

# Proteome Analysis Tools

Getiria Onsongo

*University of Minnesota, Department of Computer Science*

*200 Union Street SE, Minneapolis, MN, 55455 USA*

## Introduction

This is a short literature review on *proteome analysis tools*. It will only cover software tools used for proteome analysis. Instruments such as mass spectrometry and related techniques will not be covered.

Proteome describes the set of proteins encoded by a genome (Wilkins et al., ). Proteomics is the study of a genomes proteome and as (Tyers and Mann, 2003) points out, it encompasses most of the post-genomic analysis such as interactions between proteins and the structural descriptions of the proteins. Proteomics is a relatively new field with limited bioinformatics tools to support it. According to (White et al., 2006), as of March 2006, *ProMat* was the only open-source tool designed to support analysis of protein microarrays. Gene microarrays on the other hand have been around longer and the technology is more developed. There are also a myriad of tools available to support gene microarray analysis (Doniger et al., 2003), (Raychaudhuri et al., 2001), (Gollub et al., 2003). DNA synthesizes RNA which synthesizes proteins. Since gene microarrays can measure gene expressions levels, and given the maturity of the available technology for doing gene expression analysis, one can question the need proteome analysis. (Haynes et al., 1998) give three good reasons for doing proteome analysis. (i) mRNA expression levels do not always predict protein expression levels, (ii) proteins are dynamically modified in ways not apparent from gene sequence, and (iii) proteomes are dynamic and reflect the state of a biological system.

For the purpose of this review, proteome analysis tools will be roughly categorized into four groups. a) Tools for predicting protein interactions and protein function, b) Tools for comparing proteomes c) Tools used to analyze protein microarrays and d) Tools used for exploring and navigating proteome networks.

Below we present a few of the tools in the different categories.

## Tools for predicting protein interactions and protein functions

Most of the available software tools for doing proteome analysis attempt to predict protein interaction on a proteome scale. (Bock and Gough, 2003) extend a previously described data mining approach ( (Bock and Gough, 2001)) to prediction of protein-protein interactions on a proteome-wide scale. Using computational statistical learning theory, an analogy between the proteomes of two closely related organisms is used to predict protein-protein interactions. (Bock and Gough, 2003) use support vector machine to extrapolate from a protein interaction map in one organism to a complete map in a related organism. This approach is similar to doing sequence analysis to annotate proteins of unknown function but uses support vector machines because unlike in sequences where similarity is the key feature, in protein-protein interactions networks multiple features need to be examined. They introduce the *phylogenetic bootstrap* algorithm. Bootstrap is a numerical technique used for estimating standard errors of arbitrary test statistics (Efron and Gong, 1983). After extrapolation, predicted interaction can be used to design experiments. The authors show how a predicted protein-protein interaction network can be used to infer the function of the hypothetical protein Q9PMG7.

(Alexeyenko et al., 2006) present a tool, *MultiParanoid* for identification of orthologous groups shared by multiple proteomes. The program is automated and in addition to finding orthologs, it also finds inparalogs between pairs of proteomes. The program is free and available as a standalone program. One caveat to its application is it was designed to handle multiple proteomes that diverged at roughly the same time point during evolution. If species that diverged at different times are compared e.g yeast, human and chimpanzees, there will be no last common ancestral node in the phylogenetic tree. As a result, clusters will often contain human and chimpanzee outparalogs which are considered inparalogs relative to yeast (Alexeyenko et al., 2006). The term inparalogs indicate paralogs that arise through a gene duplication event after speciation. Outparalogs arise following a gene duplication preceding speciation. Because outparalogs are expected to have a diversified function relative to inparalogs, it is important to distinguish between the

two (is inparanoid, 2006). This application extends inparanoid (Remm et al., 2001) to multiple species while retaining the advantages of inparanoid.

(Nabieva et al., 2005) exploit the underlying structure of protein interaction maps to predict protein function. This is done through the use of a network-flow based algorithm, *FunctionalFlow*. Their approach relies on the annotations of the neighbors of a protein in a network to do the classification. They state this approach is useful when analyzing largely uncharacterized proteomes.

(Nicodème et al., 2002) present interesting results of an analysis on proteomes worth mentioning. They demonstrate that statistically over-represented or under-represented motifs in complete proteomes could be indicators of function. It is not surprising that over-represented motifs can indicate functionality. However, their assertion that under-represented motifs can indicate function bring new light to functional analysis. They give the Arg-Gly-Asp cell attachment motif (PS00016) as an example. The motif occurs in fibronectin and is important for its interaction with cell surface receptor (Ruoslahti and Pierschbacher, 1986). (Nicodème et al., 2002) point out that while there is a trend to avoid this motifs in the examined proteomes, this motif is present in snake disintegrins. Often when annotating proteins of unknown function, high similarity matches to protein of known function are used to infer function. It therefore expected that over-represented motifs can indicate importance in a proteome. This observation however points out the importance of paying attention to both extremes, over and under-represented motifs.

## Tools for comparing proteomes

This section presents tools used for comparing proteome. The European Bioinformatics Institute has developed a proteome analysis database (Apweiler et al., 2001). The database has a program that can be used to do comparisons between proteomes. Comparison that can be made includes computing a list of shared InterPro entries that are common to all selected proteomes (Kanapin et al., 2002). The comparison analysis data is presented in two different ways, as static or dynamic pages. The authors state the static pages contain data about a few of the most obvious proteome comparisons. An attempt to find out what

these obvious comparisons are failed as the link to the tool was consistently down.

The tool by (Nabieva et al., 2005) exploits the underlying structure of protein interaction maps to predict function. One would thus expect the existence of tools to support the comparison of two different proteome interaction networks to identify similar portions of the networks. None of the tools encountered so far provide this functionality. Protein interactions are not the first application to utilize networks as the data structure. Gene interaction networks and metabolic pathways are other such applications. In some of these other applications, tools already exist for comparing structural properties. These tools can potentially be used to compare two protein interaction networks. (Ogata et al., 2000) for example give an algorithm for comparing graphs that can be used to find local similarities between two graphs.

## Tools for analyzing protein microarrays

Protein microarrays are a recent technology and as a result, there are not many freely available tools for supporting their analysis. As of March 2006, *ProMat* was the only open-source tool designed for protein microarrays. ProMat is a tool for statistically analyzing data from enzyme-linked immunosorbent assay microarray experiments. It estimates sample protein concentrations and their uncertainties (White et al., 2006).

(Sundaresh et al., 2006) present software programs in *R* for analyzing protein microarrays using DNA microarray data techniques. In both DNA and protein microarrays, the goal is to effectively differentiate light signals derived from molecular binding events. (Sundaresh et al., 2006) explore the possibility of using existing DNA microarray techniques to automate the identification of humoral immune system. They show that with the appropriate modifications, DNA microarray analysis techniques can be used on protein microarrays. Specifically, they show an existing technique that has been validated (Choe et al., 2005) can be used to identify significant bindings between antibodies and antigens.

## Tools used for exploring and navigating proteome networks

This final group contains visualization tools. Protein-protein interaction networks are heavily used for predicting function. It is therefore no surprise that tools have been developed to aid the exploration and analysis of the networks. (Mrowka, 2001) developed a java applet for visualizing protein-protein interactions based on function and neighbouring distances. (Chang et al., 2005) improved this java applet to enhance navigation. With the enhanced java applet, it is possible to search for exact or similar node matches based on how the nodes have been identified.

(Strong et al., 2003) describe a method for visualizing and interpreting genome-wide functional linkages. The method involves construction of function maps and their subsequent hierarchical clustering. This method was applied to the pathogenic bacterium *Mycobacterium tuberculosis* to assign cellular functions to previously uncharacterized proteins.

## Conclusion

In this short review, proteome analysis tools were examined and roughly categorized into four main groups. a) Tools for predicting protein interactions and protein function, b) Tools for comparing proteomes c) Tools used to analyze protein microarrays and d) Tools used for exploring and navigating proteome networks. This ad-hoc classification does not cover all the different proteome analysis tools and not all tools will fit into this categorization. It however provided a framework for examining proteome analysis tools. Proteomics is a relatively new field. As it develops, techniques will improve leading to more reliable data. Analysis tools should therefore improve with time. This could be a function of time as more and more resources are devoted to proteomics. However, as is the case with most learning algorithms, analysis tools using learning algorithms are expected to improve as more reliable data is generated.

## References

- Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22(14).
- Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M., et al. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29(1):37–40.
- Bock, J. and Gough, D. (2001). Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–460.
- Bock, J. and Gough, D. (2003). Whole-proteome interaction mining. *Bioinformatics*, 19(1):125–134.
- Chang, A., McDermott, J., and Samudrala, R. (2005). An enhanced Java graph applet interface for visualizing interactomes. *Bioinformatics*, 21(8):1741–1742.
- Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16.
- Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor, S., Conklin, B., et al. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7.
- Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, 37(1):36–48.
- Gollub, J., Ball, C., Binkley, G., Demeter, J., Finkelstein, D., Hebert, J., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J., et al. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Research*, 31(1):94–96.
- Haynes, P., Gygi, S., Figeys, D., and Aebersold, R. (1998). Proteome analysis: biological assay or data archive? *Electrophoresis*, 19(11):1862–71.
- is inparanoid, W. (November 06 2006). <http://inparanoid.cgb.ki.se/ehelp.html>. *Online resource*.

- Kanapin, A., Apweiler, R., Biswas, M., Fleischmann, W., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Oinn, T., et al. (2002). Interactive InterPro-based comparisons of proteins in whole genomes. *Bioinformatics*, 18(2):374–375.
- Mrowka, R. (2001). A Java applet for visualizing protein–protein interaction. *Bioinformatics*, 17(7):669–671.
- Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(1):i302–i310.
- Nicodème, P., Doerks, T., and Vingron, M. (2002). Proteome analysis based on motif statistics. *Bioinformatics*, 18(Supplement 1):S161–S171.
- Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28(20):4021–4028.
- Raychaudhuri, S., Sutphin, P., Chang, J., and Altman, R. (2001). Basic microarray analysis: grouping and feature reduction. *Trends Biotechnol*, 19(5):189–193.
- Remm, M., Storm, C., and Sonnhammer, E. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*, 314(5).
- Ruoslahti, E. and Pierschbacher, M. (1986). Arg-Gly-Asp: a versatile cell recognition signal. *Cell*, 44(4):517–8.
- Strong, M., Graeber, T., Beeby, M., Pellegrini, M., Thompson, M., Yeates, T., and Eisenberg, D. (2003). Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Research*, 31(24):7099–7109.
- Sundaresh, S., Doolan, D., Hirst, S., Mu, Y., Unal, B., Davies, D., Felgner, P., and Baldi, P. (2006). Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics*, 22(14):1760.
- Tyers, M. and Mann, M. (2003). From genomics to proteomics. *Nature*, 422:193–197.

White, A., Daly, D., Varnum, S., Anderson, K., Bollinger, N., and Zangar, R. (2006). ProMAT: protein microarray analysis tool. *Bioinformatics*, 22(10):1278.

Wilkins, M., Pasquali, C., Appel, R., Ou, K., Golaz, O., Sanchez, J., Yan, J., Gooley, A., Hughes, G., Humphery-Smith, I., et al. From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis.