BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 20 2013, pages 2647–2648
doi:10.1093/bioinformatics/btt451

Sequence analysis

Advance Access publication August 5, 2013

Pclust: protein network visualization highlighting experimental data

Wenlin Li[1], Lisa N. Kinch[2] and Nick V. Grishin[1,2,*]

[1]Departments of Biophysics and Biochemistry and [2]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050, USA

Associate Editor: John Hancock

ABSTRACT

Summary: One approach to infer functions of new proteins from their homologs utilizes visualization of an all-against-all pairwise similarity network (A2ApsN) that exploits the speed of BLAST and avoids the complexity of multiple sequence alignment. However, identifying functions of the protein clusters in A2ApsN is never trivial, due to a lack of linking characterized proteins to their relevant information in current software packages. Given the database errors introduced by automatic annotation transfer, functional deduction should be made from proteins with experimental studies, i.e. 'reference proteins'. Here, we present a web server, termed Pclust, which provides a user-friendly interface to visualize the A2ApsN, placing emphasis on such 'reference proteins' and providing access to their full information in source databases, e.g. articles in PubMed. The identification of 'reference proteins' and the ease of cross-database linkage will facilitate understanding the functions of protein clusters in the network, thus promoting interpretation of proteins of interest.

Availability: The Pclust server is freely available at http://prodata.swmed.edu/pclust

Contact: grishin@chop.swmed.edu

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

Received on June 24, 2013; revised on July 20, 2013; accepted on July 31, 2013

1 INTRODUCTION

A common practice to formulate hypotheses for a protein of unknown function includes searching for annotations among homologous proteins. Although sequence similarity does not necessarily correlate with functional similarity (Clark and Radivojac, 2011), all-against-all pairwise similarity network (A2ApsN) works best to illustrate functional relationships among large numbers of proteins (Atkinson *et al.*, 2009), and meanwhile avoids computational complexity and problems of aligning non-homologous sequences (Frickey and Lupas, 2004). Software packages, such as CLANS (Frickey and Lupas, 2004), Pythoscape (Barber and Babbitt, 2012) and Cytoscape (Shannon *et al.*, 2003), provide powerful repositories to manage the A2ApsN. However, they require either programming basics or expertise in program setups to generate the network (detailed comparison in the Supplementary Material). Numerous efforts aim to visualize the protein–protein interaction (PPI) network (Agapito *et al.*, 2013). But these packages build the network by PPI data (not by sequence similarity) and assign functions by analysing the network structure, such as dissecting functional modules (Sharan *et al.*, 2007). Given the high misannotation rate in current databases (Schnoes *et al.*, 2009), the simple protein descriptions that current packages offer are somewhat suspect, which hinders the understanding of the protein clusters. Thus, to avoid working with a network of uncertainty, one has to tediously verify the functions of nodes in the network before getting into interesting biology.

Here, we developed a web server named Pclust for visualization of the A2ApsN, which emphasizes those 'reference proteins' with experimental studies. Pclust works with the Seq2Ref server (Li *et al.*, 2013) to identify the 'reference proteins' and highlight them in the network. The web interface bypasses the pain of software installation and the requirement of programming expertise. The highlighted 'reference proteins' and easy access to their functional studies simplify the process of relating functions to protein clusters, thus facilitating hypothesis driven research of proteins of interest.

2 METHODS AND IMPLEMENTATION

2.1 Preparation of the protein network

According to the type of user inputs, the protein sets shown in the A2ApsN are taken from (i) Seq2Ref BLAST results; (ii) user input sequences; or (iii) user customized data (e.g. http://prodata.swmed.edu/pclust/help/format.html#custom). If a single sequence is given, proteins will be taken from its BLAST result against NR. To speed up A2ApsN generation, CD-HIT (Fu *et al.*, 2012) (optional, default identity cutoff: 95%) reduces the redundancy of the protein set that is used for all-against-all BLAST clustering.

2.2 Reference protein detection

Proteins either detected by BLAST or input by the user are submitted to the Seq2Ref server to detect 'reference proteins'. As the user input format is flexible, we submit protein sequences to the Protein Identifier Cross-Reference (PICR, Wein *et al.*, 2012) service to detect their IDs in PDB, Swiss-Prot and RefSeq databases.

2.3 Protein network generation

A2ApsN is calculated with force-directed graph drawing algorithms implemented by Vivagraph (https://github.com/anvaka/VivaGraphJS) and rendered using WebGL library (supported by most browsers). Reference proteins are colored according to the data sources. Keyword search of the annotations, adjustment for the link number and on-the-fly reference panels describing functional studies are implemented by asynchronous request to our server through AJAX (Asynchronous JavaScript and XML).

3 RESULTS

Pclust has four modes; a user can specify the input as (i) a single sequence; (ii) multiple sequences; (iii) a Seq2Ref result link; and

*To whom correspondence should be addressed.

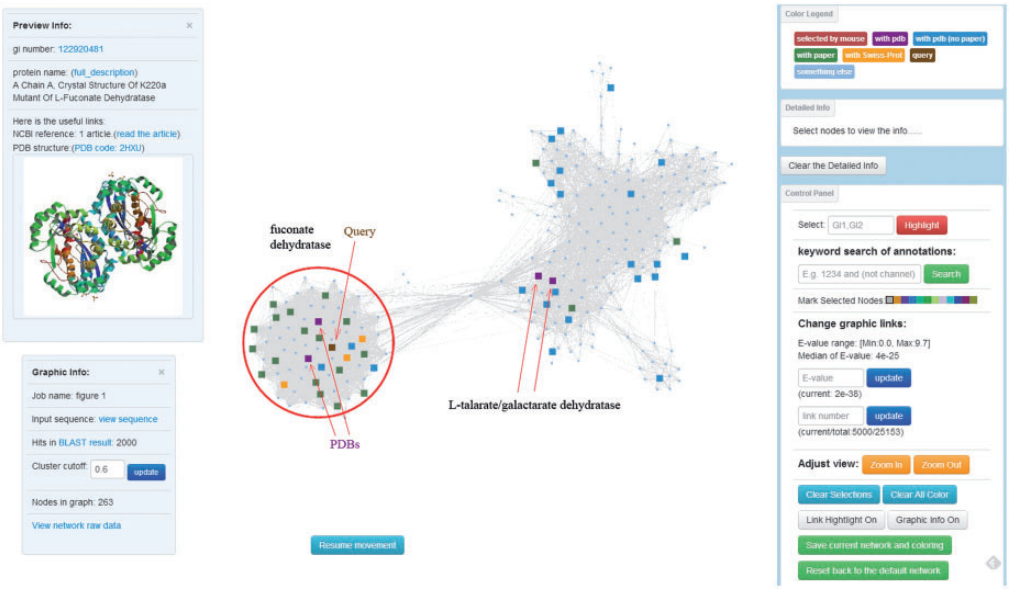

Fig. 1. Snapshot of an A2ApsN from Pclust (web link: http://prodata.swmed.edu/wenlin/server/paper_data/pclust/fig1). Protein nodes are colored, hierarchically, brown (input by the user, if applicable), red (selected by mouse), purple (with PDB structures and their articles), green (with PubMed articles), yellow (with Swiss-Prot functional comments), blue (with PDB structures of no article, such as those from structural genomics) and light blue (without any reference, smaller size). A 'preview' panel and a 'detailed Info' panel appear for the protein on which your mouse hovers and your mouse clicks, respectively. Batch selection of proteins is available by inputting the gi numbers (if applicable) separated by commas, and keyword search is available for annotations from the NCBI protein database. By default, Pclust will include the first quarter or 5000 (whichever is smaller) network links (ordered by E-value) and report the corresponding E-value cutoff; a panel to adjust network links by varying the E-value cutoff or the link number is also available. To know more about the control panel, please refer to: http://youtu.be/XLkfEg2jGOc

(iv) customized network data. (The input interface is shown in Supplementary Fig. S1.) The flow chart for each mode is available in Supplementary Figure S2–S5. The four above modes are designed to provide A2ApsNs (i) for any single sequence; (ii) with an advanced interface similar to CLANS; (iii) for previously generated Seq2Ref jobs; and (iv) with a customized input that grants users the flexibility to design. An email is required to keep track of the job submission. Once the network is ready, an email containing the result link will be sent to the provided address.

Figure 1 (current E-value cutoff: 2e-38) shows the interface for an A2ApsN, as well as an example where the input protein (brown node), annotated as 'mandelate racemase' (gi|17987990), should be a 'fuconate dehydratase', as previously described (Schnoes *et al.*, 2009). (Another practical case is available in the Supplementary Material.) Merely reading the brief protein descriptions within the cluster, such as 'RTS beta protein', 'mandelate racemase' and 'enolase superfamily member' (as in the CLANS interface), results in confusion about the function of the cluster. Pclust alleviates this confusion by highlighting the reference proteins and referring to their annotation sources linked by our server (details available in the Supplementary Material). For example, the cluster circled in Figure 1 containing the questionable 'mandelate racemase' (brown) includes two solved crystal structures of known fuconate dehydratases with provided links to their experimental data. Thus, with the convenience of locating reference proteins in A2ApsN and accessing their database links, more accurate hypotheses about the function of protein queries can be generated, potentiating biological discovery.

Funding: This work was supported by National Institutes of Health (GM094575 to N.V.G.) and the Welch Foundation (I-1505 to N.V.G.).

Conflict of Interest: none declared.

REFERENCES

Agapito,G. *et al.* (2013) Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics*, **14** (**Suppl. 1**), S1.

Atkinson,H.J. *et al.* (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*, **4**, e4345.

Barber,A.E. and Babbitt,P.C. (2012) Pythoscape: a framework for generation of large protein similarity networks. *Bioinformatics*, **28**, 2845–2846.

Clark,W.T. and Radivojac,P. (2011) Analysis of protein function and its prediction from amino acid sequence. *Proteins*, **79**, 2086–2096.

Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Li,W. *et al.* (2013) Seq2Ref: a web server to facilitate functional interpretation. *BMC Bioinformatics*, **14**, 30.

Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.

Wein,S.P. *et al.* (2012) Improvements in the Protein Identifier Cross-Reference service. *Nucleic Acids Res.*, **40**, W276–W280.