Dry microarrays: a generalized data source for genome and proteome analysis

G. Sampath

Department of Computer Science, The College of New Jersey, Ewing, NJ 08628
sampath@tcnj.edu

ABSTRACT

Microarray analysis methods can be applied to characteristic data obtained from biological sequences in matrix form (which could include data from experiments or bioassays as well). The rows correspond to a genome region (intron or exon) or a protein, the columns to data obtained from the genome or proteins under selected conditions. In such a *dry microarray*, tasks like protein classification and construction of gene networks are aided by the use of various clustering methods applied to the heterogeneous data matrix. These methods, suitably modified, can also be used for simultaneous multiple alignment of all the sequences instead of pairwise or over small numbers of sequences at a time (as is the current practice). The feasibility of the concept is shown with results obtained from the construction and analysis of a dry microarray for protein classification.

Keywords: bioarrays, proteomics, computational methods

1 GENOMICS AND PROTEOMICS: MICROARRAY AND SEQUENCE DATA

Effective and efficient biological data analysis and visualization have become necessary with increasing amounts of data becoming available through the use of rapid sequencing and bioassay techniques. Thus a variety of sequence analysis methods have been developed to study higher-level properties of biological sequences, such as secondary and tertiary structure of proteins, homology, and phylogeny [11]. Similarly microarray technology [1] has made it possible to obtain expression data for large numbers of genes under a range of conditions (such as time evolution, samples from subjects both target and control, and differing environments), with the technology now being extended to proteins as well [6]. A large number of algorithms are now available to extract meaningful information from the data such as secondary and tertiary structure from sequences [13], relationships among the genes and conditions [2], classification of proteins [7, 9, 10], and construction of gene and protein networks [14]. Some of the methods [3, 6, 11] developed include multiple alignment, dendrogram analysis, principal component analysis, eigengene analysis, biclustering (patterned submatrices), support vector machines, singular value decomposition, hidden Markov models, and self-organizing maps, with statistics playing an important role in many of them [15].

By and large the analysis and modeling of biological data is currently centered on specific kinds of data and has specific objectives. Thus computational modeling in biology is usually driven by the kind of data it is based on and has concentrated on an individual goal such as sequence analysis at different levels, phylogeny or classification, or construction of networks. Recently this has given way to computational models that attempt to extract patterns from heterogeneous data using various probabilistic/statistical [4] and data fusion models [5, 8, 12] that may include phylogenetic profiles in the form of bitmaps. In this report, a general form of microarray is proposed in which heterogeneous data from diverse sources can be brought together in a single matrix to which a wide range of microarray-based computational techniques can be applied simultaneously at all levels: sequence, genome, proteome, and cell.

2 DRY MICROARRAYS AND THEIR POTENTIAL APPLICATIONS

If sequence level data from genomes and proteins are arranged in the form of a microarray-like matrix, many of the methods used in gene expression analysis can be applied to biological sequences as well. Such an array could lead to efficient ways of classification as well as identification of relationships that may be useful in constructing networks at the gene and protein levels. This suggests further that by combining heterogeneous data from a number of sources in the microarray, a microarray can serve as a framework for data fusion. In such a *dry microarray*, the columns can be any of a number of defining properties of the genome or protein, including, for example, 1) subsequences of given lengths, 2) subsequences that are similar in some sense, or 3) secondary structure motifs. These are in addition to the more conventional types of microarray data such as experimental, tissue, and cell cycle data. The wide range of statistical and computational methods used in microarray analysis can now be applied to the dry microarray. The columns can also be weighted to bias the analysis towards properties of interest. Potential applications include:

1) *Cluster analysis* of coding and non-coding regions of a genome, which can be extended to the study of relationships between specific sites and upstream regions,
2) *Protein classification*, which could lead to identification of protein families based on different criteria on the columns, and
3) *Secondary structure* (RNA and protein) and *tertiary structure identification* (with column properties based on motifs in the primary sequences).

3 FEASIBILITY: EXAMPLE AND COMPUTATIONAL RESULT

The feasibility of the approach is illustrated by constructing a dry microarray for a set of 200 proteins randomly selected from the Swiss-PROT database (which contains 6115 unique sequences after duplicates are removed). *Hierarchical clustering* [15] was used to group proteins in an unsupervised classification. Use of the CLUSTER software of Eisen (*http://rana.lbl.gov*, see also [15]) led to Figure 1, which shows a clustered display of the 200 proteins. The y axis corresponds to protein, the x axis to 7872 subsequences of 1 to 3 residues in the sample protein set. The proteins were randomly selected from out of 6115 unique proteins (obtained after removing duplicates from over 13000 listed). The selected ones were scanned to obtain all 1-, 2-, and 3-residue subsequences and their frequencies. Each frequency was divided by the length of the sequence and entered into the dry microarray. Thus an array entry is the ratio of subsequence frequency to length of primary sequence. Light bands (visible as vertical stripes) occur for certain 2- and 3-residue subsequences as well as for individual proteins over (ordered) subranges of the subsequences (visible as short dull horizontal stripes). The vertical bands are indicative of the presence of multiple alignment among the proteins mapped by the stripes, while the horizontal stripes are suggestive of groups of subsequences (typically of the same length) occurring in a single protein.

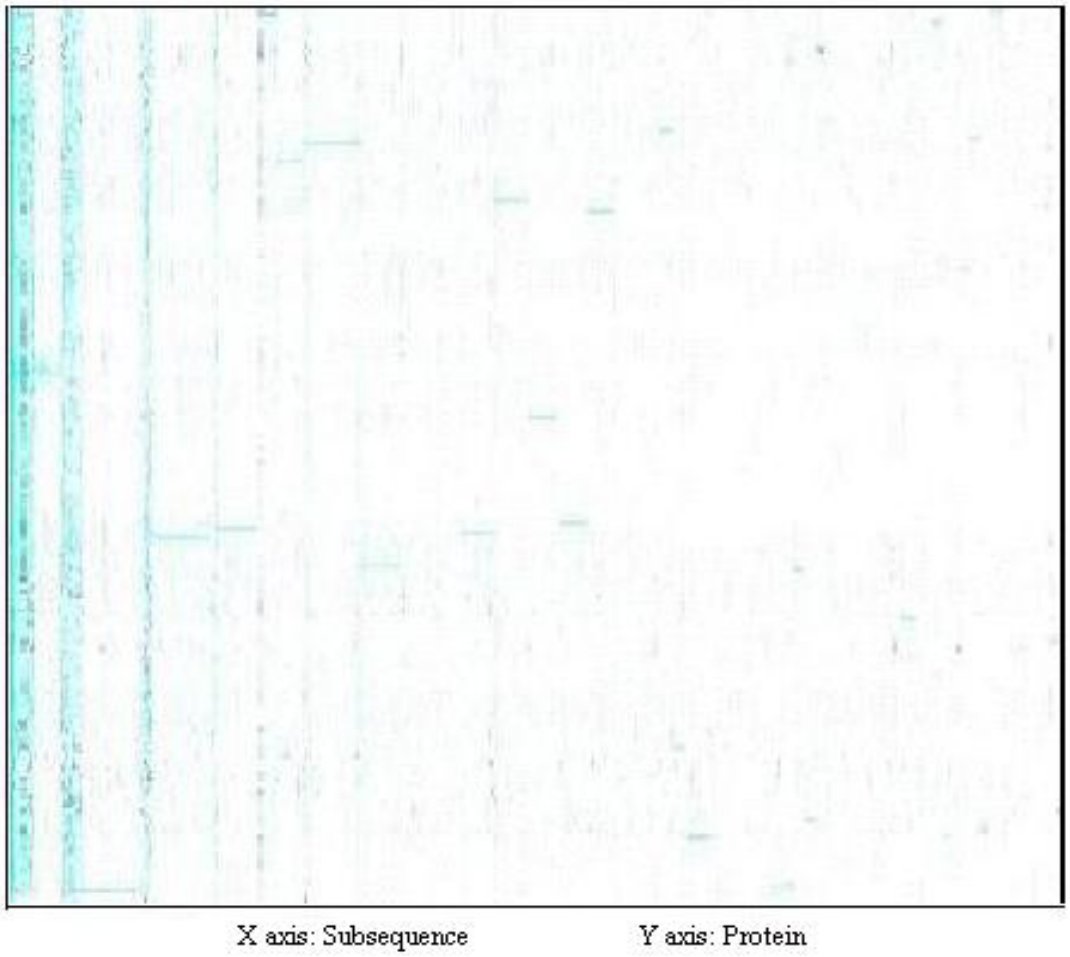

Figure 1. Hierarchical clustered display of dry microarray of 200 proteins from Swiss-PROT

29	17	19	40	29	1	17	9	6	33

Table 1. Sizes of clusters obtained by K-means clustering of 200 proteins from Swiss-PROT

Clustering was also done using K-means clustering (using the same software). Table 1 shows the cluster sizes that were obtained with 10 clusters.

Work is ongoing on data analysis using a wide range of column definitions based on one or more of the following: sequence data (based partly on scoring matrices like PAM and BLOSUM), data based on DNA/protein motifs, expression data, and secondary structure data (proteins/RNA).

REFERENCES

[1] P. Baldi and G. W. Hatfield. *DNA Microarrays and Gene Expression*. Cambridge University Press, Cambridge (U.K.), 2002.

[2] A. Brazma and J. Vilo. "Gene expression and data analysis." FEBS Letters **480**, 17-24 (2000).

[3] Y. Cheng and G. M. Church. "Biclustering of expression data." Proceedings Intelligent Systems in Molecular Biology, 2000.

[4] M. Deng, S. Mehta, F. Sun, and T. Chen. "Inferring domain-domain interactions from protein-protein interactions." http://www.genome.org/cgi/doi/10-1101/gr.153002.

[5] R. Jansen, N. Lan, J. Qian, and M. Gerstein. "Integration of genomic datasets to predict protein complexes in yeast." J. Structural and Functional Genomics **90**, 1-11 (2002).

[6] I. S. Kohane, A. T. Kho, and A. J. Butte. *Microarrays for an Integrative Genomics*. MIT Press, Cambridge (Mass.), 2003.

[7] A. Kumar, B. Smith. "A framework for protein classification." Proceedings of the German Conference on Bioinformatics, volume 2, 55-57, 2003.

[8] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan and W. S. Noble. "Kernel-based data fusion and its application to protein-protein function prediction in yeast." Proceedings of the Pacific Symposium on Biocomputing, January 3-8, 2004.

[9] A. H. Liu and A. Califano. "Functional classification of proteins by pattern discovery and top-down clustering of primary sequences. IBM Systems J. **40** (2), 379, 2001.

[10] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. "SCOP database in 2002: refinements accommodate structural genomics." Nucleic Acids Research **30** (1), 264-267 (2002).

[11] R. Mount. *Bioinformatics*, Cold Spring Harbor Press, Cold Spring Harbor (NY), 2001.

[12] P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy. "Gene functional classification from heterogeneous data." Procs. 5th International Conference on Computational Molecular Biology, April 21-24, 2001, pages 242-248.

[13] B. Rost. "Protein secondary structure prediction continues to rise." J. Structural Biology **134**, 204-218 (2001).

[14] B. Schwikowski, P. Uetz, and S. Fields. "A network of protein-protein interactions in yeast." Nature Biotech. **18** (12), 1257-1261 (2000).

[15] T. Speed (ed.). *Statistical Analysis of Gene Expression Microarray Data*. CRC Press, Boca Raton, 2003.