URL: http://www.iupac.org/symposia/proceedings/phuket97/oxender.html

Genomes, Proteomes, and Bioinformatics

Dale L. Oxender, Brian Moldover, and James D. Cavalcoli

Parke-Davie Pharmaceutical Res., Division of Warner-Lambert Co., Ann Arbor, MI 48105

Abstract: We are now facing an unprecedented flood of information as a result of complete genome sequencing projects. In the last year and a half the complete genome of the first eucaryote, *saccharomyces cerevisiae* (14 Mb, 6,340 ORFs) has been added to the database which already contains the complete genomes of a number of prokaryotes such as *Haemophilus influenzae* (1.8Mb, 1,1743 ORFs), *Mycoplasma genitalium*, *Methanococcus jannaschii* (1,738 ORFs) and *Escherichia coli* (4285 ORFs). More ambitious projects are well underway such as *C. elegans* (ca. 20,000 ORFs) and the human genome project (3,000Mb, ca, 50,000 ORFs). The next challenge after the post genome era is the assignment of function to the expressed genes, which will help identify genes associated with disease processes. Powerful new technologies are being developed to map and quantify proteins expressed within a cell. Assignment of a phenotype for each gene as a result of systematic deletion of each ORF can be an important first step in creating a gene-protein database. Bioinformatics is the scientific discipline that is essential for bringing proteome and genome analyses together. Proteome analysis by separation of complex mixtures of cellular proteins by 2-D electrophoresis is an important tool for creating the gene-protein data base. Protein identification has been greatly improved as a result of new techniques in mass spectrometry (MS). Two of these techniques are matrix assisted laser desorption and ionization (MALDI) and electrospray ionization (ESI). Using the new MS techniques it is possible to rapidly measure the masses of each of the proteins and the peptides that are generated by enzymatic digestion. A fingerprint of each protein from the 2-D analysis can now be used to identify the corresponding gene (ORF) in the database. Several biotech corporations have been formed to take advantage of the genome-proteome technology.

INTRODUCTION

The exponential increase in the amount of information as a result of the ongoing mass DNA sequencing efforts of many laboratories throughout the world, has resulted in an enormous flow

*Invited lecture presented at the International Conference on Biodiversity and Bioresources: Conservation and Utilization, 23–27 November 1997, Phuket, Thailand. Other presentations are published in *Pure Appl. Chem.,* Vol. 70, No. 11, 1998.

of information which is growing exponentially. In the last year and a half the complete genome of the first eukaryote, *saccharomyces cerevisiae* (14 mb, 6,340 ORFs) has been added to the database which already contains the complete genomes of a number of prokaryotes such as *Haemophilus influenzae* (1.8Mb, 1,1743 ORFs), *Mycoplasma genitalium, Methanococcus jannaschii* (1,738 ORFs) and *Escherichia coli* (4,285 ORFs). More ambitious projects are well underway such as *C. elegans* (ca. 20,000 ORFs) and the human genome project (3,000 Mb, ca. 50,000 ORFs).

Genome Sequencing Projects	
H. influenzae	Completed
M. genitalium	Completed
S. cerevisiae	Completed
S. aureus	Completed
M. jannischii	Completed
H. plyori	Completed
E. coli	Completed
B. subtilis	Completed
C. elegans	60%
H. sapiens	~ 1%?

DNA sequence data is now obtained on an unprecedented scale. It is now the job of computers to sort through this database to assemble these sequence fragments into cohesive units, and to determine whether or not the new sequence is biologically relevant. Bioinformatics is a growing field which applies computational methods to biological problem solving. One of the key areas where bioinformatics impacts science, is in assisting in the management and deciphering on this large flow of data.

Genomic information is useful to determine the template from which biological information is read. The importance of the linkage between the DNA sequence information and the biologically relevant expression of proteins is essential for the next dimension of our study of biology. As genomic sequence information is obtained and completed, we can use this information as a stepping stone to understand what proteins might be expressed by a cell. Since the total DNA content of an organism is referred to as the “genome”, Marc Wilkins coined the phrase “proteome” (Wasinger et al., 1995) to indicate the complete protein potential of a cell. In addition we would like to describe the “protein expression profile” of a cell as being that portion of the proteome which is expressed by a given cell under given conditions.

Not only is protein expression an important issue but also determining the level of expression of each protein, and determining the post-translational modifications for a given protein. None of these issues can be resolved from the DNA sequence alone. A review addressing this was recently published (Humphrey-Smith & Blackstock, 1997). In addition, protein expression profile would allow scientists to study the changes in protein expression with changes in cell environment or response to chemical stimuli, or changes in expression between tissues in the same organism.

One of the goals of genomics and proteomics is to provide a linkage between the proteins observed in a cell and the genes encoded by the DNA sequence. This linkage can be determined if several conditions are met: (i) the complete genome of an organism must be available, (ii) there must be a way to separate the complex mixture of proteins which are made by a cell, and identify them by some sort of criteria which makes them unique, and (iii) there must be a way of identifying which gene a protein came from.

In the case for some of the prokaryotic organisms, such as E. coli, condition I, has been met and the complete genome of those organisms is available. Condition III, can be determined to a great extent by predicting the potential proteins from the ORFs of the genome. Some post-translational modifications such as phosphorylation, methylation and acetylation are known, or can be predicted from amino acid sequence homology. While complex modifications such as lipid, or carbohydrate moiety additions are less predictable and more variable in their extent, methods are being devised which will account for these modifications.

The second condition listed above is the focus of the remainder of this paper. Current technology for identification of proteins in a complex mixture rely heavily on the techniques of two-dimensional gel electrophoresis (2D-gels). In this method, proteins are separated in the first dimension based on their charge, using iso-electric focusing. In the second dimension these proteins are separated by SDS PAGE which separates by mass. 2D-gel technology has been used for over 20 years, and in the case of E coli, over 1400 proteins have been described and published (VanBogelen et al., 1996). 2D-gels allow for the separation of complex mixtures, and then subsequent analysis through immunoblotting, co-migration with purified proteins, binding assays, genetic analysis, enzyme induction or repression, and other methods (VanBogelen & Olson, 1995). Using these methods over 300 proteins have been identified on 2D-gels.

Because of the nature of 2D-gels, it is possible to estimate the Isoelectric point (PI) or Molecular mass (MW) of a given spot on a gel. However, the accuracy of those estimations, specifically that of the MW, are not precise enough to achieve the gene-protein linkage without using the secondary techniques outlined above for confirmation of identity.

If the accuracy of estimating the MW were increased, a substantial number of unique candidates for identification could be found. There are many methods which can accurately determine the MW of a protein and which can readily be adapted to use with current 2D-gel technology. One of these methods is matrix-assisted laser desorption / ionizing mass spectrometry (MALDI-MS) determination of MW is currently accurate to 0.1-0.2% when scanning down a gel, using calibration standards located alongside the sample (Ogorzalek Loo et al., 1997) MALDI-MS is a technique which can be used for protein identification and characterization including post-translational modifications and relationships to ligands or other moieties (Nguyen et al., 1995; Stults, 1995; Aaluzec et al., 1995). The use of MALDI-MS, in conjunction with N-terminal sequencing and amino acid composition was used to identify proteins across different species (Cordwell et al., 1995). In addition a mixture of yeast proteins were clearly identified with a high degree of accuracy (Shevchenko et al., 1996). MALDI-MS has been shown to be applicable for mass determination on multiple proteins in single acrylamide gels and thus lends itself to high throughput screening (Ogorzalek Loo et al., 1997). It has already been demonstrated that MALDI-MS is a very sensitive tool for determining mass of peptide fragments cleaved from within polyacryalmide gels (Courchesne et al., 1997; O'Connell & Stults, 1997).

Therefore in a given mixture of proteins separated by 2D-gel, the use of mass spectrometry techniques can increase the number of unique candidates for identification, and identifications may be confirmed by fragmenting the proteins, and reanalyzing them using the same or alternate techniques. When the mass of a protein or it's resulting peptide map obtained through MALDI is not sufficient for identification, other techniques such as Mass Spectrometry / Mass Spectrometry (MS/MS) can be used to look at the fragments to determine primary sequence information and make identification of the protein.

Techniques such as MALDI-MS are extremely sensitive (low femtomole range), and suffer little sample loss due to handling. Thus it is possible to identify proteins which are present in very low levels within cells. In addition, methods such as MALDI-MS are capable of being automated, and the computational analyses can be performed in a pipeline manner as well. This suggests rapid identification of a large number of proteins.

The study of protein expression patterns is essential to understanding the role these proteins play in the function of cells and tissues. This process is referred to as "Functional Genomics". The linkage of protein expression back to the gene of origin is an important step in understanding the changes which can occur between the DNA and protein sequence levels. Finally, comparing expression of genes, through mRNA, and expression or proteins, it will be possible to better understand the regulation of metabolic events and cellular response to environmental changes.

REFERENCES

1. Cordwell, S., Wilkins, M., Cerpa-Pojak, A., Gooley, A., Duncan, M., Williams, K.& Humphrey –Smith, I. (1995). Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionization / time-or-flight mass spectrometry and amino acid composition. Electrophoresis 16, 438-443.
2. Courchesne, P.L., Luethy, R. & Patterson, S. D. (1997). Comparison of in-gel and on-membrane digestion methods at low to sub-pmol level for subsequent peptide and fragment-ion mass analysis using matrix-assisted laser-desorption / ionization mass spectrometry. Electrophoresis 18, 369-381.
3. Humphrey-Smith, I. & Blackstock, W. (1997). Proteome analysis: genomics via the output rather than the unput code. J.Prot. Chem. 16(5), 537-544.
4. Nguyen, D., Becker, G. & Riggin, R. (1995). Protein mass spectrometry: applications to analytical biotechnology. J. Chromatogr. A 705(1), 21-45.
5. O'Connell, K.L. & Stults, J.T. (1997). Identification of mouse liver proteins on two-dimensional electrophoresis gels by matrix-assisted laser desorption/ionization mass spectrometry of *in situ* enzymatic digests. Electrophoresis 18, 349-359.

6. Ogorzalek Loo, R., Mitchell, C., Stevenson, T., Martin, S., Hines, W., Juhasz, P., Patterson, D., Loo, J. & Andrews, P. (1997). Sensitivity and mass accuracy for proteins analyzed directly from polyacrylamide gels: Implications for proteome mapping. Electrophoresis 18, In Press.

7. Shevchenko, A., Jensen, O., Podtelejnijov, V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H.& Mann, M. (1996). Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. Proc. Natl. Acad. Sci. USA 93, 14440-14445.

8. Stults, J. (1995). Matrix-assisted laser desorption / ionization mass spectrometry (MALDI-MS). Curr. Opin. Struct. Biol. 5(5), 691-698.

9. VanBogelen, R., Abshire, K., Pertsemlidis, A., Clark, R. & Neidhardt, F. (1996). Gene-protein database of Escherichia coli K-12, Edition 6. In Escherichia coli and Salmonella 2nd edit. (Neidhardt, F.C., ed.), Vol. 2, pp. 2067-2117. 2 vols. ASM Press, Washington, D.C.

10. VanBogelen, R. & Olson, E. (1995). Application of two-dimensional protein gels in biotechnology. In Biotechnology Annual Review (El-Gewely, M.R.,ed.), Vol. 1,pp.69-103. 1 Vols. Elsevier, New York.

11. Wasinger, V. C., Cordwell, S.J., Cerpa-Polijak, A.,. Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L. &Humphrey-Smith, I. (1995). Progress with gene-product mapping of the Molecules: Mycop. Matrix-assisted laser desorption ionization mass spectrometry: applications in peptide and protein characterization. Protein Expr. Purif. 6(2), 109-123.

12. Zaluzec, E., Gage, D. & Watson, J. (1995). Matrix-assisted laser desorption ionization mass spectrometry: applications in peptide and protein characterization. Protein Expr. Purif. 6(2), 109-123.