

DAPID: A 3D-Domain Annotated Protein-Protein Interaction Database

Yung-Chiang Chen²

smolljohn.bi93g@nctu.edu.tw

Heng-Chu Chen²

maxos.bi92g@nctu.edu.tw

Jinn-Moon Yang^{1,2}

moon@faculty.nctu.edu.tw

¹ Department of Biological Science and Technology, National Chiao Tung University
Hsinchu, 30050, Taiwan

² Institute of Bioinformatics, National Chiao Tung University Hsinchu, 30050, Taiwan

Abstract

DAPID is a database of domain-annotated protein interactions inferred from three-dimensional (3D) interacting domains of protein complexes in the Protein Data Bank (PDB). The DAPID data model allows users to visualize 3D interacting domains, contact residues, and molecular details of any predicted protein-protein interactions. Our model derives these interactions by utilizing a new concept, called the “3D-domain interologs” which is similar to “interologs”. In *S. cerevisiae*, there is 18.6% overlap between our predicted protein-protein interactions and ones in the DIP database. The mean correlation coefficient of the gene expression profiles of our predicted interactions is significantly higher than that for random pairs in *S. cerevisiae*. In addition, we find several novel interactions which are consistent with the functions of the proteins. The DAPID currently holds 1008 3D-interacting domain pairs and 101511 predicted 3D-domain annotated protein-protein interactions. It is available online at <http://gemdock.life.nctu.edu.tw/dapid>.

Keywords: domain-domain interaction, protein interaction prediction, swiss-prot keyword annotation

1 Introduction

Protein-protein interactions are involved in most biological processes. Identifying their associated networks comprehensively is the key to understanding cellular mechanisms [11]. Many experimental and computational methods have been proposed to identify protein-protein interactions, which were collected in some databases, such as DIP [27], BIND [1], MIPS [22], and STRING [31]. The interaction data obtained from these methods mainly includes physical interactions (proteins interact directly) and functional associations (proteins have related biological functions) [30]. A well-known problem with most large-scale experimental methods, like the two-hybrid system [16] or affinity purifications [23], is the high false-positive rate of their generating protein interactions [32]. Computational approaches which predict protein-protein interactions have used gene expression profiles citebib10, phylogenetic profiles citebib11, known 3D complexes [2, 3, 20], interologs (two proteins will interact with each other if their orthologous proteins do as well) [21], and domain-pair profiles [33]. The development of computational approaches to map interactions seems useful in light of the shortcomings of large-scale experimental methods.

Despite a variety of the strategies and methods have used in identifying protein-protein interactions, only a few of them have paid attentions to known 3D-complexes in the PDB [12] and 3D-interacting domain databases, such as the 3did [29] and iPfam [13]. Generally, known 3D structures of interacting proteins and complexes are able to provide an atomic description of how the interaction probably

occurs and to support the visualization of protein-protein interactions at the molecular level. In addition, it is usually possible to build an interaction model of two proteins by comparative modeling if a known complex structure comprising homologs of these two proteins is available [2, 3, 20].

Here we address these questions using a new concept, the “3D-domain interologs” which is similar to “interologs” [21]. The 3D-domain interologs (Figure 1), the core idea of our DAPID method, is defined as “Domain a (in chain A) interacts with domain b (in chain B) in a known 3D complex, their inferring protein pair A’ (containing domain a) and B’ (containing domain b) in the same species would be likely to interact with each other if both protein pairs (A’ and A as well as proteins B and B’) are homologous”. We have used 3D-domain interologs to predict physical protein-protein interactions and to provide an atomic description of the interfaces ultimately responsible for interaction specificity. Large amount of our predicted protein-protein interactions were evaluated in *S. cerevisiae* based on three factors, including the TP/FP ratio (the ratio of true to false positives), enrichment, and correlation coefficient of gene expression profiles.

To the best of our knowledge, the DAPID is the first database to use 3D-domain interologs for inferring 3D-domain annotated protein interactions. The DAPID is able to provide 3D interacting domains and contact residues for visualizing molecular details of any predicted protein pairs. The DAPID currently holds 1008 3D-interacting domain pairs and 135535 protein-protein interactions, including 101511 interactions (74.9%) derived from our 3D-domain interologs, 1111 interactions (0.8%) directly extracted from PDB complexes, and 32913 interactions (24.3%) summarized from the DIP database (Table 1). The DAPID consists of eight common organism models, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Helicobacter pylori*, and *Escherichia coli*.

Table 1: Statistics of the protein-protein interactions on eight common organism models in the DAPID database. ^(a,b)Interactions are extracted from the DIP and the PDB databases, respectively.

Species	Our method	DIP ^(a)	PDB ^(b)
<i>Homo sapiens</i>	53669	1227	437
<i>Mus musculus</i>	39689	240	297
<i>Rattus norvegicus</i>	4461	91	70
<i>Drosophila melanogaster</i>	857	11847	4
<i>Caenorhabditis elegans</i>	941	3835	1
<i>Saccharomyces cerevisiae</i>	1158	14779	219
<i>Escherichia coli</i>	603	760	82
<i>Helicobacter pylori</i>	133	134	1
Total	101511	32913	1111

2 Materials and Methods

3D-domain interologs

Figure 1 shows the overview of our method using 3D-domain interologs to extract domain annotated protein-protein interactions from 3D protein complexes by performing the following steps: (i) identifying the interactive domains (domains a and b) and contact residues of a 3D complex (containing chains A and B) in the PDB [12]; (ii) projecting the Pfam domains [8] onto the interactive domains in the complex using the domain boundary defined by the Pfam database; (iii) identifying two protein families (A’ and B’) which contain the corresponding domains (a and b) from the Swiss-Prot database [6]; (iv) calculating the homologous scores between the protein templates (A and B) and their corresponding

homologous proteins (A' and B') with three factors, including knowledge annotation, contact residues matching, and sequence similarity; (v) evaluating the joint score of the protein pair (A' and B') by our scoring function. If the score exceeds a threshold, the two proteins are predicted to interact with each other .

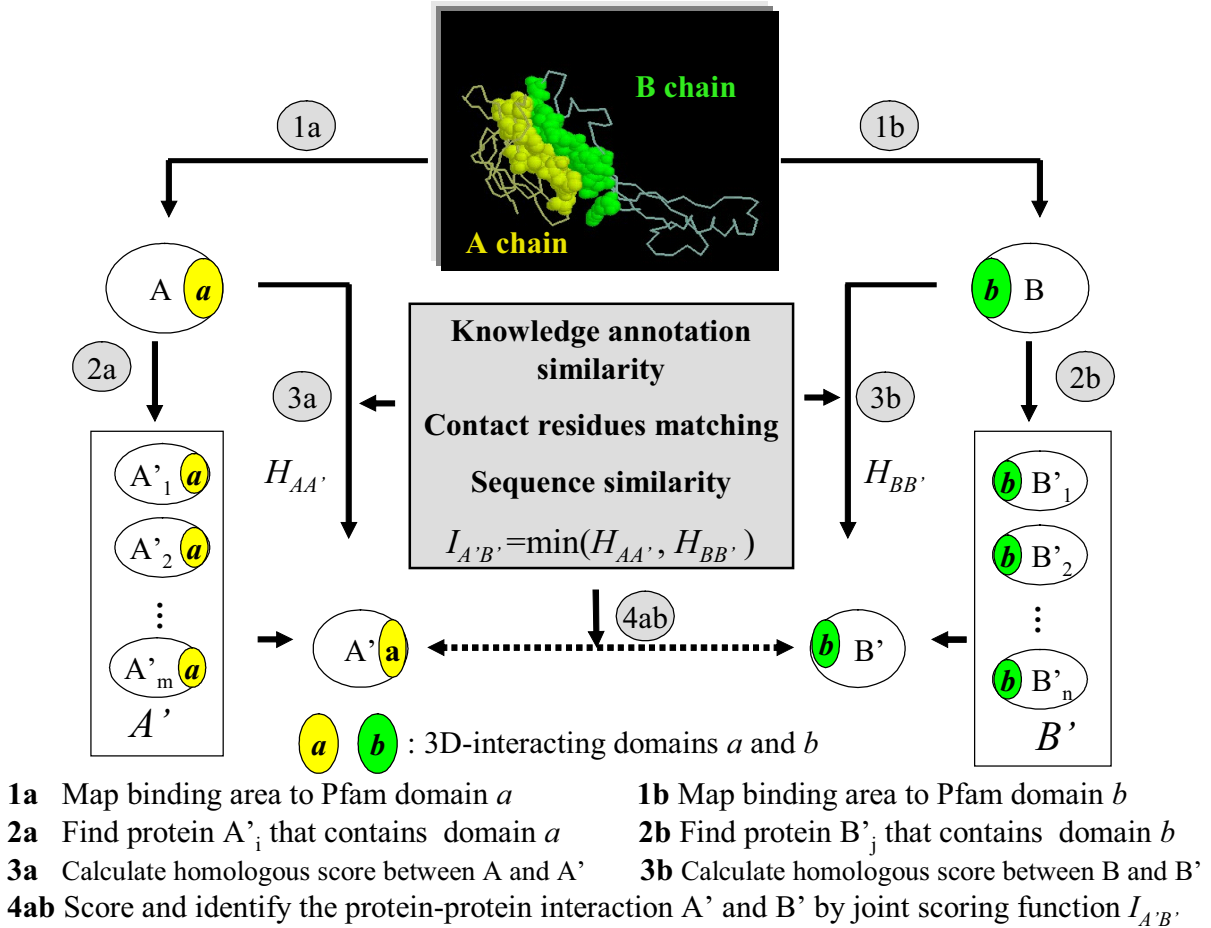


Figure 1: Overview of our method using 3D-domain interologs.

3D-interacting domains

To construct a database of 3D-interacting domains, we used known 3D complexes to identify contact residues of two chains. Contact residues, whose C_β (or C_α in glycine) should be within a threshold (distance $\leq 8 \text{ \AA}$) to any C_β of another chain, were considered as the core parts of the 3D-interacting domains in a complex. Each chain must have more than 5 contact residues and the threshold was chosen to make sure that the contact between the domains was reasonably extensive [24]. A total of 1649 nonredundant hetero-chain pairs were extracted from the PDB when the homo-chain pairs were excluded. To define the domains in the DAPID, we used the domain definitions from Pfam classification database, which well defined the domain boundary in protein sequences. The Pfam domains were then projected onto these 1649 complexes. In this way, 671 nonredundant domains and 1008 3D-interacting domain pairs were obtained.

Protein-protein interaction candidates

We used “3D-domain interolog generalized mapping method” to generate all possible protein-protein interactions as shown in Figure 1. All proteins (A'_i), containing the Pfam domain a from the Swiss-Prot database, are considered as a family (\mathbf{A}') for any given domain (a) in our 3D-interacting domain database. Two interacting families (\mathbf{A}' with m members and \mathbf{B}' with n members) are that at least one protein member (A') of a domain family (\mathbf{A}') interacts with a protein member (B') of the other domain family (\mathbf{B}'). We generated protein-protein interaction candidates by considering all possible protein pairs (e.g. mn pairs in Figure 1) of these two interacting families (\mathbf{A}' and \mathbf{B}'), called generalized 3D-domain interolog mapping method. In this way, a total of 24353629 protein-protein interactions can be yielded from 1008 3D-domain pairs. Finally, 845041 candidates of protein interactions where two interacting proteins are the same species were evaluated.

Joint scoring function

A goal of this work is to measure the reliability of protein-protein interactions derived from 3D-domain interologs. We have developed a new joint scoring function based on knowledge annotation similarity, sequence similarity, and the score of aligned contact residues between a member (A' or B') in the predicted interactive protein pair and its protein template (A or B) (Figure 1). The joint scoring function ($I_{A'B'}$) of proteins A' and B' is defined as the smaller one of two individual homologous scores and given as

$$I_{A'B'} = \min(H_{AA'}, H_{BB'}) \quad (1)$$

where $H_{AA'}$ is the homologous score between the template protein A and its homolog protein A' , $H_{BB'}$ is the homologous score between the template protein B and its corresponding homolog B' . The templates A and B are the 3D-interacting chains in a complex and proteins A' and B' are two predicted interacting proteins, which contain a domain pair a and b , respectively (Figure 1). The $H_{AA'}$ is defined as

$$H_{AA'} = K_{AA'} + 0.5(E_{AA'} + S_{AA'}) + D_{AA'} \quad (2)$$

where $K_{AA'}$ is the score of the keyword annotation defined in the Swiss-Prot database [9]; $E_{AA'}$ ($E_{AA'} = -\log(\text{e-value})$) and $S_{AA'}$ are the scores of e-value and sequence similarity, respectively, of aligning the two protein sequences (A and A') by using PSI-BLAST [4]; $D_{AA'}$ is the score of the aligned contact residues based on BLOSUM62 substitution matrix. We normalize these scores ranging from 0 to 1 to correctly combine these scores.

In order to calculate the score ($K_{AA'}$) of keyword annotations, we adapt and modify the well-known *TF-IDF* scoring scheme commonly used in document retrieval systems, where *TF* is the keyword frequency in a given protein and *IDF* is the inverse protein frequency of the keyword [5]. In the Swiss-Prot database, TF_i is 1 and IDF_i is equal to $\log_2(N/n_i)$ for a given keyword i in a protein where N (188477 in Release 47.5) is the total number of proteins and n_i is the number of proteins comprising the keyword i . The *TF-IDF* weight (W_{A_i}) of the keyword i in the protein A is defined as $TF_i \times IDF_i$. Given a protein pair A and A' , their score ($K_{AA'}$) of the keyword annotation can be calculated as

$$K_{AA'} = \sum_{i=1}^M (W_{A_i} W_{A'_i}) / \sqrt{\left(\sum_{i=1}^M W_{A_i}^2\right) \times \left(\sum_{i=1}^M W_{A'_i}^2\right)} \quad (3)$$

where M is the total number of keywords in the Swiss-Prot database; W_{A_i} and $W_{A'_i}$ are the *TF-IDF* weights of the keyword i in the proteins A and A' , respectively.

3 Results

Quality assessment

Three factors were used to evaluate the performance of our method, including the TP/FP ratio, enrichment, and correlation coefficient of gene expression profiles (Figure 2). The TP/FP ratio is defined as A_h/F_h , where A_h and F_h are the numbers of known positives (i.e., interacting protein pairs) and negatives (i.e., non-interacting protein pairs) in the predicted protein-protein interactions, respectively. The enrichment is defined as $(A_h/T_h)/(A/T)$ where T_h is the number of predicted protein-protein interactions; A is the total number of positives; T is the total number of possible interacting protein pairs.

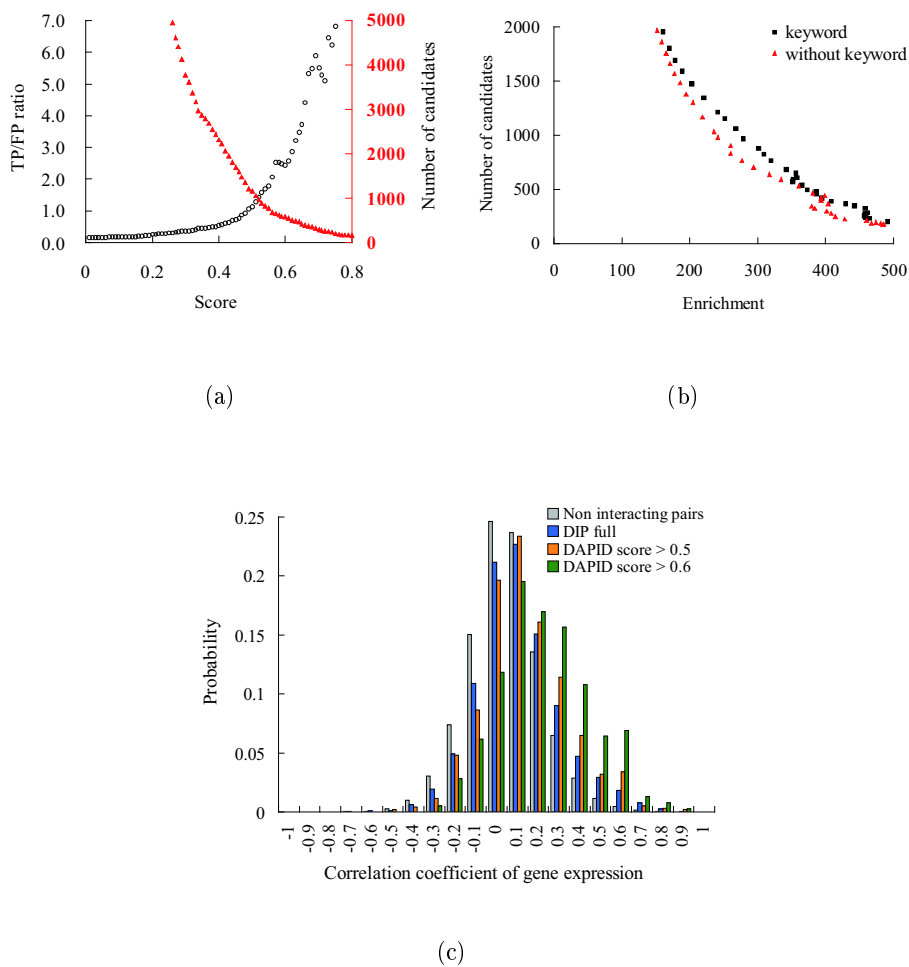


Figure 2: A summary analysis of the ability of our 3D-domain interologs in *S. cerevisiae*. (a) The TP/FP ratio (black) increases monotonically with the values of our scoring function. The TP/FP ratio is 1.126 and the number of candidates (red) is 1158 when the score is set to 0.5. (b) Comparison of the performance of our scoring function considering the score of keyword annotation (black) or excluding the keyword score (red). (c) Distributions of the correlation coefficients of gene expression profiles for four interacting protein sets: our predicted protein-pair sets with thresholds 0.5 (orange) and 0.6 (green), the DIP protein-pair set (blue), and the non-interacting protein-pair set (gray). The correlations of our predicted data sets are much higher than ones of other two data sets.

As the interactions in *S. cerevisiae* are the most extensive, reliable, and well studied, we measured the quality of our predicted interactions in *S. cerevisiae*. 14779 protein-protein interactions in the DIP database were used as the positive set and 2599785 non-interacting protein pairs defined by Jansen *et al.* [17] as the negatives if we discarded the interacting proteins which are not in the Swiss-Prot database. The total number of possible interacting pairs is ~18 millions made by ~6000 proteins in *S. cerevisiae* [17]. Therefore, A and T are 14779 and 18000000, respectively. We can yield a set of predicted protein-protein interactions by setting the threshold of our joint scoring function. For example, A_h is 215, F_h is 191, and T_h is 1158 when the threshold is 0.5. In this case, the TP/FP ratio is 1.126 and the enrichment is 226. The TP/FP ratios (black circles in Figure 2(a)) increase monotonically with the scores of our joint scoring function ($I_{A'B'}$), confirming $I_{A'B'}$ as an appropriate measure of a real interaction. Protein pairs with $I_{A'B'} > 0.5$ have a better than 50% chance (TP/FP ratio > 1) of being in a physical interaction. We suggest $I_{A'B'}=0.5$ as an appropriate threshold which is used throughout our analysis and to select 101483 protein pairs from 845041 protein pairs in the DAPID. Figure 2(b) shows that our method is better than the random pairing method about 226 times due to the number of protein-protein interaction is 1158. In addition, Figure 2(b) shows that our keyword annotation score (Equation 3) is a useful strategy to improve performance and enlarge the number of predicted protein-protein interactions.

Furthermore, the gene expression profiles of two interacting proteins were also used to access the accuracy of our method (Figure 2(c)) according to the basic assumption, that is, the gene pair with similar expression profiles is likely to encode an interacting protein pair [17]. We used the gene expression data of Hughes *et al.* [15] for *S. cerevisiae* to calculate the correlation coefficient of two interacting proteins. Figure 2(c) shows the distributions of the correlation coefficients for four interacting protein sets, including our predicted protein pairs in *S. cerevisiae* with thresholds 0.5 (1158) and 0.6 (575), protein pairs in the DIP database (14779), and the non-interaction protein pairs (2599785). The correlations of our predicted data sets are much higher than ones of other two data sets. By standard two sample T-test, the mean of correlation coefficients for our predicted protein pairs (threshold=0.5) is significantly higher than that for the negative set (p value $< \sim 10^{-50}$).

DAPID features and examples

Queries of the DAPID can be made using UniProt accession number, gene name, or keyword. An example (gene name is *bmp2* and its UniProt number is P12643) of the DAPID search and usage is shown in Figure 3. BMP2 induces bone regeneration and ectopic bone formation in adult vertebrates [26]. Our method predicted 10 interacting-protein partners of human BMP2, including 3 partners recorded in the DIP database and seven novel ones. Figure 3(b) shows the query results comprising interaction analysis, 3D-interacting domains, and some basic information of the interacting partners. In the interaction-analysis field, the DAPID indicates the sources of protein-protein interactions from 3D-domain interologs, the DIP database, and the PDB. The numbers in the parentheses represent the numbers of interaction evidences in the respective databases. If a protein-protein interaction is inferred from our 3D-domain interologs, this interaction is colored green (the DAPID score > 0.6) or orange.

The DAPID provides 3D interacting domains (Figure 3(c)) and contact residues (Figure 3(d)) for visualizing molecular details between the query protein and its interacting partner. For instance, the two interacting proteins, human BMP2 and ACVR2 (UniProt number is P27037), are predicted from the 3D protein complex (PDB code 1lx5 with chains A and B). This interaction is also recorded in the DIP database. The DAPID shows evidences about this interaction, such as the sources, references, the DAPID score, and the 3D-interacting domain pair (i.e., TGF_beta domain and Activin_recpt domain). The contact residues between chains A and B of the complex are shown in the SPACEFILL model (Figure 3(c)). The DAPID used PSI-BLAST to align the protein sequences of BMP2 and ACVR2 to their corresponding templates (i.e., 1lx5:A and 1lx5:B). The aligned contact residues are colored green



Figure 3: Screenshots of a DAPID query example. The query gene name is *bmp2* and its UniProt number is P12643 in this example. (a) Three ways to query in the DAPID: UniProt accession number, gene name, or keywords. (b) Interacting partners of BMP2. (c) The interaction between BMP2 and ACVR2 (UniProt number P27037) inferred from this 3D protein complex (PDB code 1lx5 with chains A and B). The 3D-interacting domains are TGF_beta (BMP2 with green) and Activin_recp (ACVR2 with yellow), and the contact residues are shown in the SPACEFILL model. (d) Sequence alignment results of aligning interactive proteins (BMP2 and ACVR2) into their corresponding structural templates (1lx5:A and 1lx5:B). Residues in the interaction domains are indicated as upper case, aligned contact residues are colored green (BMP2) or yellow (ACVR2), and the identical contact residues are indicated as bold. Four important residues, 316A, 370S, 372L, and 382L for BMP2 interacting with ACVR2 [18], are labeled.

or yellow and the identical contact residues are indicated as bold (Figure 3(d)). Kirsch *et al.* [18] identified four residues, 316A, 370S, 372L and 382L, which are important for BMP2 to interact with ACVR2. The DAPID can correctly identify these four residues (Figure 3(d)).

The COPII coat which consists of Sec23p/Sec24p, Sec13p/Sec31p and a small GTPase is required for vesicle budding from the endoplasmic reticulum [28]. Our method yielded eight interacting partners of the protein Sec23p (UniProt number is P15303) in *S. cerevisiae*. Three interacting partners (Sec24p, Sfb2p, and Sfb3p) are recorded in the DIP database, one partner (Sar1p) forms 3D complex with Sec23p in the PDB, one partner (Sec23p) is found in the MIPS database [22], and remaining proteins (Arf1p, Arf2p, and Arf3p) are novel interacting partners which consistently contain the Arf domain.

By analyzing the protein 3D complex (PDB code 1m2o with chains A and B), we observed that Sec23_trunk domain interacts with Arf domain. Previous works showed that ARF protein family and the protein Sec23 are implicated in vesicular transport pathway [7]. Therefore, Arf1p, Arf2p or Arf3p may be reasonable interacting to Sec23p based on protein functional views.

Discussion and future directions

It has been known that two proteins are homologous to another protein pair in a known complex does not necessarily interact with each other. We will enhance our scoring function by considering the empirical pair potentials derived from interfaces of known structures [2, 20] and other knowledge annotations (e.g. GO annotations [14] and other annotations in the Swiss-Prot database). In addition, the definitions of 3D-interacting domains extracted from some complexes may be different from the definitions of the Pfam domains. The next data release will cover other domain definitions, e.g. SMART [19] and ProDom [10]. Furthermore, we will integrate interaction data from some different public databases, e.g. DIP, BIND, and STRING, and our DAPID for studying the protein-protein networks.

In summary, the key novelty of the present work is that we utilize the 3D-domain interologs and joint scoring function incorporating knowledge annotations, sequence and structural similarity to predict physical protein-protein interactions from high-resolution structure complexes. Our DAPID is able to provide an atomic description of the interfaces ultimately responsible for interaction specificity and to visualize protein-protein interactions at the molecular level.

Acknowledgment

This work was supported by National Science Council and the University System at Taiwan-Veteran General Hospital Grant.

References

- [1] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Caverio, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Gardeman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F., and Hogue, C. W., The biomolecular interaction network database and related tools 2005 update, *Nucleic Acids Res.*, 33(Database issue):D418–D424, 2005.
- [2] Aloy, P. and Russell, R. B., Interrogating protein interaction networks through structural biology, *Proc. Natl. Acad. Sci. USA*, 99(9):5896–5901, 2002.
- [3] Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B., Structure-based assembly of protein complexes in yeast, *Science*, 303(5666):2026–2029, 2004.
- [4] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

- [5] Aung, Z. and Tan, K. L., Rapid 3D protein structure database searching using information retrieval techniques, *Bioinformatics*, 20(7):1045–1052, 2004.
- [6] Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S., The universal protein resource (UniProt), *Nucleic Acids Res.*, 33(Database issue):D154–D159, 2005.
- [7] Barlowe, C., Orci, L., Yeung, T., Hosobuchi, M., Hamamoto, S., Salama, N., Rexach, M. F., Ravazzola, M., Amherdt, M., and Schekman, R., COPII: a membrane coat formed by Sec proteins that drive vesicle budding from the endoplasmic reticulum, *Cell*, 77(6):895–907, 1994.
- [8] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R., The Pfam protein families database, *Nucleic Acids Res.*, 32(Database issue):138–141, 2004.
- [9] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, 31(1):365–370, 2003.
- [10] Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., and Kahn, D., The ProDom database of protein domain families: more emphasis on 3D, *Nucleic Acids Res.*, 33(Database issue):D212–D215, 2005.
- [11] Cary, M. P., Bader, G. D., and Sander, C., Pathway information for systems biology, *FEBS Lett.*, 579(8):1815–1820, 2005.
- [12] Deshpande, N., Address, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M., and Bourne, P. E., The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema, *Nucleic Acids Res.*, 33(Database issue):D233–D237, 2005.
- [13] Finn, R. D., Marshall, M., and Bateman, A., iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions, *Bioinformatics*, 21(3):410–412, 2005.
- [14] Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dummer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R., The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Res.*, 32(Database issue):D258–D261, 2004.
- [15] Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburttty, K., Simon, J., Bard, M., and Friend, S. H., Functional discovery via a compendium of expression profiles, *Cell*, 102(1):109–126, 2000.
- [16] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. of Sci. USA*, 98(8):4569–4574, 2001.

- [17] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M., A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, 302(5644):449–453, 2003.
- [18] Kirsch, T., Nickel, J., and Sebald, W., BMP-2 antagonists emerge from alterations in the low-affinity binding epitope for receptor BMPR-II, *EMBO Journal*, 19(13):3314–3324, 2000.
- [19] Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P., SMART 4.0: towards genomic data integration, *Nucleic Acids Res.*, 32(Database issue):D142–D144, 2004.
- [20] Lu, L., Arakaki, A. K., Lu, H., and Skolnick, J., Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome, *Genome Res.*, 13(6A):1146–1154, 2003.
- [21] Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M., Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs", *Genome Res.*, 11(12):2120–2126, 2001.
- [22] Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpfen, V., Warfsmann, J., and Ruepp, A., MIPS: analysis and annotation of proteins from whole genomes, *Nucleic Acids Res.*, 32(Database issue):D41–D44, 2004.
- [23] Pandey, A. and Mann, M., Proteomics to study genes and genomes, *Nature*, 405(6788):837–846, 2000.
- [24] Park, J., Lappe, M., and Teichmann, S. A., Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast, *J. Mol. Biol.*, 307(3):929–938, 2001.
- [25] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O., Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96(8):4285–4288, 1999.
- [26] Reddi, A. H., Bone morphogenetic proteins: an unconventional approach to isolation of first mammalian morphogens, *Cytokine & Growth Factor Rev.*, 8(1):11–20, 1997.
- [27] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D., The database of interacting proteins: 2004 update, *Nucleic Acids Res.*, 32(Database issue):D449–D451, 2004.
- [28] Schekman, R. and Orci, L., Coat proteins and vesicle budding, *Science*, 271(5255):1526–1533, 1996.
- [29] Stein, A., Russell, R. B., and Aloy, P., 3did: interacting protein domains of known three-dimensional structure, *Nucleic Acids Res.*, 33(Database issue):D413–D417, 2005.
- [30] Uetz, P. and Finley, R. L. J., From protein networks to biological systems, *FEBS Lett.*, 579(8):1821–1827, 2005.
- [31] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P., STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic Acids Res.*, 33(Database issue):D433–D437, 2005.
- [32] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P., Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, 417(6887):399–403, 2002.
- [33] Wojcik, J. and Schachter, V., Protein-protein interaction map inference using interacting domain profile pairs, *Bioinformatics*, 17(Suppl 1):S296–S305, 2001.