set of 7-dimensional real-valued vectors as described above. Then, any supervised learning model can be trained in a classification setting on labeled vectors (those deemed positive/native or negative/non-native). As described above, our analysis suggests a DT model has high classification accuracy. The DT model is trained on the following dataset.

Training Dataset Construction.

The positive dataset consists of 62 true/native interaction interfaces found on experimentally-obtained assemblies extracted from a refined PDBbind dataset [60]. The negative datasets of 76 instances combines three sets non-native interfaces. The first is constructed by randomizing the positive dataset. Units selected at random from different complexes are docked with a random rigid-body transformation. This is repeated until 25 dimers are obtained. The second set consists of 47 crystal packing structures provided in [62]. The third consists of 4 dimeric structures just $5-12\text{Å}$ away in RMSD from native structures, generated from pyDOCK [7].

3. RESULTS

idDock was run on a 2.66GHz of Opteron Processor with 4GB of memory. Eleven dimers with known native structures have been selected in this preliminary investigation, chosen to vary in size, functional classification, and used as test cases by other docking methods. These systems are detailed in Table 1 (size indicates total number of atoms). It is worth noting that these systems are not included in our training data. Hence, they constitute a true testing dataset. On each system, idDock is run until $10,000$ dimeric configurations are obtained. We present three sets of results below. First, we quantify the extent to which the tree-based score captures *nativeness* of an interface. Second, we measure the quality of the idDock-obtained ensemble in terms of lowest RMSD from the known native structure and compare that to values reported by other docking methods. Third, we analyze in greater detail three selected systems on which we expose the energy surface probed by the method.

3.1 Analysis of Model Score And Nativeness

This analysis is highlighted on three selected dimers from our testing dataset on which idDock generates $10,000$ configurations. On each system, configurations within 5Å in RMSD from the native dimeric structure (RMSD is measured after alignment) are analyzed in terms of the distribution of tree-based scores. Restricting the focus to this subset of configurations allows us to essentially analyze native-like configurations. These three dimers are selected, because they are representatives of the general trends observed on the distribution of scores for other systems in our testing dataset. One would expect that th configurations that are

Table 1: Systems in our testing dataset.

ID	1C1Y	1WWW	1FLT	1IKN	1IKN	1VCB	1VCB	1OHZ	1QAV	1G4Y	1cse
Chains	A,B	W,Y	V,Y	A,C	C,D	A,B	B,C	A,B	A,B	B,R	E,I
Size	2034	1644	528	3178	2505	1447	1846	1443	1503	1838	2442

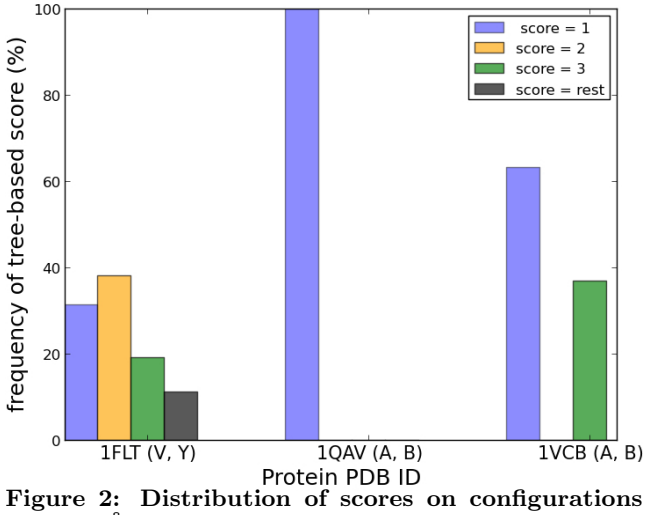

Figure 2: Distribution of scores on configurations within 5Å from known native dimeric structure obtained through idDock on three selected dimers. Distributions are shown in different colors (1 in blue, 2 in orange, 3 in green, and rest in black).

within 5Å in RMSD from the native dimeric structure would tend to have a lower rather than a higher score (we recall that lower is better in this setting). A bar graph in Figure 2 shows the percentage of these configurations with score $i \in \{1, \ldots, 5\}$ over the 5 possible score values. As expected, a higher percentage of native-like configurations have a lower score (more are in the bin corresponding to score 1), which makes the case that the tree-based score captures well the nativeness of a contact interface. This result suggests that the informatics-driven method proposed in this paper will mostly focus on the correct interaction interface then on other regions of the units' molecular surfaces.

3.2 Comparative Analysis

We now compare idDock to other docking methods in terms of lowest RMSD from the known native dimeric structure. The methods we have chosen for comparison represent different geometry- and energy-driven approaches for docking and our recent investigation of BH for docking [21, 22]. We recall that the work in [22] employed a simple physics-based energy function and focused on rigid-body transformations that superimposed geometrically- and evolutionary-conserved regions. It is worth pointing out that the comparison with our previous BH work in [22] allows directly testing the contribution of integrating the predictive learned model in enhancing the sampling capability of idDock and thus the quality of obtained dimeric configurations. BUDDA [49] is chosen as a representative of geometry-driven methods based on geometric hashing. pyDock [7] and ClusPro [8] represent highly-optimized energy-based protocols. In this context, our previous BH work in [22] represents a hybrid method.

The comparative analysis is shown in Table 2. The lowest RMSD reported by BUDDA [49] is shown in column 2,

Table 2: Comparison of idDock to other methods.

	Lowest RMSD to Native (Å)				
PDB ID	BUDDA [49]	pyDock [7]	ClusPro [8]	BH [22]	idDock
1C1Y	1.2	10.4	7.2	1.8	**2.7**
1WWW	11.4	18.2	17.2	2.6	**0.9**
1FLT	1.5	2.8	4.7	2.7	**0.6**
1IKN	1.2	20.1	19.7	2.1	**1.5**
1IKN	2.0	16.7	20.9	4.1	**2.5**
1VCB	0.7	1.4	1.9	3.4	**0.9**
1VCB	1.3	22.7	1.9	2.7	**1.4**
1OHZ	1.8	7.5	3.3	2.7	**0.7**
1QAV	1.4	9.6	1.7	2.6	**1.7**
1G4Y	0.8	26.2	1.9	4.1	**2.3**
1CSE	0.7	13.2	1.1	2.4	**1.2**

that by pyDOCK [7] in column 3, that by ClusPro [8] in column 4, our earlier BH work [22] is reported in column 5, and the lowest RMSD obtained by idDock is reported in column 6. The lowest RMSD obtained by idDock is colored in red and highlighted in bold if it is no higher than 2Å of the lowest value among all methods, and the entire row is colored in gray if it is indeed the lowest. This allows seeing that idDock is not only comparable to these state-of-the-art methods in all systems, but it outperforms these methods on $25-30\%$ of the systems. This is a promising result that motivates us to further investigate idDock and seek large-scale benchmarking.

3.3 Analysis of Dimeric Energy Surfaces Probed by idDock

We now take a closer look at the FoldX energy surface probed by idDock. In an ideal force field, there is strong positive correlation between energy values and RMSDs from the native structure; that is, lowering energy brings the search closer to the native structure. This allows offering the lowest-energy configuration as the native dimeric structure in a blind prediction setting. We show that, while the above comparative analysis makes the case that idDock approaches the native structure within a few angstroms, the energy surfaces are not always funnel-like. This is expected, as no energy function is ideal and error-free. We illustrate this on three systems, which have been selected to show negative correlation, no correlation, and positive correlation.

Figure 3 plots FoldX energy against the RMSD from the native structure of each decoy configuration generated by idDock. The dimer with native PDB id 1WWW has been selected to demonstrate negative correlation between RMSD from the native structure and FoldX energy (as energy goes down, RMSD increases). The native structure has energy of -6.67kcal/mol, and many other idDock-generated configurations have lower energy. In such a case, distortions in the energy function drive the search away from the native structure, if energy is to be used to make a prediction. The dimer with PDB id 1QAV has been selected to show no correlation and demonstrate that idDock cannot approach the energy

level of the native structure (the lowest reached is −10 as opposed to the −26.6kcal/mol of the native).

The dimer with native PDB id 1FLT represents the ideal case of positive correlation (the energy surface displays funneling), and the energy level of the native structure has been reached at −13.3kcal/mol. This analysis is useful, as it allows determining the exploration capability of the method, the extent to which the energy function guides to the correct interface, and the extent to which further refinement may be useful at driving towards the native structure.

4. CONCLUSION

We have presented idDock, a novel method for high-accuracy and computationally-efficient rigid protein-protein docking. The method integrates a machine learning model with a probabilistic search algorithm to effectively implement an informatics-driven approach. The model is based on supervised learning, identifying features found to correlate with true, native interaction interfaces in known dimeric structures. The learned model is used in a predictive setting by the proposed probabilistic search algorithm to score a sampled dimeric configuration. High-scoring configurations are further subjected to energetic refinement through a sophisticated energy function. The probabilistic search algorithm builds upon the basin hopping framework shown to have high exploration capability in high-dimensional search spaces with rugged energy surfaces.

The integration of a predictive learned model to score a sampled configuration is important, as the model is fast yet accurate, thus replacing physics-based energy functions known to be computationally demanding while prone to errors. Such functions are most appropriate for employment when the search is in the vicinity of a native-like configuration. This is the reason idDock employs FoldX, a sophisticated physics-based energy function, only to refine configurations deemed native-like by the predictive model.

Our results show that proposed method performs as well as or better than other state-of-the-art methods representative of different approaches to rigid protein-protein docking. Our ongoing work is focusing on obtaining more information about the predictive capability of the method in a larger setting of dimeric assemblies covering diverse functional classes. Future work will additionally consider enhancing the sampling capability of the search algorithm with evolutionary search techniques investigated and tested by us in the related problem of tertiary structure prediction in proteins [47, 48, 51, 52]. The model score may also be combined with terms of a physics-based energy function to conduct the search on a combined potentially less rugged surface.

While the presented work is on rigid protein-protein docking, structural flexibility can be modeled on decoy configurations with various existing tools [1–4]. The decoy ensembles can also be valuable to computational chemists in the design of more accurate physics-based or hybrid energy functions combining physics- and knowledge-based terms.

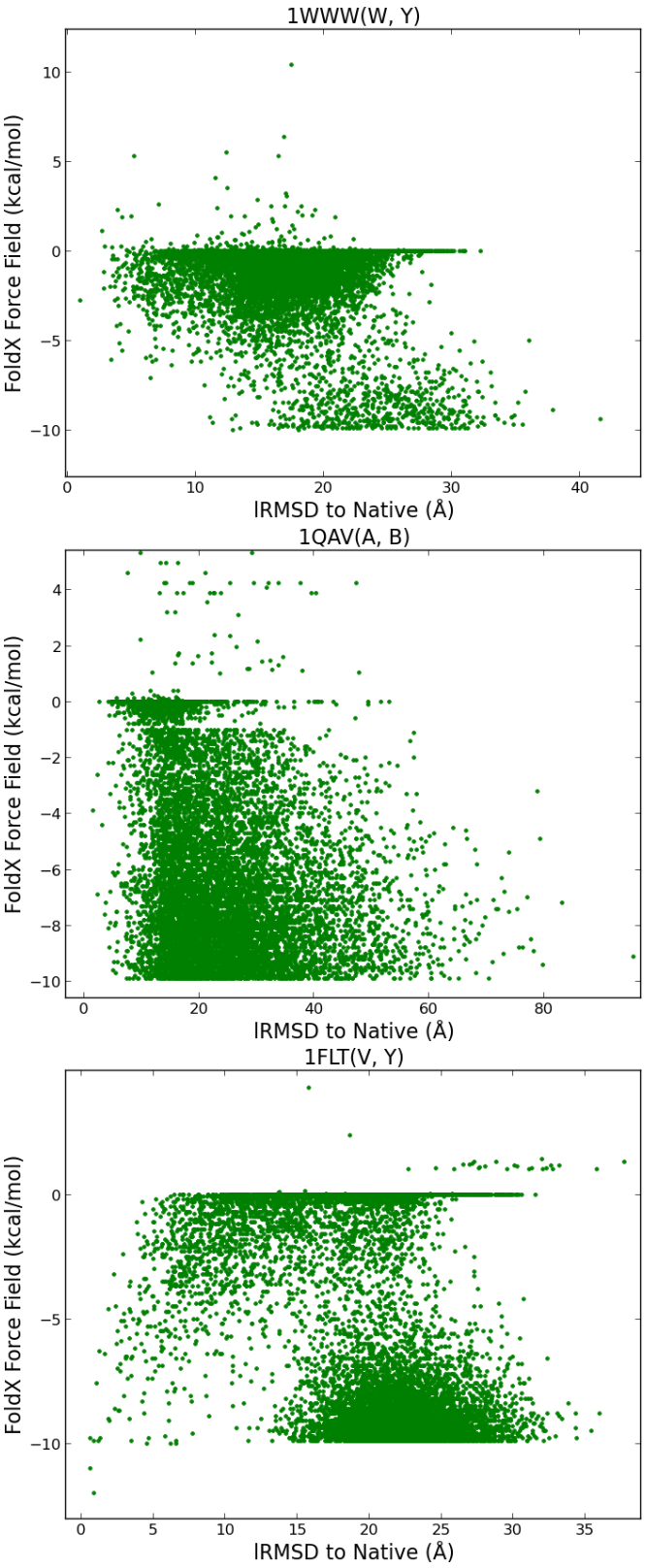

Figure 3: FoldX energy is plotted against RMSD from native structure for each idDock-generated configuration on three selected systems with native PDB ids 1WWW, 1QAV, and 1FLT, respectively.

Acknowledgment

This work is supported in part by NSF Award No. 1144106. We thank Dr. H. Rangwala for feedback on a class project initiating classification models for interfaces.

5. REFERENCES

[1] B. Akbal-Delibas, I. Hashmi, A. Shehu, and N. Haspel. Refinement of docked protein complex structures using evolutionary traces. In *Intl Conf on Biomed and Bioinf Workshops (BIBMW)*, pages 400–404. IEEE, November 2011.

[2] B. Akbal-Delibas, I. Hashmi, A. Shehu, and N. Haspel. An evolutionary conservation based method for refining and reranking protein complex structures. *J of Bioinf and Comp Biol*, 10(3):1242008, 2012.

[3] N. Akbal-Delibas and N. Haspel. Refining multimeric protein complexes using conservation, electrostatics and probabilistic selection. In *Intl Conf on Bioinf and Biomed Workshops (BIBMW)*, pages 102–108. IEEE, October 2012.

[4] N. Andrusier, R. Nussinov, and H. J. Wolfson. Firedock: fast interaction refinement in molecular docking. *Proteins: Struct. Funct. Bioinf.*, 69(1):139–159, 2007.

[5] N. J. Burgoyne and R. M. Jackson. Predicting protein interaction sites: binding hot-spots in proteinâĂŞprotein and proteinâĂŞligand interfaces. *Bioinformatics*, 22(11):1355–1342, 2006.

[6] R. Chen, L. Li, and Z. Weng. ZDock: an initial-stage protein-docking algorithm. *Proteins: Struct. Funct. Bioinf.*, 52(1):80–87, 2003.

[7] T. M. Cheng, T. L. Blundell, and J. Fernandez-Recio. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, 68(2):503–515, 2007.

[8] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho. ClusPro: a fully automated algorithm for protein-protein docking. *Nucl. Acids Res.*, 32(S1), 2004.

[9] W. L. DeLano. 12. *Curr. Opinion Struct. Biol.*, Unraveling hot spots in binding interfaces: progress and challenges.:14–20, 2002.

[10] C. Dominguez, R. Boelens, and A. Bonvin. Haddock: A protein-protein docking approach based on biochemical orbiophysical information. *J. Am. Chem. Soc.*, 125:1731–1737, 2003.

[11] D. Duhovny-Schneidman, Y. Inbar, R. Nussinov, and H. J. Wolfson. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucl. Acids Res.*, 33(S2):W363–W367, 2005.

[12] S. Engelen, A. T. Ladislas, S. Sacquin-More, R. Lavery, and A. Carbone. A large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comp Bio*, 5(1):e1000267, 2009.

[13] H. A. Gabb, R. M. Jackson, and M. J. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272(1):106–120, 1997.

[14] D. S. Goodsell and A. J. Olson. Structural symmetry and protein function. *Annu. Rev. Biophys. and Biomolec. Struct.*, 29:105–153, 2000.

[15] R. Guerois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320(2):369–387, 2002.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. In *SIGKDD*, volume 11, pages 10–18, New York, NY, USA, Nov. 2009. ACM.

[17] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, 2002.

[18] I. Halperin, H. Wolfson, and R. Nussinov. Protein-protein interactions: Coupling of structurally conserved residues and of hot spots across interfaces. implications for docking. *Structure*, 12(6):1027–1038, 2004.

[19] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu. Protein docking with information on evolutionary conserved interfaces. In *Intl Conf on Biomed and Bioinf Workshops (BIBMW)*, pages 358–365. IEEE, November 2011.

[20] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu. Guiding protein docking with geometric and evolutionary information. *J Bioinf and Comp Biol*, 10(3):1242002, 2012.

[21] I. Hashmi and A. Shehu. Hopdock: A probabilistic search algorithm for decoy sampling in protein-protein docking. *Proteome Sci*.

[22] I. Hashmi and A. Shehu. A basin hopping algorithm for protein-protein docking. In *IEEE Intl Conf on Bioinf and Biomed (BIBM)*, pages 466–469, October 2012.

[23] S. J. Hubbard, S. F. Campbell, and J. M. Thornton. Molecular recognition. conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.*, 220(2):507–530, 1991.

[24] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson. Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *J. Phys. Biol.*, 2:S156–S165, 2005.

[25] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson. Prediction of multimolecular assemblies by multiple docking. *J. Mol. Biol.*, 349(2):435–447, 2005.

[26] E. Kanamori, Y. Murakami, Y. Tsuchiya, D. Standley, H. Nakamura, and K. Kinoshita. Docking of protein molecular surfaces with evolutionary trace analysis. *proteinssfb*, 69:832–838, 2007.

[27] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, 345(5):1281–1294, 2005.

[28] J. G. Kirkwood. The forces between protein molecules in solution: A summary. *Cellular and Comparative Physiology*, 49(S1):59–62, 1957.

[29] A. T. Laurie and R. M. Jackson. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.

[30] M. F. Lensink, R. Mendez, and S. J. Wodak. Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins: Struct. Funct. Bioinf.*, 69(4):704–718, 2007.

[31] M. F. Lensink and S. J. Wodak. Docking and scoring protein interactions: CAPRI 2009. *Proteins: Struct. Funct. Bioinf.*, 78(15):3073–3084, 2009.

[32] M. F. Lensink and S. J. Wodak. Blind predictions of protein interfaces by docking calculations in CAPRI.

Proteins: Struct. Funct. Bioinf., 78(15):3085–3095, 2010.

[33] B. Li and D. Kihara. Protein docking prediction using predicted protein-protein interface. *BMC Bioinf*, 13:7, 2012.

[34] N. Li, Z. Sun, and F. Jiang. Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinf*, 9:553, 2008.

[35] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–58, 1996.

[36] Q. Liu and L. J. Propensity vectors of low-asa residue pairs in the distinction of protein interactions. *Proteins*, 78(3):589–602, 2009.

[37] Q. Liu and J. Li. Protein binding hot spots and the residue-residue pairing preference: a water exclusion perspective. *BMC Bioinf*, 11:244, 2010.

[38] S. Lyskov and J. J. Gray. The RosettaDock server for local protein-protein docking. *Nucl. Acids Res.*, 36(S2):W233–W238, 2008.

[39] R. Mendez, R. Leplae, M. F. Lensink, and S. J. Wodak. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Struct. Funct. Bioinf.*, 60(2), 2005.

[40] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British J Pharmacology*, 153(S1):S7–S27, 2009.

[41] I. S. Moreira, P. A. Fernandes, and M. J. Ramos. Hot spots-a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–812, 2007.

[42] R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. *Proteins*, 36(3):307–317, 1999.

[43] B. Olson, I. Hashmi, I. Molloy, and A. Shehu. Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules. *Advances in AI J*, 2012(674832), 2012.

[44] B. Olson and A. Shehu. Populating local minima in the protein conformational space. In *IEEE Intl Conf on Bioinf and Biomed*, pages 114–117, Atlanta, GA, November 2011.

[45] B. Olson and A. Shehu. Efficient basin hopping in the protein energy surface. In *IEEE Intl Conf on Bioinf and Biomed (BIBM)*, pages 119–124, Philadelphia, PA, October 2012.

[46] B. Olson and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Sci*, 10(Suppl 1):S5, 2012.

[47] B. Olson and A. Shehu. Multi-objective stochastic search for sampling local minima in the protein energy surface. In *ACM Conf on Bioinf and Comp Biol (BCB)*, Washington, D. C., September 2013.

[48] B. Olson and A. Shehu. Rapid sampling of local minima in protein energy surface and effective reduction through a multi-objective filter. *Proteome Sci*, 2013. in press.

[49] V. Polak. Budda: backbone unbound docking application master's thesis school of computer science, tel-aviv university. Master's thesis, Computer Science, Tel-Aviv University, Tel-Aviv,Israel, 2003.

[50] D. W. Ritchie. Recent progress and future directions in protein-protein docking. *Curr Protein and Peptide Sci*, 9(1), 2008.

[51] S. Saleh, B. Olson, and A. Shehu. A population-based evolutionary algorithm for sampling minima in the protein energy surface. In *Comput Struct Biol Workshop*, pages 48–55, Philadelphia, PA, October 2012.

[52] S. Saleh, B. Olson, and A. Shehu. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. *BMC Struct Biol*, 2013. in press.

[53] D. Schneidman-Duchovny, Y. Inbar, R. Nussinov, and H. J. Wolfson. Geometry based flexible and symmetric protein docking. *Proteins: Struct. Funct. Bioinf.*, 60(2):224–231, 2005.

[54] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The foldx web server: an online force field. *Nucl. Acids Res.*, 33(Web server issue):W382–W388, 2005.

[55] A. Shehu. Conformational search for the protein native state. In H. Rangwala and G. Karypis, editors, *Protein Structure Prediction: Method and Algorithms*, chapter 21. Wiley Book Series on Bioinformatics, Fairfax, VA, 2010.

[56] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In S. Aluru and A. Singh, editors, *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series, 2013.

[57] G. Terashi, M. Takeda-Shitaka, K. Kanou, M. Iwadate, D. Takaya, and H. Umeyama. The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins: Struct. Funct. Bioinf.*, 69(4):866–887, 2007.

[58] A. Tovchigrechko and I. A. Vakser. GRAMM-X public web server for protein-protein docking. *Nucl. Acids Res.*, 34(Web Server issue):W310–4, 2006.

[59] S. Vajda and D. Kozakov. Convergence and combination of methods in protein-protein docking. *Curr. Opinion Struct. Biol.*, 19:164–170, 2009.

[60] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang. The pdbbind database: Methodologies and updates. *J. Med. Chem.*, 48(12):411–419, 2005.

[61] H. L. Wolfson and I. Rigoutsos. Geometric hashing: an overview. *IEEE Comp Sci and Engineering*, 4(4):10–21, 1997.

[62] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinf*, 7:27, 2006.