Quantitative comparison of functional properties in protein-protein complexes

Anna Hadarovich

Belarusian State University, 220050, Minsk, Belarus, annchameleon@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 (PPI) is essential for understanding life processes at the molecular level. Because of the inherent limitations of the experimental techniques, only a fraction of known PPI has experimentally solved structures. Computational methods are needed to bridge this gap in structural information. Structural modeling of PPI (protein docking) can be roughly divided into: (i) free docking, where sampling of the binding modes is performed regardless of the possible existence of similar experimentally determined structures, and (ii) template-based or comparative docking, where such similar complexes (templates) determine docking predictions. The detection of suitable templates requires search against a diverse library of protein-protein complexes according to some measure reflecting sequence and/or structure similarity between the target and the template. The widely used TM-score [1] has been shown to perform well in the comparative protein docking [2]. However, the performance significantly deteriorates when templates share only moderate structural similarity with the target (TM-score ~ 0.4 – 0.6) [3]. We propose to complement the TM-score by Gene Ontology (GO) annotations (GO-score), which account for similarity of functional properties of interacting proteins.

The GO-score is based on a hierarchical dictionary (ontology) of the GO-terms (provided by the Gene Ontology Consortium [4]) in three domains: molecular function, biological process, and cellular component. Most studies on GO annotations consider only one domain (molecular function), because of a weaker correlation between the two remaining domains and the sequence or structure similarity [5]. We show that the use of all three ontology domains results in a better selection of protein-protein templates.

A number of existing semantic similarity-based algorithms for calculating the GO-score between single proteins [7] were tested on protein-protein complexes. All such scores were based on the concept of the information content, which relateson the significance level of each particular GO-term to the frequency of its occurrence in a reference database (UniProtKB in our case). The algorithm of Schlicker et al. [6] showed superior performance over the other methods, and thus was adopted in this study.

We tested the TM-score and the three GO-scores, one for each domain of the ontology, combined in a linear function

$$\textbf{TM-score} + \textbf{GO-score} = \boldsymbol{a} * \boldsymbol{GO_{mol}} + \boldsymbol{b} * \boldsymbol{GO_{bio}} + \boldsymbol{c} * \boldsymbol{GO_{cell}} + \boldsymbol{d} * \boldsymbol{TM\text{-}score}, \qquad (1)$$

where a,b,c,d are coefficients reflecting the contributions from the 'molecular function,' 'biological process,' and 'cellular component' domains of the GO and the TM-score, respectively. The coefficient values were obtained on a set of 807 hetero complexes from the library of docking templates [8] by maximizing the number of cases with the rank of a good model (ligand RMSD [LRMSD] from the native structure < 10 Å) scored by (**1**) better than the rank by the TM-score alone.

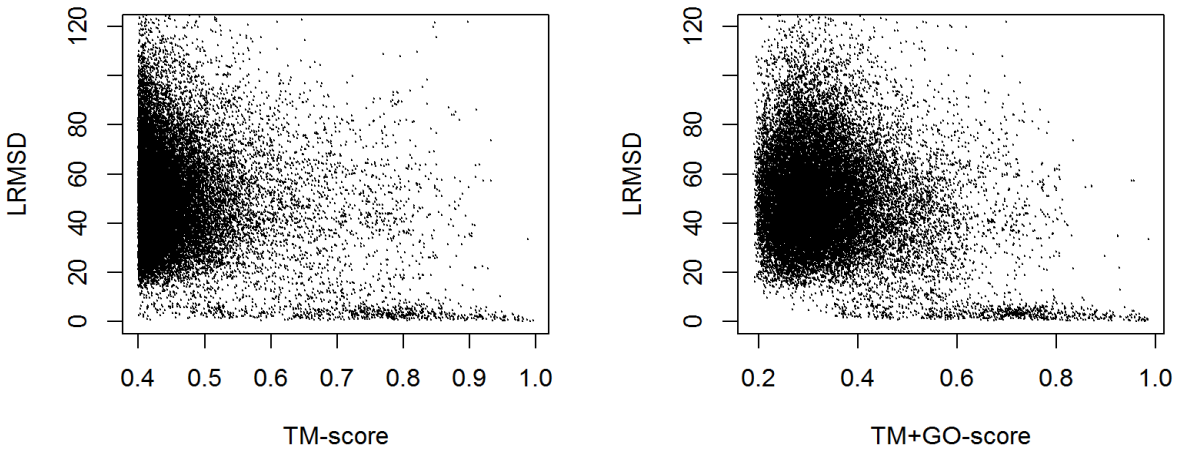

Figure 1. Ligand RMSD vs. TM-score (left) and the combined TM+GO-score (right).

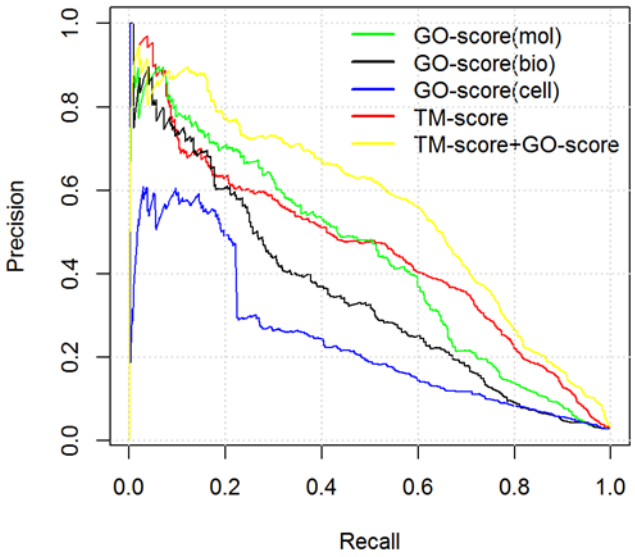

Figure 2. Precision-recall curves for the TM-score and the GO-scores for molecular function, biological process and cellular component.

The scoring function (1) was tested in actual template-based docking [9] on a set of 587 protein-protein complexes from the DOCKGROUND resource (http://dockground.compbio.ku.edu), with better separation of good models compared to the to the TM-score (better clustering of points in the LRMSD < 10 Å region).

The better performance of function (**1**) was confirmed by the precision-recall curves shown in Figure 2 (precision as the fraction of good models among all models, and recall as the fraction of good models recovered at a given threshold among all good models). At the same time, the GO-scores based on a single GO domain performed worse than the TM-score alone (smaller areas under the curves in Figure 2). Among the three GO domains the molecular function domain performs best, close to the TM-score.

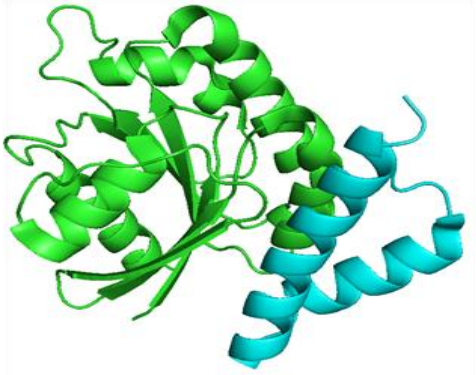

Figure 3. Example of a protein-protein complex with an improved template selection. The protein-protein target is 1j2j subunits A and B.

An example of significantly improved ranking for the good model by the scoring function (**1**) is in Figure 3. For the target complex of ADP-ribosylation factor 1 and its binding protein GGA1 (1j2j), the best model with L-RMSD 9.4 Å was built based on the template complex of AtVSP9a and AtRABF2b proteins (2efd). This model was ranked 91 by the TM-score and 6 by the function (**1**).

Our results suggest that functional attributes of protein-protein complexes can be formally compared, enhancing the conventional sequence- and structure-based measures by afunctional term. The proposed similarity measure based on GO-terms can be used in comparative modeling of protein-protein complexes by improving the choice of templates for the model building. In the future, we intend to conduct a thorough analysis of the GO-scores and their combination with TM-score.

References:

1. Y. Zhang, J. Skolnick (2004) Scoring Function for Automated Assessment of Protein Structure Template Quality, *Proteins,* **57**:702–710.
2. Y. Zhang, J. Skolnick (2005) TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Research,* **33**:2302–2309.
3. J. Negroni, R. Mosca, P. Aloy (2014) Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone, *Structure,* **22**: 1356–1362.
4. The Gene Ontology Consortium (2013) Gene Ontology Annotations and Resources, *Nucleic Acids Research,* **41**:D530–D535.
5. P. Lord et al. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics,* **19**:1275–1283.
6. A. Schlicker et al. (2006) A new measure for functional similarity of gene products based on Gene Ontology, *BMC Bioinformatics,* **7**:302.
7. F. M. Couto, M. J. Silva, P. M. Coutinho (2007) Measuring semantic similarity between Gene Ontology terms, *Data & Knowledge Engineering,* **61**:137-152.
8. I. Anishchenko, P.J. Kundrotas, A.V. Tuzikov, I.A. Vakser (2014) Structural templates for comparative protein docking, *Proteins,* doi: 10.1002/prot.24736.
9. P.J. Kundrotas, Z.W. Zhu, J. Janin, I.A. Vakser (2012) Templates are available to model nearly all complexes of structurally characterized proteins,*Proceedings of the National Academy of Sciences of the United States of America*, **109**: 9438-9441.