PINALOG: IDENTIFICATION OF PROTEIN COMPLEXES AND PREDICTION OF FUNCTION FROM THE ALIGNMENT OF PROTEIN INTERACTION NETWORKS

Michael J E Sternberg & *Hang T T Phan*

Structural Bioinformatics Group, Division of Molecular Biosciences, Imperial College London

m.sternberg@imperial.ac.uk

PINALOG is a novel approach to align two protein-protein interaction (PPI) networks which combines information available for the proteins in the networks, including sequence, function and network topology. Alignment of human and yeast PINs from the IntAct database reveals several related complexes between the two species that participate in similar biological processes. The power of function prediction based on the resulting alignment by direct transfer functions of mapped proteins was assessed. With a test set composing of proteins with low PSI-BLAST sequence percent identity, cross validation demonstrated the improved performance of function prediction made by PINALOG over PSI-BLAST, where we obtained much better recall at a similar level of precision.

INTRODUCTION

Interactions of proteins play important roles in biological processes. Although the interaction network of proteins in one species is different from that of others, there are components that perform similar function in the cells, which are likely to be conserved across species. Comparison of the protein interaction networks from different species, therefore, yields understanding of the evolution of species, as well as provides a means to predict protein function and conserved components. Several algorithms are available to align networks e.g. NetworkBlast[1], Graemlin[2], and IsoRank[3]. Here we introduce PINALOG which includes protein function in the global alignment. This facilitates the identification of conserved complexes between two species and the prediction of protein function for proteins not amenable to functional transfer by homology.

METHODS

Figure 1 summarises the algorithm of PINALOG which aligns two input PPI networks.

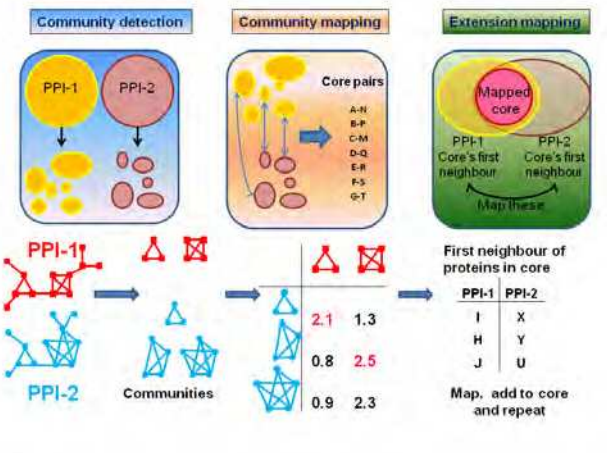

Figure 1. Method to align two protein interaction networks (PPI networks)

Step 1 – Community detection. The algorithm starts with the independent detection of communities, dense parts – likely protein complexes- in the networks, in each PPI network. As protein complexes are the key functional modules for molecular biological processes, they are often conserved among species. The use of communities as the seed for the alignment is a major new contribution in our method. CFinder using Clique Percolation Method was used to identify communities.

Step 2 – Community mapping. An initial alignment of the communities between species is obtained using function and sequence similarities of proteins. This is performed by Hungarian method[5] where communities are mapped by maximizing the total score of mapping communities. The result of this step is a list of mapped protein pairs with high similarity in sequence or function or both.

Step 3 – Extension mapping. The initial community mappings are extended to first-neighbouring proteins of those proteins already mapped. In this process, the information of network topology is included into the protein similarity to increase the possibility of the alignment of neighbouring protein pairs, which hence boosts the number of conserved interactions in the final alignment.

RESULTS & CONCLUSIONS

Our results on the alignment of human and yeast network from IntAct database revealed a large conserved network, 3,319 conserved edges, as opposed to 717 by IsoRank. The subnetworks in the conserved networks in the two species are components of similar biological processes such as the proteasome or transcription related processes.

The clusters of interacting proteins obtained from the human PPI networks are compared with human core protein complexes obtained from MIPS CORUM. These clusters in human are compared with the equivalent yeast protein complexes retrieved from MIPS CYDG. We found the agreement in function of clusters between the two species. For example, one cluster of human conserved network contains several complexes related to proteasome such as PA700 complex (20 proteins), 26S proteasome (22), 20S proteasome (14). The corresponding complex in the yeast conserved network contains 20S proteasome (14), 19/22S regulator (18) and m-AAA protease (2) complexes, all related to proteasome.

The results provide an approach to perform function prediction of proteins in difficult zone where no sequence similarity is detected. A cross-validation to test the power of function prediction in the difficult zone showed that in the GO Biological Process, PINALOG has a better recall rate of 14% in comparison with 7% of PSI-BLAST at the similar level of precision, ~30%. We made prediction for 15 unannotated human proteins from the alignment of human and yeast interactomes.

REFERENCES

1. Sharan, R. et al. *Proc Natl Acad Sci U S A* **102**, 1974-1979 (2005).
2. Flannick, J., Novak, A., Do, C.B., Srinivasan, B.S. & Batzoglou, S. *J Comput Biol* **16**, 1001-1022 (2009).
3. Singh, R., Xu, J. & Berger, B. *Proc Natl Acad Sci U S A* **105**, 12763-12768 (2008.)
4. Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. *Nature* **435**, 814-818 (2005).
5. Kuhn, H.W. *Naval Research Logistics (NRL)*, **52** 7-21 (2005).