

Prediction of protein interactions is essential for studying biomolecular mechanisms

Ilya Vakser¹

¹University of Kansas

June 30, 2023

Abstract

Structural characterization of protein interactions is essential for our ability to understand and modulate physiological processes. Computational approaches to modeling of protein complexes provide structural information that far exceeds capabilities of the existing experimental techniques. Protein structure prediction in general, and prediction of protein interactions in particular, has been revolutionized by the rapid progress in Deep Learning techniques. The work of Schweke et al. presents a community-wide study of an important problem of distinguishing physiological protein-protein complexes/interfaces (experimentally determined or modeled) from non-physiological ones. The authors designed and generated a large benchmark set of physiological and non-physiological homodimeric complexes, and evaluated a large set of scoring functions, as well as AlphaFold predictions, on their ability to discriminate the non-physiological interfaces. The problem of separating physiological interfaces from non-physiological ones is very difficult, largely due to the lack of a clear distinction between the two categories in a crowded environment inside a living cell. Still, the ability to identify key physiologically significant interfaces in the variety of possible configurations of a protein-protein complex is important. The study presents a major data resource and methodological development in this important direction for molecular and cellular biology.

Prediction of protein interactions is essential for studying biomolecular mechanisms

Ilya A. Vakser

Computational Biology Program, The University of Kansas, Lawrence, Kansas, USA

Correspondence to: Ilya A. Vakser, Computational Biology Program, The University of Kansas, Lawrence, Kansas 66047, USA. Phone: 785-864-1057; Email: vakser@ku.edu

Keywords : protein recognition; protein docking; structure prediction; cellular processes; molecular modeling

Total number of words : 1255

Abstract

Structural characterization of protein interactions is essential for our ability to understand and modulate physiological processes. Computational approaches to modeling of protein complexes provide structural information that far exceeds capabilities of the existing experimental techniques. Protein structure prediction in general, and prediction of protein interactions in particular, has been revolutionized by the rapid progress in Deep Learning techniques. The work of Schweke et al. presents a community-wide study of an important problem of distinguishing physiological protein-protein complexes/interfaces (experimentally determined or modeled) from non-physiological ones. The authors designed and generated a large benchmark set of physiological and non-physiological homodimeric complexes, and evaluated a large set of scoring functions, as well as AlphaFold predictions, on their ability to discriminate the non-physiological interfaces. The problem of

separating physiological interfaces from non-physiological ones is very difficult, largely due to the lack of a clear distinction between the two categories in a crowded environment inside a living cell. Still, the ability to identify key physiologically significant interfaces in the variety of possible configurations of a protein-protein complex is important. The study presents a major data resource and methodological development in this important direction for molecular and cellular biology.

Structural characterization of protein-protein interactions is essential for our ability to understand and modulate molecular mechanisms of life processes. Computational approaches are increasingly important for studies of protein interactions, providing structural information that far exceeds capabilities of the existing experimental techniques. Beyond generating structural data, the ability to model protein interactions leads to true understanding of the molecular mechanisms.

Prediction of protein interactions, as a term, has dual meaning. One is predicting that the proteins do interact - e.g., predicting the *existence* of interaction, or predicting the interactors, often in the context of reconstructing the networks of protein interactions in physiological processes. The other is predicting the *mode* of interaction, given the existence of the interaction established by other means, experimental or computational. Protein docking (prediction of the structure of protein complexes) [1], as a branch of life sciences, traditionally has been focusing primarily on the second problem. Docking algorithms have not been specifically designed for or capable of distinguishing interacting and non-interacting proteins, or physiological and non-physiological interactions/interfaces. Typical docking scores, resulting from the global search of the relative positions of the proteins within a complex, are correlated with the energy of interaction (otherwise they would not be able to make correct predictions of the mode of interaction). However, such correlation has been too loose for distinguishing weakly interacting (non-interacting) from strongly interacting molecular pairs, based on the docking scores alone. Re-evaluation of the global search docking predictions by more accurate, and consequently more computationally expensive, scoring functions is essential [2]. While improving docking predictions, scoring also has a potential to distinguish physiological and non-physiological predicted or experimentally determined interfaces. Combining structure-based methodologies with complementary approaches that are based on protein cellular co-localization, sequence analysis and such, should improve our ability to correctly characterize protein interactions.

Protein structure prediction in general has been revolutionized by the rapid progress of the Deep Learning techniques. Most notable is a spectacular success of the DeepMind's AlphaFold that provides high quality prediction of the tertiary structure [3]. Protein-protein docking methodology originally was based primarily on the energy-based considerations, including the concept of steric fit [4], later extended to comparative modeling [5]. A major development in the field has been the recent emergence of the Deep Learning based approaches [6, 7] that extend the prediction success from the tertiary to the quaternary structure. An important activity in the protein docking/scoring field are the community-wide Critical Assessment efforts [8], which provide platforms for the objective blind assessment of the predictive approaches. The progress in structural modeling of macromolecular complexes is propelling growing interest in larger systems, up to the level of a whole cell [9-11]. Such modeling addresses protein interactions *in vivo*, in a crowded cellular environment.

In this issue of *Proteomics*, the paper by Schweke et al. [12], centered around 3DBioInfo Elixir Community, presents a community-wide study of an important problem of distinguishing physiological protein-protein complexes/interfaces (experimentally determined or modeled) from non-physiological ones. The authors designed and generated a large benchmark set of physiological and non-physiological homodimeric complexes, and evaluated a large set of scoring functions, as well as AlphaFold predictions, on their ability to discriminate the non-physiological interfaces.

The problem of separating physiological from non-physiological interfaces (sometimes, called "biological" vs. "non-biological" respectively) is very difficult. The core reason for this difficulty is in the lack of a clear "iron-clad" distinction between these two categories in a living cell, where co-localized proteins extensively interact with each other in a crowded environment in a variety of interaction modes, heavily dominated by transient interactions/encounters. Still, the ability to identify key physiologically significant interfaces in

the variety of possible configurations of a protein-protein complex is important. The study presents a major data resource and methodological development in this important direction of molecular and cellular biology.

A major challenge for the future is developing the ability to differentiate protein-protein physiological vs. non-physiological *hetero* complexes. Due to the diversity and versatility of the protein interactions, this problem is inherently more difficult than that of the *homo* oligomers. A solution may require accounting for various aspects of macromolecular interactions in the crowded cellular environment and significant involvement of the Deep Learning techniques.

References

- [1] Vakser, I. A., Protein-protein docking: From interaction to interactome. *Biophys. J.* 2014, *107* , 1785-1793.
- [2] Moal, I. H., Moretti, R., Baker, D., Fernandez-Recio, J., Scoring functions for protein-protein interactions. *Curr. Opin. Struct. Biol.* 2013, *23* , 862–867.
- [3] Jumper, J., Evans, R., Pritzel, A., Green, T., *et al.* , Highly accurate protein structure prediction with AlphaFold. *Nature* 2021, *596* , 583-589.
- [4] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., *et al.* , Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* 1992, *89* , 2195-2199.
- [5] Kundrotas, P. J., Zhu, Z., Janin, J., Vakser, I. A., Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. USA* 2012, *109* , 9438–9441.
- [6] Evans, R., O'Neill, M., Pritzel, A., Antropova, N., *et al.* , Protein complex prediction with AlphaFold-Multimer. *bioRxiv*2022, [doi:org/10.1101/2021.10.04.463034](https://doi.org/10.1101/2021.10.04.463034) .
- [7] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., *et al.* , Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021, *373* , 871-876.
- [8] Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D., Bates, P. A., Critical assessment of methods for predicting the 3D structure of proteins and protein complexes. *Ann. Rev. Bioph.* 2023, *52* , 183-206.
- [9] Heo, L., Sugita, Y., Feig, M., Protein assembly and crowding simulations. *Curr. Opin. Struct. Biol.* 2022, *73* , 102340.
- [10] Vakser, I. A., Deeds, E. J., Computational approaches to macromolecular interactions in the cell. *Curr. Opin. Struct. Biol.* 2019, *55* , 59-65.
- [11] Vakser, I. A., Grudinin, S., Jenkins, N. W., Kundrotas, P. J., Deeds, E. J., Docking-based long timescale simulation of cell-size protein systems at atomic resolution. *Proc. Natl. Acad. Sci. USA*2022, *119* , e2210249119.
- [12] Schweke, H., Xu, Q., Tauriello, G., Pantolini, L., *et al.* , Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. *Proteomics* 2023.