Gene Ontology terms in scoring of docked protein models

Anna Hadarovich

United Institute of Informatics Problems, National Academy of Sciences, 220012, Minsk, Belarus, `ahadarovich@gmail.com`

Ivan Anishchenko, Petras J. Kundrotas

Center for Computational Biology, The University of Kansas, Lawrence, Kansas 66047, USA

Alexander V. Tuzikov

United Institute of Informatics Problems, National Academy of Sciences, 220012, Minsk, Belarus

Ilya A. Vakser

Center for Computational Biology and Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas 66047, USA

Structural characterization of protein-protein interactions is essential for understanding life processes at the molecular level. Experimental techniques, due to their inherent limitations, can determine structures only for a fraction of known proteins and for even a smaller fraction of known protein interactions. Thus, structures of most protein-protein complexes have to be determined by free or template-based docking. In the template-based docking, the detection of templates requires search against a diverse library of co-crystallized protein-protein complexes according to sequence and/or structure target/template similarity. The docking based on structural similarity often has to be performed on modeled structures of the interactors, which are typically less accurate than the experimentally determined ones [1].

The TM-score that quantifies structural similarity between proteins has been shown to perform well in the template-based docking [2]. However, its performance significantly deteriorates when the templates are only moderately similar to the target (TM-score ~ 0.4 – 0.6) [3]. We propose to complement the TM-score by a scoring function that quantifies similarity of the protein functional properties utilizing hierarchical dictionary (ontology) of the Gene

Ontology (GO) terms (provided by the GO Consortium [4]) in molecular function (MF), biological process (BP), and cellular component (CC) domains.

The combined scoring function consists of the TM-score and a linear combination of three GO-scores:

$$\mathbf{PC} = (\boldsymbol{a} \times \mathbf{GO_{MF}} + \boldsymbol{b} \times \mathbf{GO_{BP}} + \boldsymbol{c} \times \mathbf{GO_{CC}}) \times \mathbf{TM}, \qquad (1)$$

where *a*, *b*, and *c* are weights of the GO-terms from the MF, BP and CC domains, respectively. The proposed values of the coefficients were determined by maximizing the AUC (area under precision-recall curve) for the docking results of 587 protein-protein complexes from the DOCKGROUND resource (http://dockground.compbio.ku.edu) [5]. Precision was calculated as the fraction of near-native models (ligand RMSD from the native structure < 10 Å) among all models with the scoring function larger than a threshold, and recall is the fraction of the near-native models above this threshold among all near-native models.

The scoring function (1) was benchmarked by template-based docking [6] using DOCKGROUND model set 2 as targets and 4,950 DOCKGROUND templates [8]. The model set consists of 165 protein-protein complexes along with six models for each structure with predefined values of C^{α} RMSD from the native structure in 1 - 6 Å range [7].

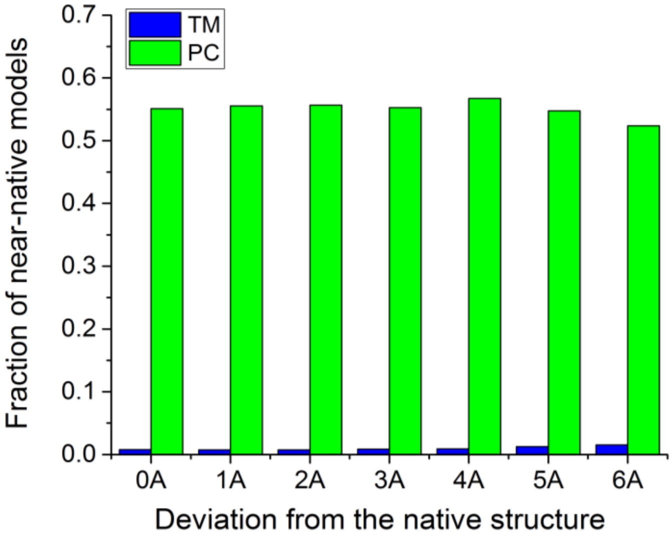

Figure 1. Fraction of near-native models in 0.4 - 0.6 scoring interval.

The AUC value shows cumulative docking performance consisting of three parts: (***i***) docking success rate (fraction of targets, for which at least one model in top *n* models is a near-native one, defined here as a model with ligand C^{α} RMSD from the native structure < 10 Å); (***ii***) change in ranking of the near-native models, and (***iii***) reliability of the scoring function value (probability that a model with that value is a near-native one). The combined scoring function (1) did not show significant improvement in the docking success rate, and its impact on ranking was somewhat ambiguous (although with some impressive examples shown below). However, the function dramatically increases the reliability of the score for the X-ray structures (deviation 0 Å) and the models (1 - 6 Å), especially for the lower values of the scoring functions (Figure 1). Importantly, as Figure 1 shows, the combined scoring function is reliable even for the low accuracy models.

An example of significantly improved ranking by the scoring function (1) is in Figure 2. For the target complex of CD1-2 antigen with Beta-2-microglobulin (3dbx), the second best model with L-RMSD 1.8 Å was based on the template complex of MHC H2-TL-T10-129 and Beta-2-microglobulin proteins (1r3h). Because of the structural inaccuracies in the 6 Å model of 3dbx complex, the TM-score between this model and the template was not very high (0.64). Thus the model built on that template was ranked 11 by the TM-score. The combined scoring function (1) increased the rank of this model to 2, because the function related GO-scores are not affected by the distortions of the complex structure.

The results suggest that the scoring function combining the TM-score and the functional GO-scores discriminates the incorrect predictions better than the structural score alone, especially for the less accurate protein models.

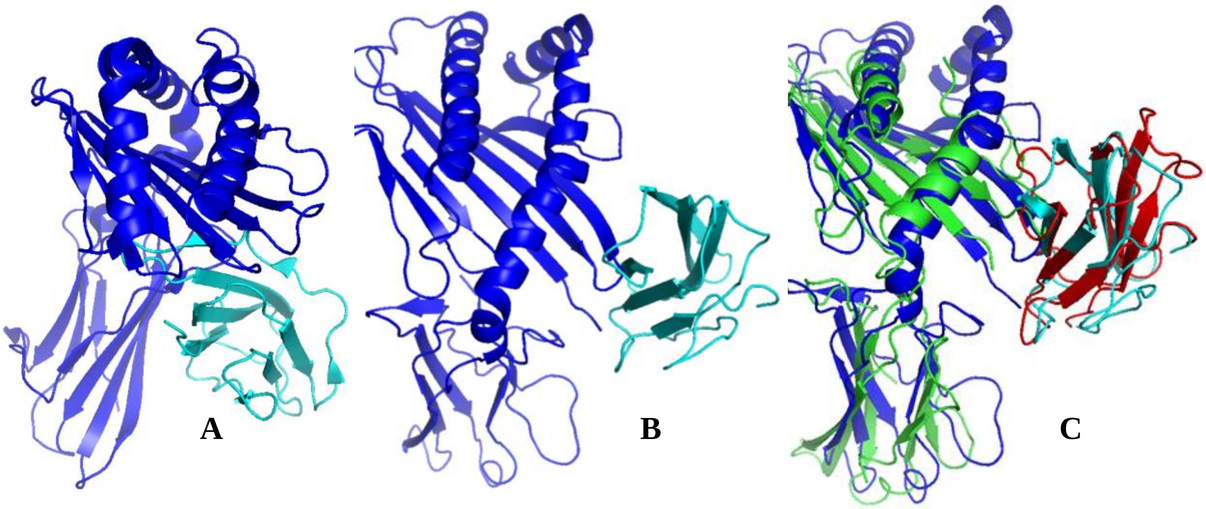

Figure 2. Example of a protein-protein complex with improved template selection. (A) Protein-protein target 3dbx subunits A (blue) and B (cyan). (B) Model of 3dbx with 6 Å deviation from the native structure. (C) Aligned 6 Å model of 3dbx and template 1r3h subunits A (green) and B (red).

References:

1. Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., Vakser, I. A. (2014) Protein models: the Grand Challenge of protein docking. *Proteins*, 82(2): 278-287.
2. Zhang, Y., Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7): 2302-2309.
3. Negroni, J., Mosca, R., Aloy, P. (2014) Assessing the applicability of template-based protein docking in the twilight zone. *Structure*, 22(9): 1356-1362.
4. Gene Ontology, C., Blake, J. A., Dolan, M., Drabkin, H., et al. (2013) Gene Ontology annotations and resources. *Nucleic Acids Res*, 41(Database issue): D530-535.
5. Douguet, D., Chen, H. C., Tovchigrechko, A., Vakser, I. A. (2006) DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics*, 22(21): 2612-2618.
6. Kundrotas, P. J., Zhu, Z., Janin, J., Vakser, I. A. (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *PNAS*, 109(24): 9438-9441.
7. Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., Vakser, I. A. (2015) Protein models docking benchmark 2. *Proteins*, 83(5): 891-897.
8. Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., Vakser, I. A. (2015) Structural templates for comparative protein docking. *Proteins*, 83(9): 1563-1570.