Jason Ptacek

Bioinformatics and Proteomics: Partners in Progress

Proteomics is a field that is blossoming, both in applications and possibilities. The number of genomes sequenced is rapidly increasing, leaving the task of understanding how the genome's biochemical relative, the proteome, works. To do this, a proteome chip, similar to the DNA chips, was made. While the fruits of this project are just being realized, it is clear that bioinformatics will play a key role in the data analysis and its storage in databases. This paper summarizes the field of proteomics, giving particular attention to the proteome chip, and proposes whole-proteome analysis of the yeast's 125 kinases.

Proteomics Before The Proteome Chip

Large-scale protein analysis became possible with the development of 2D-PAGE. 2D-PAGE is most used for protein separation as well as identifying the presence or absence of proteins from different genetic backgrounds. Another tool, two-hybrid analysis, allows potential protein-protein interactions to be identified. This technique was used to identify 957 putative interactions in S. cerevisiae, allowing proteins of unknown function to be classified on the basis of their interactions (Uetz, Giot et al. 2000).

Of considerable interest has been the development of DNA microarrays. This technology generates expression profiles from different backgrounds (DeRisi, Iyer et al. 1997), identifies variations in gene sequence (Winzeler, Richards et al. 1998), and when combined with chromatin immunoprecipitation it can identify all the binding sites of transcription factors within a genome (Iyer, Horak et al. 2001). Yet, one of the major goals

of science continues to be the elucidation of the function of individual proteins. Function is itself a vague term with a protein's molecular function being related to, but different from, a protein's cellular function (Gerstein and Jansen 2000). Yeast proteome microarrays, or chips, were developed to help determine a protein's molecular function in relation to all other proteins in the proteome. Each chip contains nearly all of the proteins in yeast individually spotted onto a special glass slide. These chips allow proteins to be characterized by their enzymatic activity and by their binding partners.

The Proteome Chip of Today

MacBeath and Schreiber showed that it was possible to characterize proteins spotted onto glass slides by testing three known protein-protein interactions, three known kinase-substrate reactions, and three known protein-ligand reactions (MacBeath and Schreiber 2000). Zhu *et al*. demonstrated that the yeast kinase family could be characterized by their substrate specificity and site of phosphorylation (Zhu, Klemic et al. 2000). Zhu and colleagues then developed the yeast proteome chip (Zhu, Bilgin et al. 2001). Many biological challenges were overcome in the development of these chips, and are discussed along with the advantages of the chips by Zhu and Snyder (Zhu and Snyder 2001). Proteome chips also present challenges in the creation and storage of large amounts of data.

This paper will focus on the former challenge, that of the initial analysis of the microarray data. For protein-protein and protein-lipid interactions, the protein chip is probed with the biotinylated protein or lipid of interest generating spots of interaction on the chip (Zhu, Bilgin et al. 2001). A positive calling algorithm was developed to identify spots of interaction (Zhu, Bilgin et al. 2001). This process removed spots of

contamination, corrected for global variations in intensity from chip to chip and experiment to experiment, and produced the ratio $r = G/R$ as a measure of the binding per amount of protein per spot. Here, G represents the filtered lipid-binding signal and R represents the amount of protein at each spot, as determined by probing with GST antibody. Positive "hits" were identified in the various experiments, allowing further analysis.

The protein chip can accommodate a wide range of experiments, enabling many types of analysis to be completed. For example, probing the chip with biotinylated calmodulin revealed 33 potential calmodulin-binding proteins as well as six of the known calmodulin-binding proteins (Zhu, Bilgin et al. 2001). By using multiple sequence alignment, 14 of the 39 proteins were shown to share a common motif. This result will allow new potential calmodulin-binding proteins to be discovered in other organisms by searching for this conserved motif.

The Future of Proteome Chips

"[The proteome chip] is a biochemist's dream, to be able to look for any activity over the entire proteome," says Eric Phizicky, a biochemist at the University of Rochester Medical Center (Service 2001). One activity that will be looked at is kinase activity using the entire yeast proteome. Each kinase will be individually purified in a high-throughput manner and substrates identified by individually incubating the proteome chip in the presence of each kinase. Lists of potential substrates for each of the 125 yeast kinases will be compiled. Analysis of this data will include multiple-sequence alignment that searches for common motifs among the substrates of each individual kinase. This can be extended to searching for potential substrates of kinases in other organisms based on homology to

yeast kinase-substrate pairs. This 'homology method' is often used to extend results from one organism to another. Future methods will involve structural homology searches, such as looking for common folds that a kinase recognizes or folds in kinases that predict its activity or substrate specificity. These techniques are much more computationally intense than sequence homology searches, and more structural data is needed before this can be used on a proteome-wide level.

Clustering algorithms, used extensively for DNA microarrays, could be used to cluster substrates according to folds or other characteristic. Phylogenetic trees could potentially be created that organize the yeast kinases according to the clustered substrates as a way to visualize structure-function information. For example, it would be interesting to see if substrates that are phosphorylated on tyrosine residues share common folds. However, for the present, analysis will focus on finding common motifs of the substrates for each kinase.

Conclusion

Tackling the human genome with its estimated 30,000 to 40,000 genes has been relatively simple compared to the proteome, with experts predicting between 200,000 and 2 million proteins (Service 2001). The proteome chip is still in its infancy, but the expectations of these chips are that they will change the way science is conducted. Studying the proteome as a whole should be much more informative by revealing activities and interactions of proteins in parallel rather than as an isolated protein removed from its cellular network. With the large amount of complex data, it will be up to computational analysis to sift through the data and give it meaning, both to the proteome of study and to the proteomes of other organisms.

Reference

DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.

Gerstein, M. and R. Jansen (2000). "The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function?" Curr Opin Struct Biol **10**(5): 574-84.

Iyer, V. R., C. E. Horak, et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**(6819): 533-8.

MacBeath, G. and S. L. Schreiber (2000). "Printing proteins as microarrays for high-throughput function determination." Science **289**(5485): 1760-3.

Service, R. F. (2001). "PROTEOMICS: High-Speed Biologists Search for Gold in Proteins." Science **294**(5549): 2074-2077.

Service, R. F. (2001). "PROTEOMICS: Searching for Recipes for Protein Chips." Science **294**(5549): 2080-2082.

Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-7.

Winzeler, E. A., D. R. Richards, et al. (1998). "Direct allelic variation scanning of the yeast genome." Science **281**(5380): 1194-7.

Zhu, H., M. Bilgin, et al. (2001). "Global analysis of protein activities using proteome chips." Science **293**(5537): 2101-5.

Zhu, H., J. F. Klemic, et al. (2000). "Analysis of yeast protein kinases using protein chips." Nat Genet **26**(3): 283-9.

Zhu, H. and M. Snyder (2001). "Protein arrays and microarrays." Curr Opin Chem Biol **5**(1): 40-5.