Refinement of Docked Protein Complex Structures Using Evolutionary Traces

Bahar Akbal-Delibas*, Irina Hashmi†, Amarda Shehu†, and Nurit Haspel*
* *Department of Computer Science,*
University of Massachusetts Boston, Boston, MA, 02125
abakbal@cs.umb.edu, nurit.haspel@umb.edu
† *Department of Computer Science,*
George Mason University, Fairfax, VA 22030
ihashmi@gmu.edu, amarda@gmu.edu

***Abstract*—Detection of protein complexes and their structures is crucial for understanding the role of protein complexes in the basic biology of organisms. Computational methods can provide researchers with a good starting point for the analysis of protein complexes. However, computational docking methods are often not accurate and their results need to be further refined to improve interface packing. In this paper, we introduce a novel refinement method that incorporates evolutionary information by employing an energy function containing Evolutionary Trace (ET)-based scoring function, which also takes shape complementarity, electrostatic and Van der Waals interactions into account. We tested our method on docked candidates of three protein complexes produced by a separate docking method. Our results suggest that the energy function can help biasing the results towards complexes with native interactions, filtering out false results. Our refinement method is able to produce structures with better RMSDs with respect to the known complexes and lower energies than those initial docked structures.**

***Keywords*-protein docking; energy refinement; evolutionary trace analysis; evolutionary-conserved amino acids; effective distance restraints**

I. Introduction

Protein complexes play a central role in cellular organization and function, ion transport and regulation, signal transduction, protein degradation, and transcriptional regulation [1]. Since the three dimensional structure and the functionality of proteins are closely related to each other, detection of protein complexes and their structures is crucial for understanding the role of protein complexes in the basic biology of organisms.

Predicting the structure of a complex formed by assembling multiple chains is a difficult problem to solve in wet labs. Computational methods can become very useful where experimental methods fall short and provide researchers a good starting point for the analysis of protein complexes. Computational docking methods use structural and geometric search techniques and physico-chemical filters to model complex binding and rank computed structures according to energetic criteria using scoring functions. These scoring functions typically focus on electrostatic, Van der Waals, and hydrostatic interactions, similarity to experimental structures, or agreement with other experimental data [2]–[6]. The results generated by such computational methods are expected to be low-energy structures that are similar to the native complex structures.

Unfortunately, computational docking methods are not complete: low-energy structures often disagree with NMR data [7]. Recent CAPRI (Critical Assessment of PRedicted Interactions) rounds show an important observation [2]: even the most accurate methods predict only about 50% of the targets. Therefore, the results of computational docking methods need to be further refined in order to obtain native-like structures. Usage of refinement methods on protein complexes is not limited to computational docking methods; structures obtained by experimental methods can also be refined.

In this paper we present a novel docking refinement method that combines shape, physico-chemical and evolutionary information to better discriminate native-like from decoy structures for the protein-protein docking results and improve interface packing. The novelty of our work comes from its ability to combine two different existing concepts called *evolutionary conservation scores* and *effective distance restraints*, which are explained in detail in section II, and incorporate them into the energy function. The main idea is that proteins tend to preserve their functionally-important amino acids, which play a part in interacting with its partner proteins, throughout the evolution [8], [9], and functionally-important amino acids of different interacting chains are expected to be close to each other on the interface. Recent methods using sequence conservation through evolutionary traces (ET) [9]–[11] allow detecting binding interfaces in silico. Our method makes use of this information by employing an energy function that incorporates an ET-based scoring function to detect evolutionarily conserved amino acids, and drives the search towards conformations which have those functionally-important amino acids positioned close to each other on each chain. The energy function iteratively detects top-scoring transformations at each stage of the refinement process to improve interface packing. As shown, our results are more similar to the native complex than the initial results. The method explained below can run on complexes containing any number of chains and not just

978-1-4577-1613-3/11/$26.00 ©2011 IEEE

dimers.

II. Methods

A. Overview

The input to our program is a complex structure generated by a docking method. Figure 1 shows a flowchart of the refinement process for a given N-chain complex. The refinement proceeds in cycles. At the beginning of the process, each chain within the complex is considered as a separate unit. Each cycle picks two random units and seeks to improve the conformation of one unit with respect to the other one. The improvements are done via rigid-body rotations. The search to improve these conformations focuses on their vicinity in order to keep the computational costs low and avoid large changes to the structure. Units are rotated around X, Y and Z coordinate axes from d_0 to d_1 (e.g. -5 to 5) degrees, resulting in $3 \times (d_1 - d_0 + 1)$ new conformations. Each rotation is performed around one of the three axes that goes through the centroid of the unit that is under consideration. After a set of new top conformations is obtained for the selected pair of units, the two units are considered a fixed single unit. The next cycle starts with this new combined unit (Q) and the rest of the units in the complex. The reason for combining the refined units is to avoid impairing their refined relative conformations in the next cycles. The results are energy minimized for 1000 steps to resolve local clashes using NAMD [12] at the end of each cycle. The process ends when all of the chains in the complex are in one combined unit.

As one can expect, the number of obtained conformations will increase exponentially. For this reason, an ET-based scoring function is used to detect *K* top-scoring (lowest energy) transformations at the end of each cycle (see Scoring Function section below). Only the *K* conformations with lowest energy values are fed to the next cycle to be further refined. We used *K*=10 for the experiments of which results are shown in this paper. The output of the program is the top-K conformations generated at the last cycle, which are all refined versions of the input structure.

B. Scoring Function

The energy function that the search seeks to minimize is described below. A new term, E_{dsf}, that consists of effective distance restraints as in [3] and molecular surface complementarity function as in [11], based on evolutionary conservation of residues, is added to the usual Van der Waals and electrostatic terms. The energy function is computed for the set of interface atoms, which is defined for each chain as the atoms within at most 6Å distance to the adjacent chain atoms.

$$E = E_{VdW} + E_{electrostatic} + E_{dsf}$$

A distance restraint is defined as an intermolecular distance d_{iAB} between any interface atom m of an active residue i (see the definition below for active/passive distinction of residues) of chain A (m_{iA}) and any interface atom n of each residue k of chain B (n_{kB}) and vice versa.

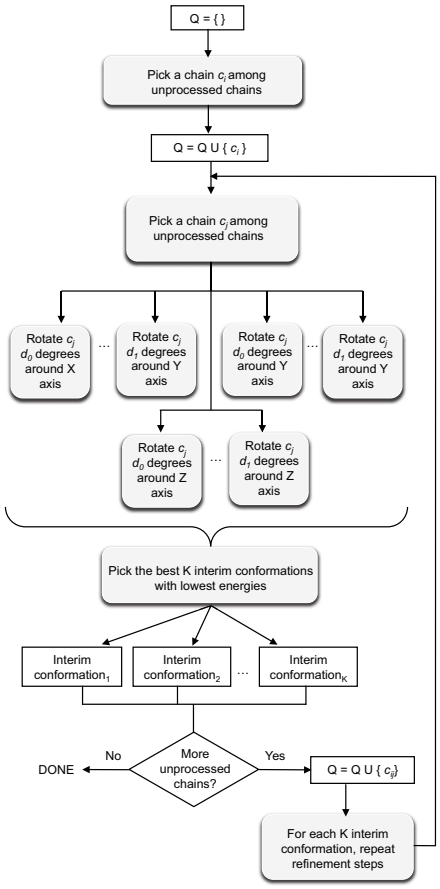


Figure 1. Flowchart of the refinement process.

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{N_{Atoms}} \sum_{k=1}^{N_{ResB}} \sum_{n_{kB}=1}^{N_{Atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{(-1/6)}$$

The sum of each chain's effective distance restraints gives the total effective distance restraints (d^{eff}) for a protein complex. For each active residue on protein A, the contribution of atoms on chain B to the total effective distance restraint is greater if they are in the vicinity of the current active residue; the effect is smaller otherwise. Effective distance restraints are used to prevent active residues to drift apart from the protein interface. If a conformation yields to a high value of effective distance restraint, this means the active residues are too far from the neighbouring chain, hence the native-like interface is not achieved.

Active/passive distinction of residues is made by an ET-

based sequence analysis. Evolutionary conservation values for the interface amino acids are computed as in [11]:

$$c_i = \frac{\int_{-\infty}^{s_i} e^{(t-m)^2/2\sigma^2} dt}{\int_{-\infty}^{\infty} e^{(t-m)^2/2\sigma^2} dt}$$

where s_i is the associated ET score for the i^{th} amino acid as in [9], while m and σ are the mean and standard deviation for the score, respectively. The s_i scores are parsed from ET rank files that were retrieved from the Evolutionary Trace Server [13] for each complex. The score ranges from 0.0 (most variable) to 1.0 (most conserved). The assumption is that evolutionarily-conserved amino acids carry a more important functional role than variable amino acids. Therefore, evolutionarily-conserved amino acids are expected to be found densely on the interface. Amino acids that have conservation values greater than 0.5 are considered active amino acids. The rest are passive.

The same conservation values are also used in computing the degree of interface conservation and surface complementarity [11] :

$$S = \sum_{atompairs} ((-n_A \cdot n_B + 1)/2)^2 \cdot (c_A c_B)^2 \cdot D$$

where n_A and n_B are the normal vectors at each atom on surface A and surface B, respectively; c_A and c_B are the conservation values for each atom pair as explained above; and D is the damping factor. The normal vectors are computed from the corresponding molecular surfaces, which are extracted using the msroll program in Connolly Molecular Surface Package [14]. If the normal vectors of an atom pair have 180° between them, then $-n_A \cdot n_B = 1$, meaning that those atoms are most likely to interact, otherwise the product will be a fraction. If the distance between the pair of atoms, r_{AB} , is greater than 0.5, $D = \frac{1}{4{r_{AB}}^2}$ otherwise $D = 1$.

By using the total effective distance restraints along with the interface conservation and shape complementarity information, we compute E_{dsf} term of our energy function.

$$E_{dsf} = w \times \frac{S}{d^{eff}}$$

where w is a negative signed weight that is adjusted to its optimum value from our experiments. This term evaluates how well two surfaces fit to each other, along with the degree of evolutionary conservation on the surfaces and how close the evolutionarily conserved atoms to the protein interface. As evolutionarily conserved atoms on different surfaces come closer, the value of the denominator will decrease. Structures with higher values of this ratio are desired, and therefore w is a negative weight to have a step-down impact on the total interface energy. The core idea of combining the evolutionary conservation scores and effective distance restraints in order to bias the results towards complexes with native interactions is embodied in this term.

In order to estimate the energies of the resulting complex we add hydrogen atoms to the structures with assumed ideal positions by using Chimera [15], an external molecular structure analysis tool which uses the AMBER Force Field [16]. However, there is always some margin of error in such theoretical models which can lead to a very high energy, especially in the VdW term. For this purpose the Van der Waals term is computed by using a soft Lennard-Jones potential [17], with an attenuated repulsion term. Reducing the repulsion term's power from 12 to 9 reduces the growth rate of the function, resulting in more permissive VdW interactions.

$$E_{VdW} = \sum_{atompairs} \varepsilon[(\frac{r_{ij}}{d_{ij}})^9 - (\frac{r_{ij}}{d_{ij}})^6]$$

The electrostatic term is computed based on Coulomb's law.

$$E_{electrostatic} = 332 \times \sum_{atompairs} \frac{q_i \times q_j}{e \times r_{ij}}$$

where q_i and q_j are the electrostatic charges of atoms i and j taken from AMBER force field [16], e is the dielectric constant (vacuum constant 1 is used for this paper), and r_{ij} is the distance between the ij atom pair. The total value is converted from Coulomb to kcal/mol by multiplying with 332.

Molecular figures shown in this paper and some of the analysis are done by using Visual Molecular Dynamics tool (VMD) [18].

III. Experiments

A. Experimental Setup

Experiments are carried out by two different machines; one having 2.70 GHz Pentium(R) Dual Core processor with 6 GB memory, and the other one having 2 GHz Intel Core Duo processor with 1 GB memory. In this paper we used four structures produced by I. Hashmi (unpublished results) for the Rho/rhogap/gdp(dot)alf4 complex (PDB ID: 1TX4), crystal structure of a Rac-RhoGDI complex(PDB ID: 1DS6), two chains (A and C) of the IkappaBalpha/NF-kappaB (PDB ID: 1IKN), and two chains (V and Y) of the VEGF in complex with domain 2 of the Flt-1 receptor (PDB ID: 1FLT) as input.

B. Results

The results were evaluated in terms of their RMSD values and non-bonded energy values to the corresponding known complexes. The goal was to obtain lower RMSD values as well as lower electrostatic and VdW energy values for the refined structures than the original corresponding docked complexes that we started with. All top ten results with lowest energy values had better or similar RMSD values than the input structure with respect to the known native

Table I
RMSD, ELECTROSTATIC ENERGY AND VAN DER WAALS ENERGY EVALUATIONS OF DOCKED SOLUTIONS AND THEIR REFINED VERSIONS FOR 1DS6, 1TX4, 1IKN AND 1FLT.

	RMSD	Electrostatic	VdW
Docked solution for 1DS6	4.24	-3830.19	1E+10
Refined solution	2.90	-8274.54	-1255.13
Docked solution for 1TX4	5.21	-6022.28	1E+10
Refined solution	3.20	-9807.5	-779.01
Docked solution for 1IKN	5.15	-8487.87	5.98382E+06
Refined solution	3.85	-13563.57	-1334.39
Docked solution for 1FLT	5.82	-2869.33	1E+10
Refined solution	4.50	-4690.03	-705.28

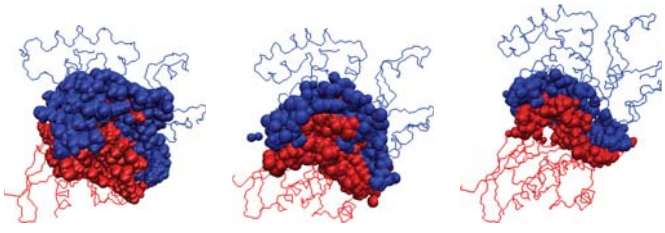
(a) (b) (c)

Figure 2. Initial docked solution for 1DS6 has 1546 interface atoms (a), the refined version of the initial docked solution has 1125 interface atoms (b), and the native structure for 1DS6 has 976 interface atoms (c). Interface atoms are drawn as spheres. Chain A is colored in blue and chain B in red.

structure. Table I shows the lowest-RMSD result for each structure.

In the future, we plan to investigate the effect of conservation scores of residues that are not on the interface versus the residues that are on the interface, in order to evaluate how good a refinement candidate is and assess the predictive ability of our method. If one can tell that a docked structure is an extremely poor candidate to start with (for example, the chains are docked onto each other not at native interface surfaces) in advance, then such candidates can be eliminated from the refinement process to save computation time.

In addition, we would like to enhance our method by detecting flexible parts of chains and applying local minimizations. This way, we can address the issue of refining native-like solutions without causing any displacement of chains from their ideal relative positions.



(a) (b) (c)

Figure 3. Initial docked solution for 1TX4 has 2855 interface atoms (a), the refined version of the initial docked solution has 1797 interface atoms (b), and the native structure for 1TX4 has 891 interface atoms (c). Interface atoms are drawn as spheres. Chain A is colored in blue and chain B in red.

IV. CONCLUSION

Protein complexes play a central role in living organisms. Detection of protein complexes and their structures is crucial for understanding the role of protein complexes in the basic biology of organisms. Computational methods can provide researchers with leverage on the analysis of protein complexes.

We present a method for refining computational docking solutions based on evolutionary information by employing an ET-based scoring function that takes shape complementarity, electrostatic and Van der Waals interactions into consideration. The idea is that evolutionarily-conserved amino acids carry a more important functional role than non-conserved amino acids. Therefore, interfaces that are densely populated with evolutionarily-conserved amino acids are favorable. The method presented here can be employed to refine protein complex structures obtained by docking methods such as [3], [19], [20]

We tested our method on four complexes consisting of two chains. The initial tests show about 2 Angstrom improvements of docked solutions within the best results, in terms of RMSD, and the energy values are also significantly reduced. These improvements are promising for the future of the method.

ACKNOWLEDGMENT

This work was supported by the Theoretical and Computational Biophysics group, NIH Resource for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign.

Hydrogen atoms were added to structures using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081).

Finally, we would like to thank the Lichtarge Computational Biology Lab members at Baylor College of Medicine, Houston, Texas, for providing support with the Evolutionary Trace Server.

REFERENCES

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

[2] S. Vajda and D. Kozakov, "Convergence and combination of methods in protein-protein docking," *Curr. Opinion Struct. Biol.*, vol. 19, no. 2, pp. 164–170, 2009.

[3] C. Dominguez, R. Boelens, and A. Bonvin, "Haddock: A protein-protein docking approach based on biochemical or-biophysical information," *J. Am. Chem. Soc.*, vol. 125, no. 1, pp. 1731–1737, 2003.

[4] D. Schneidman-Duchovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "Geometry based flexible and symmetric protein docking," *Proteins: Struct. Funct. Bioinf.*, vol. 60, no. 2, pp. 224–231, 2005.

[5] C. J. Camacho and S. Vajda, "Protein-protein association kinetics and protein docking," *Curr. Opinion Struct. Biol.*, vol. 12, no. 1, pp. 36–40, 2005.

[6] J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. T. Eyck, "Protein docking using continuum electrostatic and geometric fit," *Protein Eng.*, vol. 14, no. 2, pp. 105–113, 2001.

[7] S. Potluri, A. K. Yan, J. J. Chou, B. R. Donald, and C. Bailey-Kellogg, "Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space, using NMR restraints and van der waals packing," *Proteins: Struct. Funct. Bioinf.*, vol. 65, no. 1, pp. 203–219, 2006.

[8] H. Madaoui and R. Guerois, "Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking," *Proc. Natl. Acad. Sci.*, vol. 105, no. 22, p. 7708, 2008.

[9] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families," *J. Mol. Biol.*, vol. 257, no. 2, pp. 342–358, 1996.

[10] S. Engelen, A. T. Ladislas, S. Sacquin-More, R. Lavery, and A. Carbone, "Joint evolutionary trees: A large-scale method to predict protein interfaces based on sequence sampling," *PLoS Comp Bio*, vol. 5, no. 1, p. e1000267, 2009.

[11] E. Kanamori, Y. Murakami, Y. Tsuchiya, D. Standley, H. Nakamura, and K. Kinoshita, "Docking of protein molecular surfaces with evolutionary trace analysis," *Proteins: Struct. Funct. Bioinf.*, vol. 69, no. 4, pp. 832–838, 2007.

[12] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with namd," *J. Comput. Chem.*, vol. 26, pp. 1781–1802, 2005.

[13] Baylor College of Medicine Lichtarge Computational Lab. Evolutionary Trace Server. [Online]. Available: http://mammoth.bcm.tmc.edu/ETserver.html

[14] M. L. Connolly, "Analytical molecular surface calculation," *J Appl Cryst.*, vol. 16, no. 5, pp. 548–558, 1983.

[15] E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng, and T. Ferrin, "UCSF Chimera–a visualization system for exploratory research and analysis," *J. Comput Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

[16] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, 1995.

[17] A. Ferrari, B. Wei, L. Costantino, and B. Shoichet, "Soft docking and multiple receptor conformations in virtual screening," *J. Med. Chem.*, vol. 47, no. 21, pp. 5076–5084, 2004.

[18] Theoretical and Computational Biophysics group, NIH Resource for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign. Visual Molecular Dynamics. [Online]. Available: http://www.ks.uiuc.edu/Research/vmd/

[19] 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.*, vol. 2, pp. S156–S165, 2005.

[20] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu, "Protein docking with information on evolutionary conserved interfaces," in *Proc. IEEE International Workshop on Comput Struct Biol Workshop (CSBW)*, 2011, submitted.