BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 12 2013, pages 1577–1579
doi:10.1093/bioinformatics/btt181

Databases and ontologies

Advance Access publication April 18, 2013

INstruct: a database of high-quality 3D structurally resolved protein interactome networks

Michael J. Meyer[1,2,3,†], Jishnu Das[1,2,†], Xiujuan Wang[1,2,†] and Haiyuan Yu[1,2,*]

[1]Department of Biological Statistics and Computational Biology and [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853 and [3]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY 10065, USA

Associate Editor: Mario Albrecht

ABSTRACT

Summary: INstruct is a database of high-quality, 3D, structurally resolved protein interactome networks in human and six model organisms. INstruct combines the scale of available high-quality binary protein interaction data with the specificity of atomic-resolution structural information derived from co-crystal evidence using a tested interaction interface inference method. Its web interface is designed to allow for flexible search based on standard and organism-specific protein and gene-naming conventions, visualization of protein architecture highlighting interaction interfaces and viewing and downloading custom 3D structurally resolved interactome datasets.

Availability: INstruct is freely available on the web at http://instruct.yulab.org with all major browsers supported.

Contact: haiyuan.yu@cornell.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 24, 2012; revised on March 31, 2013; accepted on April 15, 2013

1 INTRODUCTION

Protein–protein interactions demonstrate functional principles of biological processes because most proteins carry out their cellular functions by interacting with other proteins. The set of all protein interactions within an organism, known as the 'interactome', is often represented as a network (Pawson and Nash, 2000; Vidal, 2005). Interactome networks are powerful resources for biologists because they help elucidate the interconnected nature of signaling and communication within cellular systems. It has also been suggested that mechanistic explanations of many human diseases can be obtained by studying alterations to this network (Barabasi *et al*., 2011; Vidal *et al*., 2011). For example, a global guilt-by-association principle has been widely used to predict disease genes by dissecting molecular networks (Oliver, 2000).

However, for an interactome network to be successfully applied in biological studies, it is imperative that it incorporate the intricate structural details of proteins within the network and not simply treat the proteins as mathematical points in a graph-theoretic network (Schuster-Bockler and Bateman, 2008; Wang *et al*., 2012). As structure is the basis of protein function (Lahiry *et al*., 2010), elucidating structural details of interactions can help refine our current understanding of biochemical function from protein–protein interaction networks (Barabasi and Oltvai, 2004).

Here, we present INstruct (http://instruct.yulab.org), a comprehensive database of high-quality, 3D, structurally resolved protein interactome networks in human and six widely studied model organisms. To our knowledge, INstruct is the first online repository containing structurally resolved interaction interfaces between proteins for which no co-crystal structure is available. To accomplish this, we used an interaction interface inference method (Wang *et al*., 2012) to structurally resolve interactions based on 37 210 known co-crystal structures in the PDB (Berman *et al*., 2000). In total, INstruct currently contains 6585 human, 644 *Arabidopsis thaliana*, 120 *Caenorhabditis elegans*, 166 *Drosophila melanogaster, 119 Mus musculus,* 1273 *Saccharomyces cerevisiae* and 37 *Schizosaccharomyces pombe* structurally resolved interactions. As a comprehensive database providing structural details not previously annotated in protein interactome networks, INstruct will be an invaluable resource in a wide array of biological research.

2 METHODS

Binary protein–protein interaction data used to build INstruct was curated from eight major interaction databases—BioGrid (Stark *et al*., 2011), DIP (Salwinski *et al*., 2004), HPRD (Keshava Prasad *et al*., 2009), IntAct (Kerrien *et al*., 2012), iRefWeb (Turner *et al*., 2010), MINT (Licata *et al*., 2012), MIPS (Mewes *et al*., 2011) and VisAnt (Hu *et al*., 2009). Not all organisms included in INstruct derived interaction data from every database. These interactions were then filtered to meet strict high-confidence criteria (Das and Yu, 2012) resulting in 61 108 high-quality binary interactions for all seven organisms (Fig. 1 and Supplementary Note S1). It should be noted that none of the protein–protein interactions with co-crystal structures are filtered out of INstruct.

To add structural resolution to our high-quality binary interactome networks, we leveraged the information in several protein databases. Using protein domain definitions from Pfam (Punta *et al*., 2012), we identified 'Pfam-A' domains, which are both 'significant' and 'in-full' as defined by Pfam that also appear in proteins in our high-quality binary interactome networks. To determine the domains mediating the protein–protein interactions in our network, we gathered domain interaction data from 3did (Stein *et al*., 2009) and iPfam (Finn *et al*., 2005), which in turn derive their domain–domain interaction evidence from

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

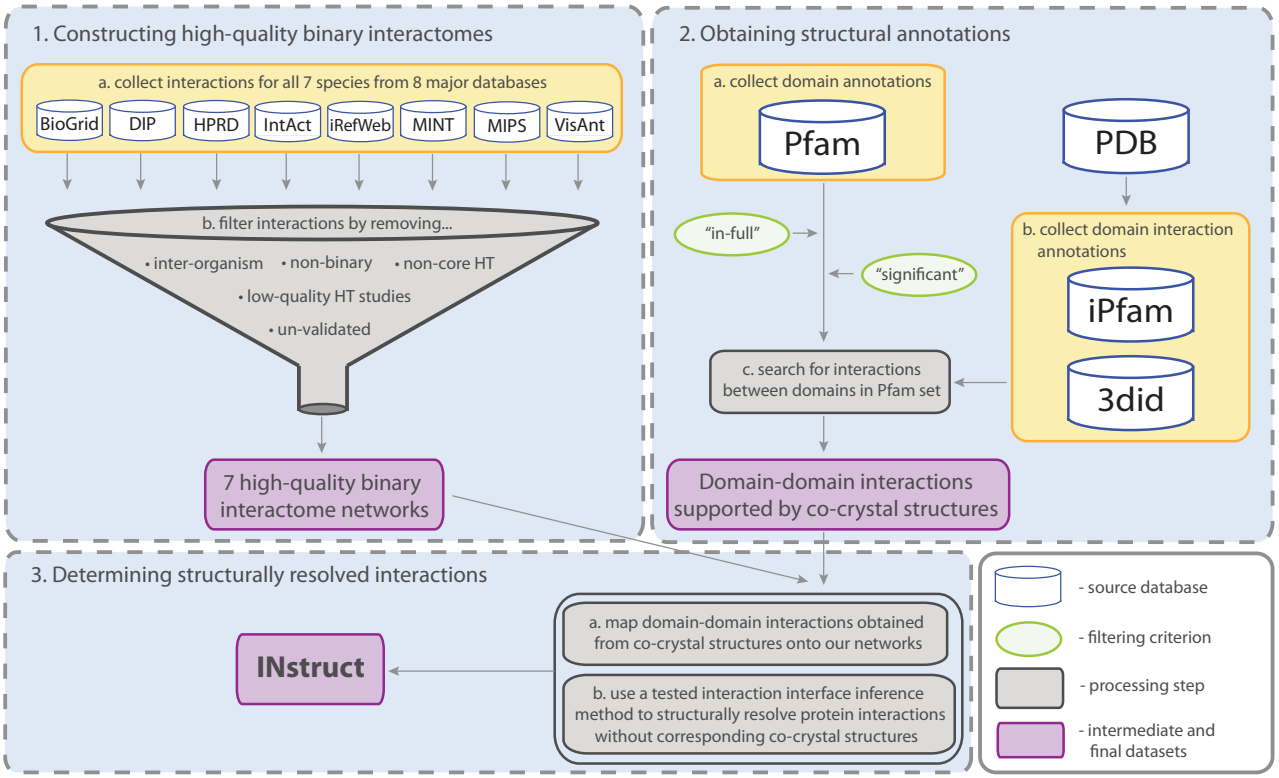

Fig. 1. A flow chart showing the sources and three stages of data processing used to create the 3D interactomes in INstruct. (1) Constructing high-quality binary interactomes. Interactomes for each of the seven organisms were created by collecting protein–protein interactions from each of the shown databases. We removed inter-organism interactions, non-binary interactions, interactions from high-throughput (HT) studies that are not a part of the author's high-confidence dataset (core), interactions from low-quality HT studies and unvalidated interactions. (2) Obtaining structural annotations. By collecting high-quality structural annotation data, we produced a set of domain–domain interactions supported by atomic-resolution co-crystal structures. (3) Determining structurally resolved interactions. We used a method of homology-based interaction interface inference to structurally resolve interaction interfaces for interacting proteins

37 210 existing 3D atomic-resolution co-crystal structures in the PDB. In all, 1708 protein–protein interactions in our binary interactomes are directly represented by one of these co-crystal structures, in which case it is straightforward to determine where the pair of proteins interacts.

For 7236 protein–protein interactions not supported by direct co-crystal evidence, we applied a tested interaction interface inference method (Wang *et al.*, 2012) to extend the scope of the interaction data provided by 3did and iPfam (Fig. 1). For these interactions, we predicted the interface domains based on co-crystal structures of homologous domains for one or both partners (Supplementary Note S2). Although 3did and iPfam indicate pairs of homologous domains that have been shown to interact in co-crystal structures of pairs of proteins, INstruct is the first source to predict that these domain–domain interactions facilitate protein–protein interactions for which no co-crystal structure exists.

Although we have demonstrated high confidence in the ability of our method to identify the domains at protein interaction interfaces, it is important to note the inherent difference in resolution available for interfaces determined directly from co-crystal evidence versus those that were inferred using homologous structures. Atomic-resolution information is only available for interactions with co-crystal structures, whereas interaction interfaces inferred from homology are resolved to the level of protein domains. To maintain uniformity, INstruct displays only domain-level information for all interactions. When available, atomic-resolution information is easily accessible through direct links to the PDB.

In total, these methods yielded 8944 3D protein–protein interactions with structurally resolved interaction interfaces. Full network statistics are available in Supplementary Table S1.

3 USAGE

A web-based interface is deployed for accessing these interactomes, which includes five basic features: (i) searching for proteins, (ii) retrieval of interaction data, (iii) visualization of protein domains, (iv) creation of custom downloadable datasets and (v) downloading of entire interactome datasets.

A protein in any organism can be queried using its Universal Protein Resource (UniProt) accession ID (UniProt Consortium, 2011) or its corresponding standard gene symbol. Additionally, each organism has a single searchable alternate identifier used by a popular organism-specific database. The standard query interface (shown in Supplementary Fig. S1) accepts multiple simultaneous queries using any combination of the three available identifiers for each organism.

Users are taken directly to the results page if one or more of their queries match an entry in INstruct. This page shows the results for each query linearly down the page in the order that they appeared in the query. Each matching query returns three types of output: (i) naming information, (ii) schematics showing

the domain-level interactions and (iii) sortable tables providing information about domain–domain interactions. The sidebar always contains the search box, for refinement of search terms, and a dialog for downloading the complete set of interactions that match all terms in the query.

For each protein that interacts with a query protein, a schematic is shown, displaying the domain architecture of both proteins side-by-side as linear models according the order of their amino acids (an example is shown in Supplementary Fig. S1). Between each pair of proteins, domains that interact are indicated in green and those that do not are indicated in gray. Network edges are drawn between domains that interact on the two proteins, with edges colored in red indicating domain interactions derived directly from co-crystal evidence, and edges in gray indicating domain–domain interactions inferred by homology. Regardless of whether a domain is involved in a structurally resolved interaction, all domain information is interactive and linked to further information provided on the Pfam website.

Interactions involving the queried protein and each of its interacting partners are shown in a table below each schematic and can be sorted by domain on either protein, on the number of publications supporting this interaction, or on the number of supporting PDB structures. PDB structures that provide direct co-crystal evidence for the indicated domain–domain interaction are shown in red, and all other PDB structures provide homology-based evidence and are shown in blue. External references, including Pfam links to each domain in the interacting pair, publication information from PubMed, and PDB structures supporting each interaction, can be accessed by clicking on their corresponding entry in the table.

Additionally, each search produces a tab-delimited text file containing all relevant information about the interactions as shown on the results page, including all available naming conventions and amino acid locations of the interacting domains. A custom dataset can be created simply by searching for up to five proteins at once, separating each term with a delimiter in the search field, or by uploading a file of identifiers on the downloads page. All valid search terms will be added to the downloadable file and displayed on the results page.

When available, the domain–domain interaction pairs that facilitate the shown protein interactions on the results page will appear in the sidebar on the right side of the results page. Underneath each domain–domain pair is given a list of proteins that interact via the same domain–domain pair. Proteins in red provide the direct co-crystal evidence for the given domain–domain interaction. In other words, a co-crystal containing each protein in red proves the existence of the domain–domain interaction in the first place and allows us to resolve other protein–protein interactions to include this domain–domain interaction.

4 CONCLUSION

INstruct will be a useful tool in a broad spectrum of fundamental biological research. With the continued growth of data sources, especially available co-crystal structures in the PDB, we expect the coverage of our structurally resolved interactome networks to increase over time (Chandonia and Brenner, 2006). We plan to update INstruct at least once per year to incorporate newly available information and to expand our repertoire of model organisms. As more structural data become available, we will build more comprehensive 3D, structurally resolved interactome networks for different organisms.

Funding: Tri-Institutional Training Program in Computational Biology and Medicine [via NIH training grant 1T32GM083937 to M.J.M.]; National Institute of General Medical Sciences [R01 GM097358 and R01 CA167824 to H.Y.]; Tata Graduate Fellowship (to J.D.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences.

Conflict of Interest: none declared.

REFERENCES

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Barabasi,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.

Das,J. and Yu,H. (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.*, **6**, 92.

Finn,R.D. *et al.* (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Hu,Z. *et al.* (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.*, **37**, W115–W121.

Kerrien,S. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.

Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Lahiry,P. *et al.* (2010) Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.*, **11**, 60–74.

Licata,L. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.

Mewes,H.W. *et al.* (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.*, **39**, D220–D224.

Oliver,S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.

Pawson,T. and Nash,P. (2000) Protein-protein interactions define specificity in signal transduction. *Genes Dev.*, **14**, 1027–1047.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Schuster-Bockler,B. and Bateman,A. (2008) Protein interactions in human genetic diseases. *Genome Biol*, **9**, R9.

Stark,C. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.

Stein,A. *et al.* (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.

Turner,B. *et al.* (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)*, **2010**, baq023.

UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.

Vidal,M. (2005) Interactome modeling. *FEBS Lett.*, **579**, 1834–1838.

Vidal,M. *et al.* (2011) Interactome networks and human disease. *Cell*, **144**, 986–998.

Wang,X. *et al.* (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–164.