BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **231,** 1–6 (1997)
ARTICLE NO. RC966045

BREAKTHROUGHS AND VIEWS

Of Genomes and Proteomes

Peter James
Protein Chemistry Laboratory, Universitätstrasse 16, Swiss Federal Institute of Technology, ETH-Zentrum, 8092 Zürich, Switzerland

Received December 26, 1996

The era of complete genome sequences has arrived and with it vast amounts of data which must be annotated, cross referenced, and placed within the regulatory networks which define the physiology of an organism. One eucaryotic and three procaryotic genomes have been completed and the data made available and another 50 sequences are expected to be completed by the end of the decade. One of the first steps in the new post genome era will be to decipher the functions of the huge numbers of new open reading frames. Various approaches to investigate what unknown genes do and how genes interact together within an organism are being undertaken, including (1) the simultaneous measurement of the expression levels of all genes in a cell and (2) the mapping and quantitation of all proteins expressed within a cell. The idea of systematically mapping and identifying the total protein complement of the genome (the "proteome") arose over 20 years ago when the separation of proteins from total cell extracts by two dimensional (2D) gel electrophoresis was developed. This review will focus on the use of 2D gel electrophoresis as the basis for constructing proteome maps and on the rapid advances in mass spectrometry which will allow the large-scale, automated identification of proteins which is necessary for the creation of such databases.

THE GENOME REVOLUTION

In 1977, Fred Sanger published the first complete DNA sequence of an organism, virus ϕX174 (5386 bp, coding for 9 proteins [1]). Since this first genome was determined, rapid advances in computing and robotics have enabled sequencing of much larger genomes by a brute force approach, shotgun sequencing [2]. This entails the automated sequencing of hundreds of thousands of small (300-500 bp) randomly generated DNA fragments which are assembled by computer to reconstruct the whole genome. In the last year and a half, the genomes of three procaryotes, *Haemophilus influenzae* (1.8 Mb, 1,743 open reading frames (ORFs) [2]), *Mycoplasma genitalium* (0.7 Mb, 470 ORFs [3]), and *Methanococcus jannaschii* (1.7 Mb, 1,738 ORFs [4]) as well as the first eucaryotic genome *Saccharomyces cerevisiae* (14 Mb, 6,340 ORFs [5]) have been completed. Large scale projects are underway and the genome of the nematode *Caenorhabditis elegans* (100 Mb, ca. 20,000 ORFs), is tagged for completion by mid 1998, and the most ambitious, the human genome project (3,000 Mb, ca. 50,000 ORFs) could be finished as soon as 2005. Rapid advances continue to be made in DNA sequencing technology and it has been estimated that 50 genomes will have been completed by the turn of the century [John Reeve, personal communication]. Industrial concerns are now having organisms sequenced to order by companies like TIGR (The Institute for Genomic Research, http://www.tigr.org/tigr home/index.html) and GTC (Genome Therapeutics Corporation, http://pandora.cric.com/). The genomes which are being targeted for sequencing include those of pathogens such as *Staphylococcus aureus* to help in the development of new antibiotics and extremophile genomes such as that of *Aquifex*, a chemoautolithotroph that grows at 90° C, in order to find enzymes stable enough for use in industrial environments.

2D GEL ELECTROPHORESIS

These complete genome sequences, besides being fonts of information, pose many new questions such as: how to interpret and co-ordinate this new data: how to systematically find roles for the 40 to 60% of new bacterial ORFs that cannot be assigned a function based on homology searches, and how to connect the fairly static genome with the dynamics of gene transcription and translation ? A technique first developed in the mid '70s by Patrick O'Farrel, two dimensional (2D) gel electrophoresis [6], may be the key which will allow one to begin to tackle these problems. In this method, protein

0006-291X/97 $25.00

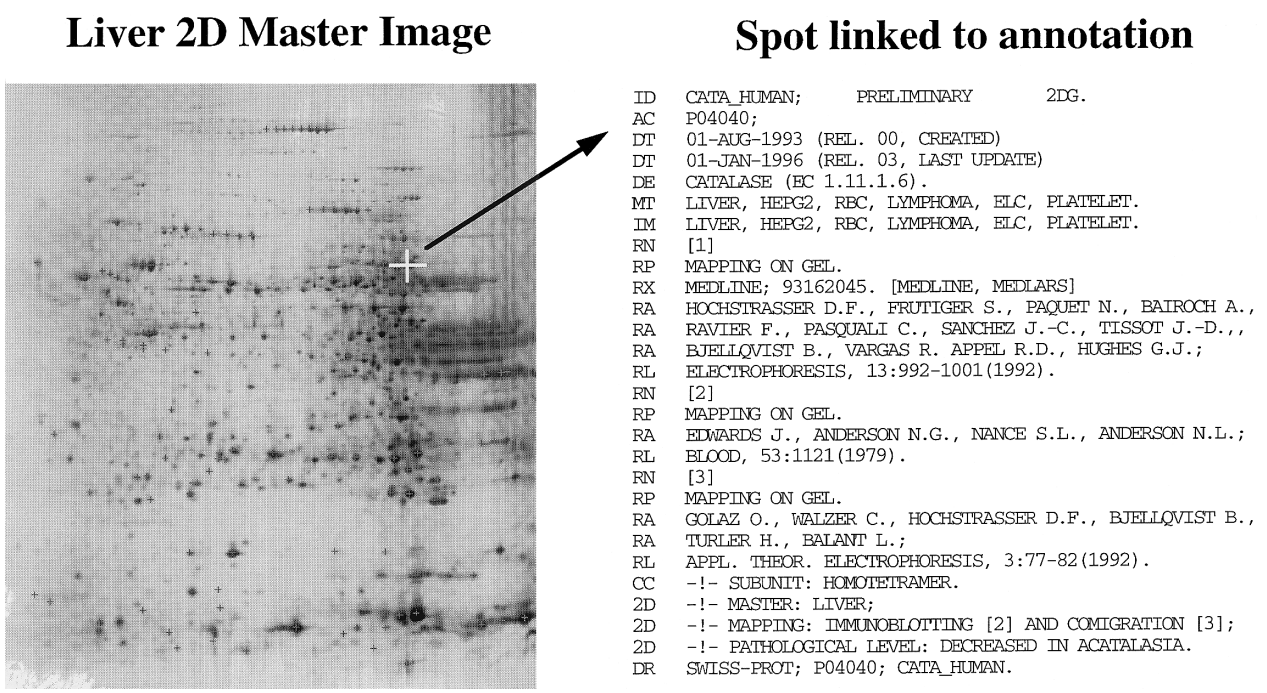

FIG. 1. The 2D gel master image of liver from the ExPASy 2D gel database. The gel image is available from the URL http://expasy.hcuge.ch/cgi-bin/map1. 2D spots can be selected and the annotation linked to the spot displayed. The spot marked with the white cross is catalase. The annotation indicates how it was identified (by immunoblotting) and how it alters in disease states (decreased in acatalasia) and provides a futher link to the entry in the SwissProt database.

extracts from cells are separated on polyacrylamide gels, first according to charge and then by size. The resulting gel patterns, representing up to 10,000 proteins [7], reveal not only the amount of each protein that is being expressed but allow one to determine if they have been post-translationally modified (for example, the visualisation of phosphoproteins using ^{32}P labelling). Scanned images of 2D-gels of cells in different physiological states can be analysed using computer programs and the changes quantitated [8]. For example, 2D gel databases of cells or body fluids in normal and pathological states are being developed by Ron Appel and Denis Hochstrasser for use in medical diagnosis as a sort of "clinical molecular scanner" [9]. The Swiss-2DPAGE database (Figure 1) contains 2D master gel images of cells representing the "normal" state, whereas the complementary Swiss-2Disease database consists of annotated gel images obtained from cells and tissues in various pathological states such as renal failure and myeloma. Both are available via the ExPASy world-wide web server at the URL: http://expasy.hcuge.ch.

One of the main problems with 2D gel databases has been the reproducibility of the protein separation pattern between different laboratories. The development of first dimension gels with covalently immobilised pH gradients, Immobiline™ strips, has greatly reduced the protocol dependant variability associated with dynamic pH gradient tube gels formed using ampholyte buffers [10]. This has greatly facilitated the exchange of 2D gel information between labs and a WWW federation [11] is being set up to actively promote and standardise exchange protocols. The common interface for data accession allows 2D gel databases from all WWW federation sites to be searched and in the near future, it may be possible to match gel images over the network [12].

PROTEIN ANALYSIS AND DATABASE SEARCHING

The tremendous resolving power and sensitivity of 2D electrophoresis and the ability to combine it with other methods by electroblotting to inert supports for Edman sequencing to identify proteins [13, 14, 15], has allowed the construction of 'cell' maps [16] of the most abundant proteins. Standard 2D gels visualised by silver staining show around 1-2,000 spots, 75% of which are in amounts less than 500 femtomoles. Large format gels allow the visualisation of up to 10,000 spots, though little advantage can be taken of this for mapping (other than by genetic methods), since over 95 % occur in amounts beyond the current limits of commercial high sensitivity Edman sequencers.

Protein identification has taken a quantum leap in speed and sensitivity through the advent of new 'soft' ionisation techniques in mass spectrometry (MS). Two

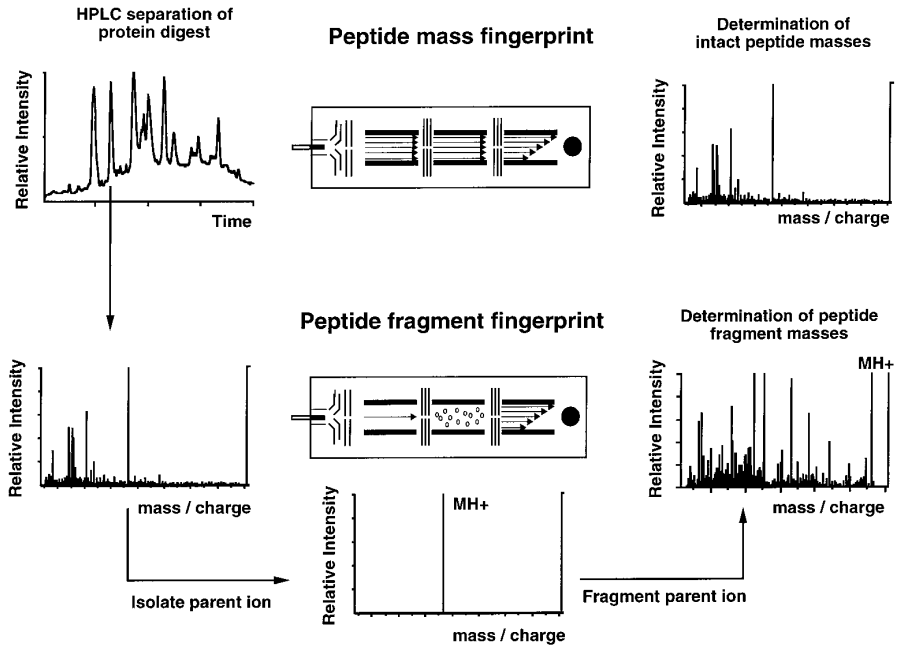

FIG. 2. Peptide mass and peptide fragment fingerprint data collection. The peptides arising from a protein digest are partially separated by capillary HPLC which is interfaced by an electrospray ionisation (ESI) interface to a triple quadrupole mass spectrometer and the masses of the intact peptides determined. The set of masses is the peptide mass fingerprint of the protein. During the HPLC run, peptides can be isolated by the first quadrupole for fragmentation. The set of fragment ion masses is the peptide fragment fingerprint of that particular peptide.

techniques, matrix assisted laser desorption and ionisation (MALDI) developed by Franz Hillenkamp [17] and electrospray ionisation (ESI) by John Fenn [18], allow the accurate measurement of peptide and protein molecular weights. The sensitivity level for peptide detection by MALDI in time of flight (TOF) mass spectrometers is of the order of tens of femtomoles [19] whilst triple quadrupole and ion trap mass spectrometers extend the range further with sensitivity levels in the attomole range [20]. Using the new MS techniques it has become possible to rapidly measure the masses of the peptides produced by residue specific enzymatic or chemical digestions. The recognition that the set of masses produced by such a digestion is unique to a protein, gave rise to the concept of 'Peptide mass fingerprinting: the identification of a protein in a database using a set of molecular masses of peptides generated by a specific digestion (Figure 2). This is emerging as a reliable and rapid alternative to peptide sequencing by Edman degradation. A flurry of papers from the groups of Bill Henzel, Darryl Pappin, Matthias Mann, John Yates and ourselves appeared in the middle of 1993 [21, 22, 23, 24, 25]. It is both a rapid and sensitive technique, for example, by using a multiple position sample plate and automatic data collection, one can obtain peptide mass fingerprints for 100 protein digests in an unattended overnight run by MALDI-TOF MS. Peptide mass fingerprinting is useful for identifying proteins in protein databases but the confidence level drops rapidly when searching six frame translations of DNA databases. This can be remedied by the use of a second, orthogonal data set, such as the masses from a digest using an enzyme or chemical with a different specificity to the first, or by deuterium exchange of the first digest [26].

Triple quadrupole and ion trap mass spectrometers can be used to obtain partial and complete sequence information can be obtained by analysing fragmentation spectra of peptides (MS/MS) [27]. Peptides can be fragmented by isolating the 'parent' peptide ion from other ions and then accelerating the parent into a region containing an inert gas such as argon. The ion undergoes multiple collisions with the gas and fragments in a random manner along the amide backbone of the peptide. The resulting 'daughter' ions are detected and a fragmentation pattern which contains sequence specific information is collected. This can be achieved with less than femtomole amounts of material [20]. An analogous procedure to MS/MS can be carried in a MALDI-TOF MS by Post source decay (PSD) analysis though around 100 femtomoles of peptide is required [28]. The MS/MS or PSD spectra can be manually interpreted to give a sequence for database searching.

The idea of database searching using MS data was taken further with the development of Peptide fragment fingerprinting, the identification of a peptide in a database using the MS/MS fragmentation spectrum of a peptide as the search parameter. This technique is very useful when inaccurate or small DNA sequence

stretches such as Expressed Sequence Tags (error rate 2%, average length 400 bases) and genomic sequences (error rate 0.1-5%, size 1.8-100 Mb) are being searched where normal protein fingerprinting data fails. One approach to the correlation of tandem mass spectral data of peptides to sequences in a protein database was developed by John Yates [29]. MS/MS spectra are automatically stripped from an auto-MS/MS HPLC run and the mass of each parent ion is used to search for isobaric amino acid stretches in the database. The experimentally determined spectrum is compared with the spectrum predicted from the sequence and the best matches are subject to cross correlation analysis. Matthias Mann took a different approach to peptide identification using peptide sequence tags [30]. The algorithm can handle non-standard amino acids which do not have to be specified. The spectrum must be manually inspected to find a group of ions which form a series from which a small sequence (the tag) can be read and used with the intact peptide mass, and the tag sequence start and end masses, to search the databases.

THE PROTEOME APPROACH

Despite the sensitivity limitations imposed by Edman degradation, several extensively annotated 2D gel databases have been built up in combination with other techniques for protein identification such as comigration with a known protein, gene inactivation or coexpression, and antibody recognition. One of the first projects to attempt to systematically map all of the spots on a 2D gel to a genomic sequence, that of *Escherichia coli*, was begun in the laboratory of Frederick Neidhardt in 1978 [31] and now ca 460 gene products have been identified from the estimated 4,100 ORFs. The availability of complete genome sequences and the new mass spectrometric identification methods will allow the construction of complete 'proteome maps'. Several projects have been announced: *S. cerevisiae*, Centre for Proteome Analysis, Denmark; *Cyanobacterium synechocystis*, Kazusa DNA Research Institute, Japan; *Bacillus subtilis*, Ernst-Moritz-Arndt-Universität, Germany; *Methanobacterium thermoautotrophicum*, Swiss Federal Institute of Technology, Switzerland and *M. genitalium*, Centre for Proteome Research, Australia.

However, there is one major caveat, the proteome coverage that can be achieved using 2D gel analysis must be defined. Up to 10,000 protein spots can be seen on large format 2D gels [7]. This can be extended by using prefractionation techniques such as the preparative isolation of cell organelles like nucleii and mitochondria. Large increases in sample loading with IPG gels [ ] and resolution can be achieved by dividing up the pH separation range of the ca 20 cm first dimension (2-12) into small overlapping pH windows (3-4, 3.5-4.5 etc.). The separation is reproducible enough to allow the windows to be put back reconstruct a proteome map over a metre and a half long. One must also define the amount of the genome that is being expressed at any one time since in the case of both procaryotes and eucaryotes large portions of the genome are only expressed at specific times such as during host infection or embryogenesis respectively. A detailed comparison of the mRNAs being expressed with the level of their translation products will be necessary in order to answer this question. Recently two new techniques have been described, the microarray assay for gene expression [33] and the Serial Analysis of Gene Expression (SAGE) which allow the expression levels of thousands of genes to be monitored simultaneously [34]. Plans are already underway to mass produce microarrays containing all the ca. 6,500 genes of S. cerevisiae in order to monitor its total gene expression. This approach combined with proteome analysis will allow a correlation of gene activity and protein expression in a cell at any one moment.

These new methods for proteome and gene expression analysis are quantitative and will allow new systematic approaches to investigation the function and regulation of unknown genes. N.L. Anderson has defined three major areas for the analysis of gene function and regulation [35]: molecular anatomy (protein composition of cells and tissues); molecular pathology (analysis of disease in terms of changes in protein expression and modification); and molecular pharmacology/toxicology (the effects of drugs and xenobiotics on protein expression and modification). A fourth area, molecular physiology, can be added, the change in protein expression in response to changes in the cells micro- or macro-environment. For example, *E. coli* usually obtains sulphur in the form of inorganic sulphate and assimilates it by reduction to sulphide which is then used for cysteine biosynthesis. In the soil however, a high percentage of sulphur is present in an organic form such as sulphate esters, sulphamates, amino acids and sulphonic acids.

We have recently described the identification of, and assigned possible functions to four previously unidentified ORFs in the *E. coli* genome [36]. By comparing 2D gels of *E. coli* grown on minimal medium in the presence of inorganic sulphate or ethanesulphonate as the sole sulphur sources, a set of eight proteins were seen to be upregulated in the presence of ethanesulphonate (Figure 3). All the proteins were identified by MS analysis. Of the four known proteins, three, sulphate binding protein, *2; cysteine synthase A, *5; and cystine binding protein, *7, belong to the cys regulon whereas the fourth was found to be a general stress protein, alkylhydroperoxide reductase, *8. Proteins *1 and *3 occur in the 8.5′ region of the genome, which contains four open reading frames with a single promoter region. Proteins *4 and *6 were found in the 21.5′ separated by the ycbE gene, an ORF encoding a hypothetical ABC transporter protein. Both 8.5′ and 21.5′ regions show many features common to ABC-type transport operons;

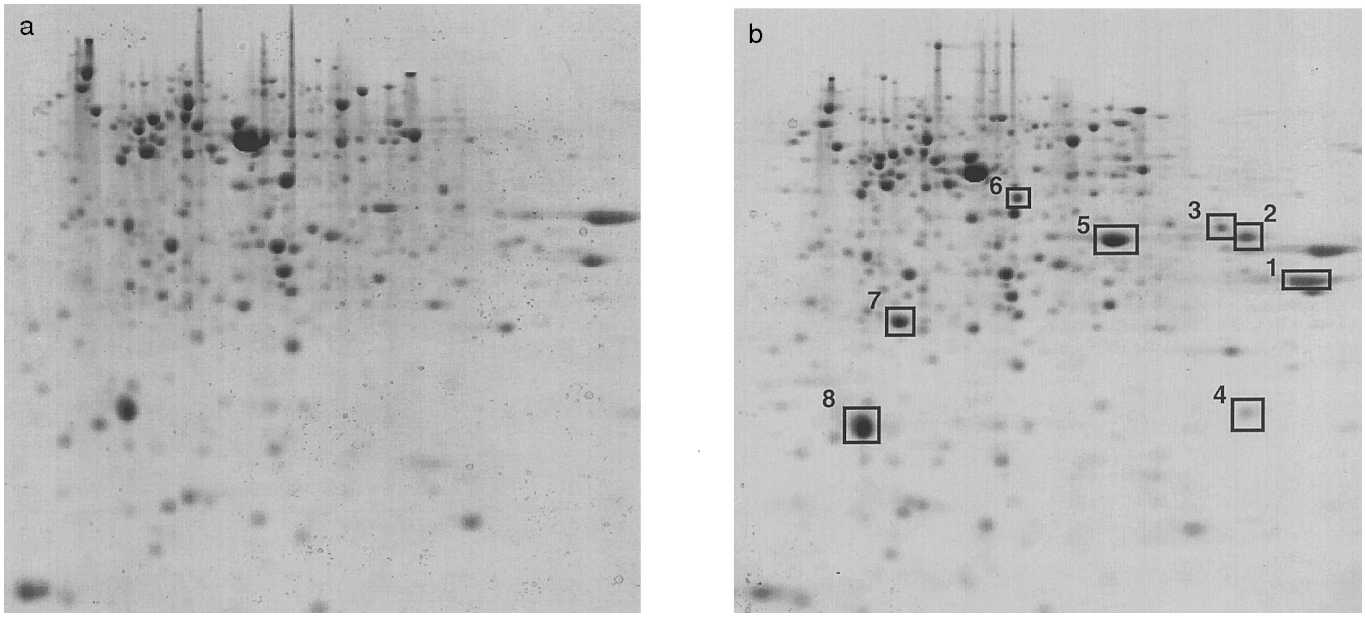

FIG. 3. Two dimensional gel electrophoresis mapping of *Escherichia coli. E. coli* was grown with (a) 500 μM sulphate or (b) 500 μM ethanesulphonate as the sole sulphur source. The proteins which were induced during growth with ethanesulphonate by more than a factor of $\times$1.7 are labelled [reproduced from M. Quadroni, W. Staudenmann, M. Kertesz, and P. James (1996) *Eur. J. Biochem.* **239,** 773–781 with permission].

a periplasmic substrate binding protein, a channel forming membrane protein and a cytoplasmic nucleotide binding protein. Promoter analysis and gene disruption experiments showed the four 8.5′ ORFs are probably a sulphate regulated, taurine uptake operon [37], whereas the 21.5′ ORFs may form a sulphate regulated, alkylsulphonate uptake operon.

OUTLOOK

The next step in the post-genome era must be the allocation of functions to the many new ORFs that have been discovered. This will require a quantitative analysis of both the transcripts (perhaps using the SAGE or microarray methodology) and their products (proteome analysis by mass spectrometry). The first steps to systematically assign functions to ORFs on a large scale has been announced: the goal of the EUROFAN (European Functional Analysis) project is to elucidate the physiological and biochemical functions of all the newly discovered ORFs in *S. cerevisiae.* Roughly 30% of the ORFs have a known function, a further 20-30% can be tentatively assigned a function based on similarity to other proteins of known function whilst the remaining ca. 40% remain unknown. EUROFAN is intended as a pilot project to test the feasibility of allocating functions to ORFs by systematically deleting, one by one, each of the genes of chromosome III and analysing the mutants. The first results of this exciting project were presented by Stephen Fey at the 2D Electrophoresis meeting in Siena last year [38]. A particular elegant method which complements the DNA and protein based gene identification approaches, is a genetic approach based on mapping DNA and protein polymorphisms. Joachim Klose and co-workers have presented results from interspecies backcross experiments between the two mouse species, *Mus musculus* and *Mus spretus* [39]. For each of the 982 animals obtained in the B1 generation, the genotype of 78 polymorphic DNA markers was determined. Using these markers, the genetic segregation, both qualitative and quantitative, of protein coding regions were mapped based on the 2D gel analysis of protein fractions from each of 5 different organs from 64 B1 animals. The brain proteins showed 1076 polymorphisms out of the 8458 proteins observed per gel, of which, genes for 258 proteins (548 spots) could be mapped.

Proteome analysis is not just of academic interest but is commercially useful. Proteome Inc. (Beverly, MA), has been founded by James Garrels to map the complete yeast Proteome and to provide a yeast protein identification service. The Large Scale Biology Corporation (Rockville, MD) has been using 2D gels of rat liver protein extracts to investigate the molecular pharmacology and toxicology of drug effects on targets. For example, lovastatin (produced by Merck as Mevacor) is designed to lower plasma cholesterol levels by inhibiting 3-hydroxy-3-methylglutaryl (HMG) CoA reductase. Analysis showed that HMG CoA reductase and synthase are both upregulated but that the main cholesterol lowering effect is due to the upregulation of the

low density lipoprotein receptor which increase cholesterol removal from the blood [40]. Thus by unravelling regulatory networks new drugs may be developed which regulate the target protein by interacting with a second coregulated one.

Maybe now we are verging on the edge of being able to harness the flood of information coming from the genome projects, to put it in order using proteome and microarray/SAGE projects, in a way that we may finally see how all the fine threads are pulled together to make the biochemical web which defines life. As an amateur detective once succinctly put it:

> "My dear fellow," said Sherlock Holmes [41], "life is infinitely stranger than anything which the mind of man could invent. If we could fly out of that window and hover over this great city, gently remove the roofs, and peep in at the queer things which are going on, the strange coincidences, the plannings, the cross-purposes, the wonderful chain of events, working through generations, and leading to the most outré results, it would make all fiction with its conventionalities and foreseen conclusions most stale and unprofitable."

ACKNOWLEDGMENTS

The author thanks Paola Dainese, Michael Kertesz, Manfredo Quadroni, and Werner Staudenmann for helpful discussions and critical reading of the manuscript.

REFERENCES

1. Sanger, F., *et al. Nature* **265,** 687–695.
2. Fleischmann, R. D., *et al.* (1995) *Science* **269,** 496–512.
3. Fraser, C. M., *et al.* (1995) *Science* **270,** 397–403.
4. Bult, C. J., *et al.* (1996) *Science* **273,** 1058–1073.
5. Goffeau, A. *et al.* (1996) *Science* **274,** 546–567.
6. O'Farrell, P. H. (1975) *J. Biol. Chem.* **250,** 4007–4021.
7. Klose, J., and Kobalz, U. (1995) *Electrophoresis* **16,** 1034–1059.
8. Taylor, J., Anderson, N. L., Scandora, A. E., Villard, K. E., and Anderson, N. G. (1982) *Clin. Chem.* **28,** 861–866.
9. Appel, R. D., *et al.* (1991) *Electrophoresis* **12,** 722–735.
10. Bjellqvist, B., Ek, K., Righeti, P. G., Gianazza, E., Görg, A., Westermeier, R., Postel, W. (1982) *J. Biochem. Biophys. Methods* **6,** 317–339.
11. URL http://expasy.hcuge.ch/ch2d/2d-index.html.
12. Lemkin, P. Presentation at the Siena 2D Electrophoresis Meeting (to be published as a special edition of *Electrophoresis* in March 1997).
13. Aebersold, R. H., Teplow, D. B., Hood, L. E., and Kent, S. B. (1986) *J. Biol. Chem.* **261,** 4229–4238.
14. Bauw, G., De Loose, M., Inze, D., Van Montagu, M., and Vandekerckhove J. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84,** 4806–4810.
15. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84,** 6970–6974.
16. Celis J. E., *et al.* (1990) *Electrophoresis* **11,** 989–1071.
17. Karas, M., Bachmann, D., and Hillenkamp, F. (1985) *Anal. Chem.*, **57,** 2935.
18. Whitehouse, C. D., Dreyer, R. N., Yamashita, M., Fenn, J. B. (1985) *Anal. Chem.* **57,** 675–681.
19. Vorm, O., and Mann, M. (1994) *J. Am. Soc. Mass Spec.* **5,** 955–958.
20. Figeys, D., Ducret, A., Yates, J. R., and Aebersold, R. (1996) *Nature Biotechnology*, **14,** 1579–1583.
21. Henzel, W., Billeci, T., Stults, J., Wong, S., Grimley, C., and Watanabe, C. (1993). *Proc Natl Acad Sci USA* **90,** 5011–5015.
22. Pappin, D., Hojrup, P., and Bleasby, A. (1993). *Current Biology* **3,** 327–332.
23. Mann, M., Hojrup, P., and Roepstorff, P. (1993). *Biol. Mass Spec.* **22,** 338–345.
24. Yates, J. R., Speicher S., Griffin P. R., Hunkapillar T. (1993). *Anal. Biochem.* **214,** 397–408.
25. James, P., Quadroni, M, Carafoli, E., and Gonnet, G. (1993) *Biochem. Biophys. Res. Comm.* **195,** 58–64.
26. James, P., Quadroni, M, Carafoli, E., and Gonnet, G. (1994) *Protein Science* **3,** 1347–1350.
27. Hunt, D. F, Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986) *Proc Natl Acad Sci U S A.* **83,** 6233–6237.
28. Kaufmann, R., Spengler, B., Lützenkirchen, F. (1993) *Rapid Commun. Mass Spectrom.* **7,** 902–910.
29. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) *J. Am. Soc. Mass Spec.* **5,** 976–989.
30. Mann M., and Wilm, M. (1994) *Anal. Chem.* **66,** 4390–4399.
31. Pederson, S., Bloch, P. L., Reeh, S., and Neidhardt, F. C. (1978) *Cell* **14,** 179–190.
32. Rabilloud, T., Valette, C., Lawrence, J. J. (1994) *Electrophoresis* **15,** 1552–1558.
33. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1996) *Science* 467–470.
34. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1996) *Science* 484–487.
35. Anderson, N. L. Presentation at the Siena 2D Electrophoresis Meeting (to be published as a special edition of *Electrophoresis* in March 1997).
36. Quadroni, M., Staudenmann, W., Kertesz, M., and James, P. (1996) *Eur. J. Biochem.* **239,** 773–781.
37. Van der Ploeg, J., Weiss, M., Saller, E., Nashimoto, H., Saito, N., Kertesz, M., Leisinger, T. (1996) *J. Bacteriol.* **178,** 5438–5446.
38. Fey, S. Presentation at the Siena 2D Electrophoresis Meeting (to be published as a special edition of *Electrophoresis* in March 1997).
39. Klose, J. Presentation at the Siena 2D Electrophoresis Meeting (to be published as a special edition of *Electrophoresis* in March 1997).
40. Anderson, N. G., and Anderson, N. L. (1996) *Electrophoresis* **17,** 443–453.
41. Doyle, A. C. (1930) The Complete Sherlock Holmes, pp. 190, Penguin Books, London.