Bioinformatics, 2019, 1–5
doi.10.1093/bioinformatics/btz587
Application Notice

Data and text mining

DISPOT: A simple knowledge-based protein domain interaction statistical potential

Oleksandr Narykov [1,*], Dmytro Bogatov [2], Dmitry Korkin [1,3,*]

[1] Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA, [2] Department of Computer Science, Boston University, Boston, MA, USA and [3] Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 20, 2109; revised on June 17, 2019; editorial decision on July 15, 2019; accepted on July 22, 2019

Abstract

Motivation: The complexity of protein-protein interactions (PPIs) is further compounded by the fact that an average protein consists of two or more domains, structurally and evolutionary independent subunits. Experimental studies have demonstrated that an interaction between a pair of proteins is not carried out by all domains constituting each protein, but rather by a select subset. However, finding which domains from each protein mediate the corresponding PPI is a challenging task.
Results: Here, we present Domain Interaction Statistical POTential (DISPOT), a simple knowledge-based statistical potential that estimates the propensity of an interaction between a pair of protein domains, given their SCOP family annotations. The statistical potential is derived based on the analysis of more than 352,000 structurally resolved protein-protein interactions obtained from DOMMINO, a comprehensive database on structurally resolved macromolecular interactions.
Availability and implementation: DISPOT is implemented in Python 2.7 and packaged as an open-source tool. DISPOT is implemented in two modes, *basic* and *auto-extraction*. The source code for both modes is available on GitHub: https://github.com/korkinlab/dispot and standalone docker images on DockerHub: https://hub.docker.com/r/korkinlab/dispot. The web-server is freely available at http://dispot.korkinlab.org/.
Contact: korkin@korkinlab.org or onarykov@wpi.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large-scale characterization of protein-protein interactions (PPIs) using high-throughput interactomics approaches, such as yeast-two-hybrid (Y2H) and tandem-affinity purification/mass spectrometry (TAP/MS) approaches (Gavin *et al.*, 2002; Rolland *et al.*, 2014), have provided the scientists with the new insights of the cell functioning at the systems level and allowed to better understand the molecular machinery underlying complex genetic disorders (Barabasi and Oltvai, 2004; Cui *et al.*, 2015; Mitra *et al.*, 2013). Structural studies of protein-protein interactions have revealed that a protein-protein interaction is often carried out by smaller structural protein subunits, protein domains (Ekman *et al.*, 2005; Jin *et al.*, 2009; Vogel *et al.*, 2004). Roughly two-thirds of eukaryotic and more than one-third of prokaryotic proteins are estimated to be multi-domain proteins (Ekman *et al.*, 2005), and thus it is not surprising that $\approx 46\%$ of structurally resolved interactions are domain-domain interactions (Kuang *et al.*, 2016). A high-throughput breakdown of the interactome at this, domain-level, resolution is a much more experimentally challenging task, currently unfeasible at the whole-system level and requiring computational methods to step in (Deng *et al.*, 2002; Finn *et al.*, 2004; Ohue *et al.*, 2014; Segura *et al.*, 2015).

Here, we present a simple knowledge-based Domain Interaction Statistical POTential (DISPOT), a tool that leverages the statistical information on interactions shared between the homologous domains from structurally defined domain families. The knowledge-based potentials are extracted from our comprehensive database of structurally resolved macromolecular interactions, DOMMINO (Kuang *et al.*, 2016). Our statistical potential can be integrated into protein-protein interaction prediction methods that deal with multidomain proteins by ranking all possible pairwise combinations of domain interactions between the two or more proteins.

2 Methodology

The development of DISPOT is driven by several observations. First, an average interaction between a pair of proteins is not carried out by all domains constituting each protein, but only by a select subset. Indeed, each domain has its unique structure and biological function and may not be designed to interact with a particular domain from another protein (Banappagari *et al.*, 2010; Shimizu *et al.*, 2016). Second, the domain-domain interactions often share homology: when two homologous domains interact with their partners, these partners frequently also share the homology with each other (Kuang *et al.*, 2016). Thus, one can introduce the domain-domain interaction propensity in terms of the frequency of domain-domain interactions between the two domain families. Lastly, the propensity of domain-domain interaction is expected to vary across different families, thus allowing to provide the finer resolution of the protein-protein interaction network.

The quantification of the odds for a domain from one domain family to interact with a domain from another family is defined in this work as a knowledge-based statistical potential. Statistical potentials are widely used in biophysical applications, often for characterizing the residue contacts between the protein chains (Huang and Zou, 2008; Krüger *et al.*, 2014; Lu *et al.*, 2003). One of the main applications of the residue-level statistical potentials is in protein docking (Kozakov *et al.*, 2006). Our domain-domain statistical potential complements the residue-level potentials by considering structural units from the higher-level of protein structure hierarchy and requiring no structural information about the protein domains. Specifically, the input for DISPOT includes the protein sequences of the two proteins interacting with each other.

First, the domain architecture of each protein is obtained. To do so, a region of the protein sequence is annotated to a family of homologous domains. For the definition of domain families, we leverage the Structural Classification of Proteins (SCOP) family-level classification (Andreeva *et al.*, 2004). SCOP represents a structure-based hierarchical classification of relationships between protein domains or single-domain proteins with 'family' being the first level of SCOP classification and 'superfamily' being the second level. Protein domains from the same SCOP family are evolutionary closely related and often share the same function. Since a protein with no structural information cannot be directly annotated by SCOP, we use SUPERFAMILY (Gough and Chothia, 2002), a Hidden Markov Model (HMM) based approach that maps regions of a protein sequence to one or several SCOP families or superfamilies. SUPERFAMILY allows us to cover a substantial subset of known proteins: the HMM coverage at the protein sequence and amino acid levels for the UniProt database were reported at 64.73% and 58.78% respectively in 2014 (Oates *et al.*, 2014).

Second, for each pair of SCOP families we count a number of non-redundant protein-protein interactions between the members of these families that have been experimentally determined. Our source of data is DOMMINO (Kuang *et al.*, 2016, 2011) a comprehensive database of structurally resolved macromolecular interactions. It contains information about interactions between the protein domains, interdomain linkers, terminal sequences, and protein peptides. In this work, we use exclusively domain-domain interactions because the data about this type of interactions is the most abundant. To remove redundancy in the data, we use ASTRAL compendium (Brenner *et al.*, 2000) which is integrated into the SCOPe database (Fox *et al.*, 2013). From ASTRAL, we obtain a set of domains, where each domain shares less than 95% sequence identity to any other domain in the set. This set is then used to determine pairs of redundant domain-domain interactions in the original DOMMINO dataset. Two domain-domain interactions are determined as redundant if both corresponding pairs of domains share 95% or more sequence identity. For each pair of redundant domain-domain interactions, one interaction is randomly removed. The process continues until no pair of redundant interactions can be detected.

Third, for each domain family from each protein, a statistical potential is calculated (Fig. 1A, 1B and Supplementary Materials Figure S1). There are two types of statistical potentials introduced in this work: (1) calculated for a domain from a specific domain family, and (2) calculated for a pair of domains, one domain from each of the two interacting proteins. The statistical potential P_i for a single domain D_i is calculated based on the total number of interactions N_{D_i} extracted from the non-redundant DOMMINO dataset for the specific SCOP family this domain belongs to. The statistical potential P_{ij} for a pair of domains, D_i and D_j, is calculated based on the total number of occurrences N_{ij} of the interactions between all domains from the same two SCOP families as D_i and D_j. Those numbers are then transformed into probabilities as follows:

$$P_i = \frac{1}{Z_1} \ln \frac{N_{p_i}}{N_{\text{mean}}} \qquad Z_1 = \sum \ln \frac{N_{p_k}}{N_{\text{mean}}}$$

$$P_{ij} = \frac{1}{Z_2} \ln \frac{M_{p_{ij}}}{M_{\text{mean}}} \qquad Z_2 = \sum \sum \ln \frac{M_{p_{kl}}}{M_{\text{mean}}}$$

where N_{mean} is an average number of interactions for a domain family and M_{mean} is an average number of interactions for a pair of domain families, both calculated from the non-redundant DOMMINO set.

DISPOT potentials are derived following a standard strategy for calculating a statistical potential. The statistical potentials for the atomic contact pairs are traditionally derived based on Boltzmann relation (Huang and Zou, 2008):

$$P_{ij} = -k_B T \ln \frac{p_{ij}(r)}{p^*_{ij}}$$

where k is the Boltzmann constant, T is the system's temperature, p_{ij} is an experimentally observed density of atom pairs from different partners in a complex at distance and p^*_{ij} is corresponding density in the reference state. Since we do not work with the atomic-level physical interactions, we replace the Boltzmann constant from DISPOT equations and substitute temperature with the inverse of normalization constant Z. In addition, p_{ij} and p^*_{ij} are substituted with the number of interactions between domains in DOMMINO database.

DISPOT can also provide integrated protein-level statistics. There are multiple ways to combine the domain-level statistics into a protein-level statistics. Two simple approaches to integrate domain-domain interactions for a given protein-protein interaction in terms of a standalone (single protein) and interaction (protein pair) potentials are:

$$P_{M_u} = \max_i P_i \qquad \text{and} \qquad P_{M_{uv}} = \max_{i,j} P_{ij}$$

respectively, where i and j correspond to the domains from protein u and v. The rationale behind these definitions lies in the assumption that a single strongest domain-domain interaction is the one of the most important defining factor for the PPI. These definitions of cumulative potentials were tested in terms of their

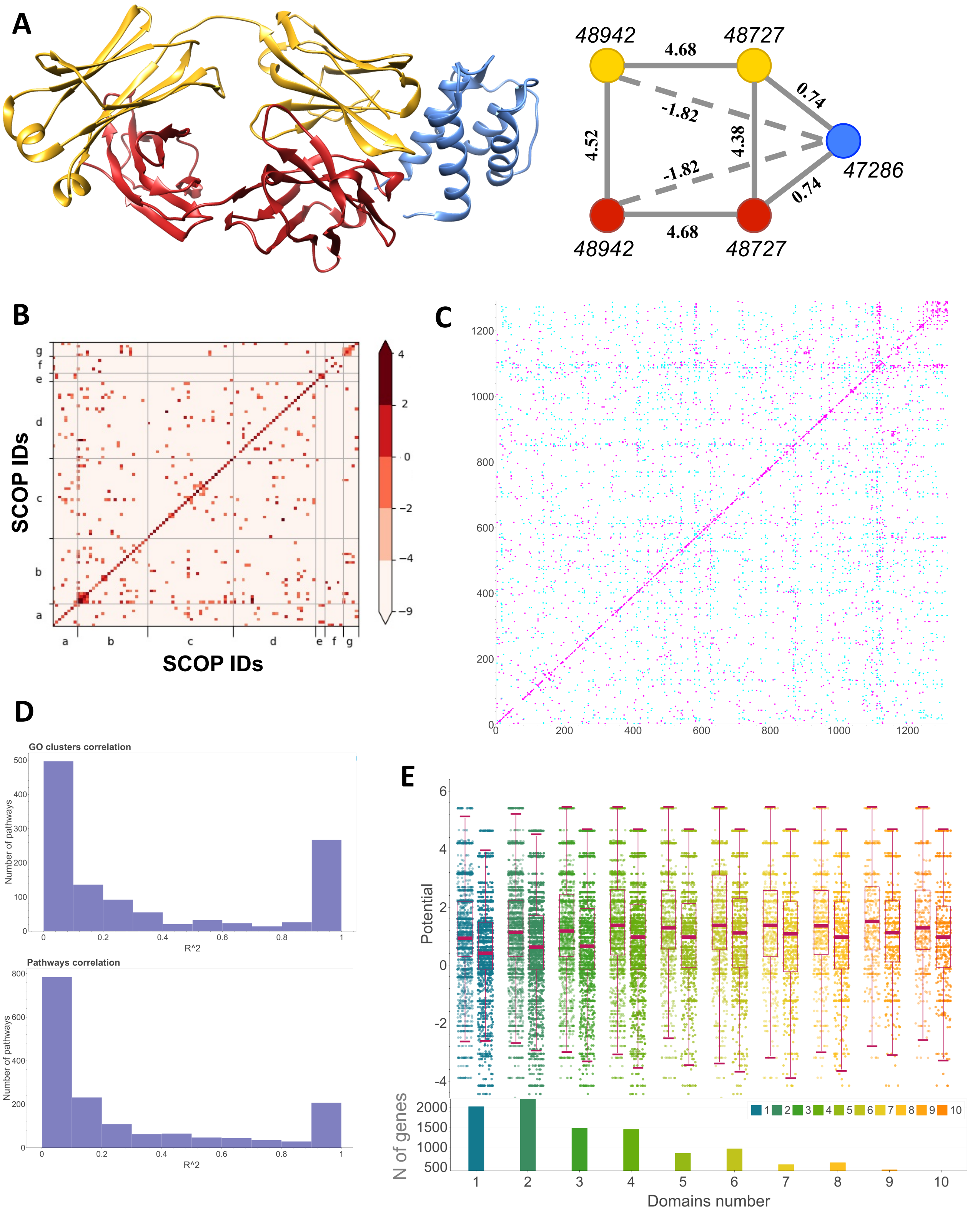

Fig. 1. DISPOT statistical potential and its application. A. A crystal structure (left) of the protein complex between CNTO607 Fab human monoclonal antibody (yellow and red colors denote two different chains) and interleukin-13 (IL-13), and the corresponding domain-domain interaction network (right). Shown in italics are SCOP Family IDs, and in bold are DISPOT values for the corresponding interactions. Nodes colored with the same color belong to the same chain. Solid lines connecting nodes correspond to the physical interaction, while dashed lines connect nodes corresponding to the protein domains that do not physically interact. B. A heatmap showing DISPOT values calculated for each pair of SCOP families, where only potentials for pairs of SCOP families with 5 and more non-redundant interactions are plotted. The families are grouped based on the SCOP class (a-g) and are ordered within each fold based on their IDs. C. A contact map showing the correlation between experimentally obtained human interactome HI-I-05 and DISPOT-based PPI prediction. A prediction that calls a PPI correctly is shown in magenta, while PPIs that were missed are shown in cyan. D. Correlation calculated using R2 correlation coefficient between the hu.MAP interaction probability score and DISPOT statistical potential for KEGG pathways (bottom) and GO clusters (top). E. Distribution of the protein-level DISPOT statistical potentials grouped by the number of SCOP domains in a protein defined using SUPERFAMILY.

ability to predict a PPI using several experimental sources. First, we obtained the coverage landscape by the cumulative potentials on the experimental protein-protein interactomes one obtained using high-throughput yeast-two-hybrid screening (HI-I-05) (Rual *et al.*, 2005) and another one obtained using curated literature-based search (LitBM-17, http://interactome.baderlab.org/data/LitBM-17.psi). As expected, while this naïve method was able to recover 2,944 PPIs in HI-I-05, it missed 1,188 PPIs even using a lenient threshold of -20 (Fig. 1C). Similarly, the cumulative potential was able to recover only 1,718 PPIs while 1,453 PPIs were not recovered (Supplementary Materials Fig S1). We then apply the same pairwise cumulative potential to the large-scale mass spectrometry study (Drew *et al.*, 2017). Specifically we study the correlation between the hu.MAP probability score and cumulative pairwise score among KEGG pathways (Kanehisa and Goto, 2000) and GO clusters produced by GeneSCF on 13,855 genes with SUPERFAMILY annotation (Subhash and Kanduri, 2016) (Fig. 1D). While the number of highly correlated pairs was substantial, the number of pairs with very little correlation still prevailed. Finally, the analysis of the cumulative single potential for a protein showed that it can obtain a diverse range of values and this property seems to be independent of how many domains this protein has (Fig. 1E). Similar behavior was observed when looking at the other basic cumulative measures (Supplementary Materials, Figure S3).

Overall, we have analyzed and summarized interactions from 3,619 SCOP family pairs that were extracted from 352,199 PPIs. In total, domains from 1,384 SCOP families were characterized that form domain-domain interactions in 1,384 'homo-SCOP' interaction pairs (i.e., both domains are annotated with the same SCOP family) and 2,235 'hetero-SCOP' pairs (Fig. 1B and Supplementary Materials Fig. S1). The analysis of the calculated statistical potentials showed a wide diversity across different families.

Finally, we would like to make a cautionary note of using the developed tool. DISPOT was designed not as a PPI prediction tool, but rather a tool that provides additional information on the likelihood of specific domain-domain interactions in a given physical PPI. The main reason is the fact that structural coverage of the protein-protein interaction space is still far from being full, which leads to the presence of a high number of false-negatives if one was to use DISPOT as a stand-alone predictor. This intuition has been supported by our evaluation of DISPOT against the two interactomics golden standards. Thus, if a researcher wants to employ DISPOT in a PPI prediction method, we recommend adding the DISPOT potentials as features to the overall feature vector, that would include other parameters, such as secondary structure, evolutionary conservation of the sequence, predicted residue hydrophobicity, *etc*.

3 Implementation and usage

The basic mode is implemented in Python with the dependency on packages *pandas* and *numpy*. It takes SCOP identifiers (IDs) for either 'family' (`fa`) or 'superfamily' (`sf`) hierarchy levels as an input and produces statistical potential for corresponding pair of domains. Switching between the SCOP levels is implemented in command line option `sf`. One of the possible input options is a command line option domains, which provides a list of space-separated SCOP identifiers. Based on this list, the program produces all possible unique pairwise combinations of identifiers and the corresponding statistical potentials. Option `max` produces the highest value of statistical potential for a selected domain and a SCOP ID for the corresponding interaction domain partner. Option `output` specifies the output file. If no file path is specified, then program opens a console output prompting a user to input the data. A detailed description of all acceptable input formats and options is available in *README* file and help menu of the main script `dispot.py`.

The auto-extraction version relies on the SUPERFAMILY models and scripts and HMMER program for extracting the corresponding SCOP IDs for either family or superfamily levels of hierarchy. The Perl programming language interpreter is an additional dependency. HMMER is compatible with the major linux distributions (it has been tested on Ubuntu 16.04 and Alpine 3.7 with additional installation of `alpine-glibc`). Windows users are advised to use the docker image. The main script is `dispot.py`, and it includes several options: `fasta_folder` — to specify a path to the folder with FASTA files; `output_folder` — to specify a path to the results; and `max` — to substitute the regular output of all pairwise statistical potentials with the highest statistical potential for a given domain family and a SCOP ID of the interaction partner on which this value is achieved. Additional script `batch_process.py` provides almost the same functionality, except it uses the default locations: `./data/` for the input and `./data/results/` for the output. For each FASTA sequence, we extract a SUPERFAMILY-derived SCOP ID and the location(s) of the corresponding domain on the protein sequence. It is stored in the `./tmp/` folder and is available until the next run of any of the scripts mentioned in this section. The data are stored in the Python dictionary objects serialized by package *pickle*.

DISPOT has been also implemented as a web-server that carries the full functionality of the developed methods and comes with a tutorial. The web-server is freely available at http://dispot.korkinlab.org/.

Funding

This work was supported by National Science Foundation (1458267) and National Institute of Health (LM012772–01A1) to D.K.

References

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic acids research*, **32**(1), D226–D229.

Banappagari, S., Ronald, S., and Satyanarayanajois, D. S. (2010). A conformationally constrained peptidomimetic binds to the extracellular region of HER2 protein. *Journal of Biomolecular Structure and Dynamics*, **28**(3), 289–308.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, **5**(2), 101.

Brenner, S. E., Koehl, P., and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic acids research*, **28**(1), 254–256.

Cui, H., Dhroso, A., Johnson, N., and Korkin, D. (2015). The variation game: Cracking complex genetic disorders with NGS and omics data. *Methods*, **79**, 18–31.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. In *Proceedings of the sixth annual international conference on Computational biology*, pages 117–126. ACM.

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., Ma, Y., Wallingford, J. B., and Marcotte, E. M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, **13**(6), 932.

Ekman, D., Bjorklund, A. K., Frey-Skott, J., and Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of molecular biology*, **348**(1), 231–243.

Finn, R. D., Marshall, M., and Bateman, A. (2004). iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**(3), 410–412.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2013). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research*, **42**(D1), D304–D309.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., and Cruciat, C.-M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141.

Gough, J. and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic acids research*, **30**(1), 268–272.

Huang, S.-Y. and Zou, X. (2008). An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Structure, Function, and Bioinformatics*, **72**(2), 557–579.

Jin, J., Xie, X., Chen, C., Park, J. G., Stark, C., James, D. A., Olhovsky, M., Linding, R., Mao, Y., and Pawson, T. (2009). Eukaryotic protein domains as functional units of cellular evolution. *Science signaling*, **2**(98), ra76–ra76.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, **65**(2), 392–406.

Krüger, D. M., Garzón, J. I., Chacón, P., and Gohlke, H. (2014). DrugScorePPI knowledge-based potentials used as scoring and objective function in protein-protein docking. *PloS one*, **9**(2), e89466.

Kuang, X., Han, J. G., Zhao, N., Pang, B., Shyu, C.-R., and Korkin, D. (2011). DOMMINO: a database of macromolecular interactions. *Nucleic acids research*, **40**(D1), D501–D506.

Kuang, X., Dhroso, A., Han, J. G., Shyu, C.-R., and Korkin, D. (2016). DOMMINO 2.0: integrating structurally resolved protein-, RNA-, and DNA-mediated macromolecular interactions. *Database*, **2016**.

Lu, H., Lu, L., and Skolnick, J. (2003). Development of unified statistical potentials describing protein-protein interactions. *Biophysical journal*, **84**(3), 1895–1901.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, **14**(10), 719.

Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J., Sardar, A. J., Zaucha, J., Thurlby, N., Fang, H., and Gough, J. (2014). The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic acids research*, **43**(D1), D227–D233.

Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T., and Akiyama, Y. (2014). MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data. *Protein and letters, peptide*, **21**(8), 766–778.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., and Mosca, R. (2014). A proteome-scale map of the human interactome network. *Cell*, **159**(5), 1212–1226.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., and Ayivi-Guedehoussou, N. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**(7062), 1173.

Segura, J., Sorzano, C. O. S., Cuenca-Alba, J., Aloy, P., and Carazo, J. M. (2015). Using neighborhood cohesiveness to infer interactions between protein domains. *Bioinformatics*, **31**(15), 2545–2552.

Shimizu, M., Noguchi, Y., Sakiyama, Y., Kawakami, H., Katayama, T., and Takada, S. (2016). Near-atomic structural model for bacterial DNA replication initiation complex and its functional insights. *Proceedings of the National Academy of Sciences*, **113**(50), E8021–E8030.

Subhash, S. and Kanduri, C. (2016). GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. *BMC bioinformatics*, **17**(1), 365.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Current opinion in structural biology*, **14**(2), 208–216.