½ general pfp background
relate pfp problem to networks
relate pfp problem and phenotype ?

In recent years, the numerous large scale sequencing projects combined with fast sequencing techniques have generated enormous amounts of sequence data. This has led to the identification of thousands of previously unseen genes. A fundamental goal is therefore to identify the function of uncharacterized genes on a genomic scale. It is difficult to design functional assays for uncharacterized genes so a major current challenge in bioinformatics is to devise algorithmic methods that, given a gene, can predict a hypothesis for its function that can then be validated experimentally.

As gene databases grow in size the diversity among the sequences increases and homology based methods become less effective [1]. Luckily, great amounts of data are now available, which offer clues about protein function, such as gene expression, metabolite expression, protein-protein interaction and these could be exploited for improving the predictions. Many of these types of data can have a natural representation as networks and therefore recently scientists have focused on developing methods that make use of the topology of these networks for functional inference.

This approach has been pioneered in a most interesting work by Marcotte et al.[2], in which the authors built a network in which each node corresponded to a protein in the *S. cerevisiae* genome, and the links between two proteins represented correlated evolution (through phylogenetic profiles), patterns of domain fusion, co-expression and protein-protein interaction. Treating these links as independent, their method consisted in assigning to an uncharacterized protein the function shared by the proteins it was connected to. Since this work appeared, other approaches have been developed, which use networks topologies to infer functional annotation. Most of them use networks built from protein-protein interaction (PPI) data and they could be broadly divided into two categories. A first group of methods breaks the networks into modules and then identifies the function of an unknown protein based on the function of the known member in its module [3], [4], [5], to name a few. A second group of methods, more related to the approach by Marcotte described earlier, assign a function to a protein by directly considering the function of its neighbours (e.g. [6], [7], [8], [9]).

REFERENCES

1. Friedberg, I., *Automated protein function prediction--the genomic challenge.* Brief Bioinform, 2006. **7**(3): p. 225-42.
2. Marcotte, E.M., et al., *A combined algorithm for genome-wide prediction of protein function.* Nature, 1999. **402**(6757): p. 83-6.
3. Rives, A.W. and T. Galitski, *Modular organization of cellular networks.* Proc Natl Acad Sci U S A, 2003. **100**(3): p. 1128-33.
4. Arnau, V., S. Mars, and I. Marin, *Iterative cluster analysis of protein interaction data.* Bioinformatics, 2005. **21**(3): p. 364-78.
5. Spirin, V. and L.A. Mirny, *Protein complexes and functional modules in molecular networks.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12123-8.
6. Vazquez, A., et al., *Global protein function prediction from protein-protein interaction networks.* Nat Biotechnol, 2003. **21**(6): p. 697-700.

7. Deng, M., et al., *Prediction of protein function using protein-protein interaction data.* J Comput Biol, 2003. **10**(6): p. 947-60.
8. Letovsky, S. and S. Kasif, *Predicting protein function from protein/protein interaction data: a probabilistic approach.* Bioinformatics, 2003. **19 Suppl 1**: p. i197-204.
9. Nabieva, E., et al., *Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps.* Bioinformatics, 2005. **21 Suppl 1**: p. i302-10.