Quarterly Review of Biophysics **30**, 4 (1997), pp. 279–331. Printed in the United Kingdom

Protein identification in the post-genome era: the rapid rise of proteomics

PETER JAMES

Protein Chemistry Laboratory, Swiss Federal Institute of Technology, Universitätsstrasse 16, CH-8092 Zürich, Switzerland

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

1. INTRODUCTION

1.1 *Innovation, automation and miniaturization*

Most advances in biology can usually be traced back to the development of a new technique: the recent explosion in sequence information in the databases arose from the pioneering work on separation methods by Frederick Sanger which paved the way for the development of protein (Sanger, 1945) and DNA/RNA (Maxam & Gilbert, 1977; Sanger, 1981) sequencing and culminated in the receipt of two Nobel prizes by Sanger. The initial phase of sequence database expansion was slow due to the tedious and slow nature of protein sequencing. Peptide sequencing was carried out manually and the complete analysis of a protein was tiresome, requiring the isolation of sufficient peptides from several digests of the target protein using proteases of different specialities to collect an overlapping set of fragments which cover the whole sequence. Protein sequencing gained momentum when the phenylisothiocyanate sequencing chemistry developed by Edman in 1949 was automated (Edman & Begg, 1967) and a commercial instrument requiring lower amounts (nanomoles) of sample was put on the market. Further technical advances such as novel valves to deal with small volumes of aggressive chemicals, the introduction of high pressure liquid chromatography (HPLC), and novel supports for sample immobilization, were all combined in the first gas phase sequencers, greatly increasing the sensitivity and allowing automated data collection (Hewick *et al.* 1981) and analysis. The new instruments with a sensitivity in the low picomole range appeared as rapid advances in DNA technology such as the development of restriction mapping (Danna *et al.* 1973), cloning (Cohen *et al.* 1973) and the dideoxynucleotide sequencing chemistry were threatening to make protein chemistry a relic of the past (Malcolm, 1978).

The second phase of database expansion had begun. Since DNA sequencing did not require expensive instrumentation, it was open to anyone who could run a gel

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

and buy a kit of enzymes. Instead of years, sometimes decades of painstaking work to isolate the protein of interest, the development of expression cloning allowed the isolation of the piece of DNA coding for the activity of interest, sometimes in a matter of months, and the entire coding region and hence the protein sequence could be determined in a similar amount of time. The DNA approach seemed much more fruitful, since not only did one get the entire sequence from a clone, the development of expression and mutagenesis systems allowed a systematic approach to dissecting the function of the protein. A new approach to biological problem solving was taking place. The initial euphoria over these methods overshadowed the more classical work being done in the 'mundane' areas of basic biochemistry such as cell signalling. This changed rapidly as the fields converged due in part to the literally encyclopaedic contributions of people like Shosaku Numa and Alfred Gilman in the areas of neurotransmitter receptor/membrane channels and G-proteins respectively (Gilman, 1987; Numa, 1989). One could sequence entire families of proteins and, by comparison, find areas of sequence which were conserved and thus could be inferred to be important for the function of this family. These areas were then targets for the subsequent mutation-function experiments.

1.2 *The genome era*

As with protein sequencing, DNA sequencing became rapid and simple and with the introduction of fluorescent dideoxynucleotides, the separation and data read-out could be fully automated. Sensitivity made a quantum leap with the introduction of the polymerase chain reaction technology (Saiki *et al.* 1988) for the amplification of DNA, allowing even single copy mRNAs to become accessible. Behind all this activity, the idea that entire pro- and eucaryotic genomes could be sequenced was slowly growing and gaining acceptance. The ball had been started rolling by the publication of the first complete sequence of an organism, virus ϕX174 (Sanger *et al.* 1977). In the early 1980s the determination of a physical genome map using the restriction enzyme patterns of all of the chromosomes of *Saccharomyces cerevisiae* was proposed, and the map of *Escherichia coli* was already under way. Rapid advances in robotics and automated DNA sequencing fuelled the formation of large-scale sequencing projects. Originally these projects were based on sequencing of small clones ($<$ 40 kb) derived from extensively mapped restriction fragments. The situation changed dramatically with the demonstration that shotgun sequencing of clones, from random DNA fragment libraries, was feasible for sequencing entire genomes. This approach was mainly catalysed by the work of Craig Venter with 'expressed sequence tags' (Adams *et al.* 1991). These tags (ESTs) are small cDNA fragments isolated and single pass sequenced from random primed cDNA libraries, which if the libraries are normalized, can provide a representative picture of what is being expressed in a tissue. Around two gels per day could be run with 20 lanes per gel yielding 450 bases per lane, thus generating 40 tags per DNA sequencer per day. As funding became available, labs with 10–20 sequencers could produce over 100 kb per day. Currently huge EST

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

libraries are under construction which now represent well over half of the total number of sequence records in the Genbank database. At first this data source was received with scepticism but now it has proved its worth; in finding new genes involved in human disease; for identifying exons in the vast expanses of genomic DNA and especially as a source of genetic based mapping reagents for the construction of physical chromosome maps.

The new automated sequencing, data collection and assembly was applied by Venter to the determination of the entire genome sequences of the bacteria *Haemophilus influenzae* and *Mycoplasma genitalium* which were published in 1995 (Fleischmann *et al.* 1995; Fraser *et al.* 1995). To achieve enough accuracy and coverage for genome sequencing, six fragments covering any particular stretch had to be sequenced. Huge quantities of sequence data for *H. influenzae* were rapidly generated using 14 DNA sequencers for 3 months giving a total of 11·34 Mb, six times the genome size. The final assembly of the sequence stretches required gap closure by non-automated approaches. These genomes opened the flood gates and now eight bacterial and archael genomes have been completed and 34 are under way. The first eucaryotic genome *S. cerevisiae* (14 Mb) was completed in 1996, and the most ambitious, the human genome (3000 Mb) project, is tagged for completion in 2004. The bottleneck has become data management and analysis.

2. THE PROTEOME

2.1 *Two-dimensional gel electrophoresis*

The idea of analysing the complete complement of proteins being produced by a cell arose over 20 years ago with the development of two-dimensional (2D) gel electrophoresis. Kenrick & Margolis (1970) combined native isoelectric focusing with pore gradient sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS–PAGE) to obtain a separation of serum proteins. The 2D technique which is mostly used today originated from the work of Patrick O'Farrell (1975) and Joachim Klose (1975). The tremendous resolving power and sensitivity of the technique and the ability to combine it with other methods by electroblotting to insert supports for testing with antibodies or for Edman sequencing to identify proteins, has allowed the construction of 'cell' maps (Pederson *et al.* 1978; Celis *et al.* 1990).

The technique is based on the separation of proteins according to their isoelectric points, pI, in the first dimension, either O'Farrell type isoelectric focusing (IEF) if the pH gradient is maintained dynamically with ampholytes or immobilized pH gradient electrophoresis (Bjellqvist *et al.* 1982) if the gradient is covalently immobilized in a gel matrix. The proteins are then transferred to the second dimension SDS–PAGE where they are further separated according to mass (see Fig. 1). By using large format gels (40 × 30 cm as opposed to the 'standard' 20 × 20 cm format) 10000 proteins have been resolved from mouse testis extracts (Klose & Kobalz, 1995). The only drawback to this type of analysis is that membrane proteins are poorly represented due to solubility problems occurring during the transfer between the two dimensions.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

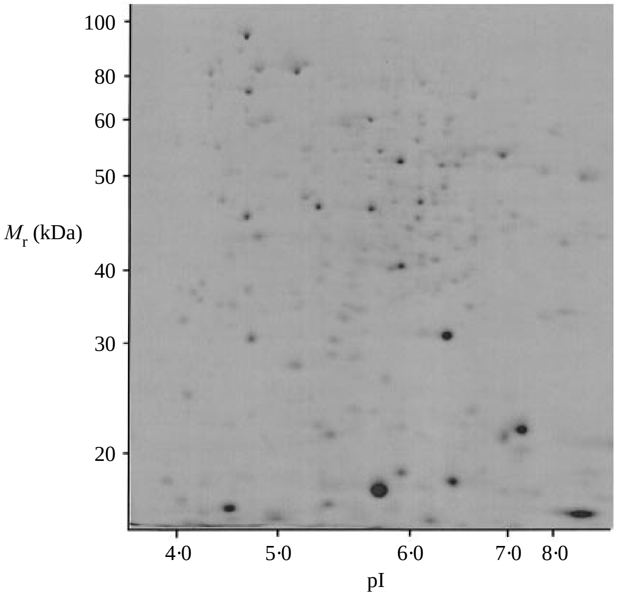


Fig. 1. Protein separation by two dimensional gel electrophoresis. A 2D gel of proteins extracted from the bacterium *B. japonicum*. The cell suspension was centrifuged and the cells disrupted by passage through a French press. The proteins were denatured and solubilized in a buffer containing 8 M urea before loading on the first dimension IEF strip. The proteins were focused for 100000 volt hours before transferring to the second dimension SDS–PAGE gel for separation according to size. The gel was then stained with the dye Coomassie blue to visualize the proteins.

The main advantages of 2D gels is their huge resolving power and their dynamic range for protein detection allowing the gels to be analysed both qualitatively and quantitatively. Standard Coomassie blue stained gels can detect proteins down into the subpicomole range while silver staining extends this by one or two orders of magnitude, and the latest fluorescent stains could take the detection limit down to the high attomole limit. Since the binding of these dyes is to a large extent proportional to the proteins size and independent of the sequence, proteins can be quantitated. Scanned images of 2D gels of cells in different stages can be analysed by computer and used to quantitate the changes occurring in protein expression (Taylor *et al.* 1982). The main problem in 2D PAGE was the qualitative description of the spots, how to link a numbered spot on a computer image to an entry in a sequence database. The methods that are evolving form the centre of interest of this review which could be subtitled as 'Can proteome analysis be made compatible with mRNA expression analysis in terms of speed and sensitivity?'

2.2 *Applications of 2D PAGE*

2D gel protein databases offer a systematic approach to the study of complex processes such as cell proliferation and differentiation. Many factors cannot be taken into account by an analysis of the mRNA expression levels in a cell:

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

since the mRNA level for a protein may not reflect the amount of that protein in the cell because this depends on the relative turnover rates of the mRNA and protein and the rate of translation. It is not only the level of protein expression but the degree to which temporal covalent modifications such as glycosylation and phosphorylation (to name but two from hundreds occurring in the cell, see the delta mass web site for a compilation of post-translational modifications at

http://www.medstv.unimelb.edu.au/
WWWDOCS/SVIMRDocs/MassSpec/deltamassV2.html).

Since many processes such as cell signalling do not involve protein synthesis in the short term response and are dominated by phosphorylation, mRNA analysis will not be of much use in dissecting the signalling processes involved. 2D PAGE allows these modifications to be monitored and analysed in combination with mass spectrometry.

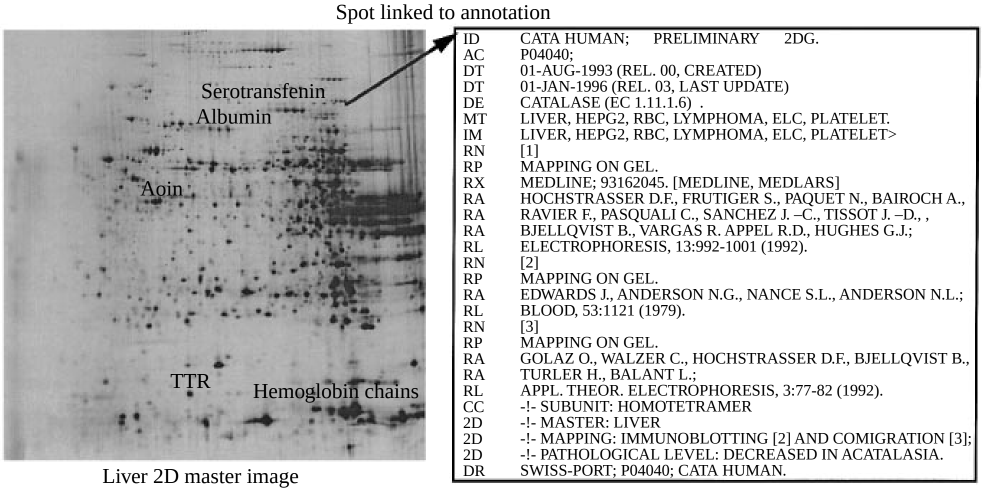
Comparative protein maps of cells and tissues in normal and pathological states are being developed for use as a diagnostic tool, for example the MELANIE project of Denis Hochstrasser in Geneva, Switzerland, in which the 2D gel is envisaged as a kind of molecular scanner used to monitor changes in disease to allow early diagnosis from biopsies (Appel *et al.* 1991).

2D PAGE analysis has been applied clinically to distinguish between healthy and tumour tissue from mammary epithelium (Steinbeck *et al.* 1984). Two prominent spots were found which were present in all malignant tissues examined and absent in all eight non-neoplastic tissues. 2D gel technology is being applied in many other fields of medicine; in the molecular epidemiology of viruses and bacteria (Cash, 1991), to determine changes in protein synthesis patterns in murine organs during post-implantation development (Praxmayer *et al.* 1992) and in studying the immune response (Kovarova *et al.* 1992). 2D PAGE is also finding extensive use in testing the effects of toxic xenobiotics (Anderson *et al.* 1986). Liver proteins of male C57BL/6T mice treated with 0, 50 or 250 mg kg^{-1} Aroclor 1254 were analysed and the resulting patterns were processed using a computerized imaged analysis system and quantitative data selected for a total of 150 protein spots. Thirty-one proteins were found that showed quantitative differences attributable to treatment with chlorinated hydrocarbons. Similar results can be obtained from analogous tissue culture experiments, suggesting that one can cut down drastically on the number of animals required for toxicity testing.

2.3 *Global analysis of protein expression, the Proteome approach*

The availability of complete genome sequences for organisms will allow the construction of complete protein maps. The 'Proteome' is a term that has recently been coined by Marc Wilkins (Wilkins *et al.* 1995) to describe the total set of proteins encoded by a genome. The main prerequisites for the construction of 2D gel databases are: a highly reproducible 2D gel pattern, the ability to identify spots, and a software system for analysing the gels, calculating the spot densities

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press



```
ID     CATA HUMAN;     PRELIMINARY     2DG.
AC     P04040;
DT     01-AUG-1993 (REL. 00, CREATED)
DT     01-JAN-1996 (REL. 03, LAST UPDATE)
DE     CATALASE (EC 1.11.1.6)  .
MT     LIVER, HEPG2, RBC, LYMPHOMA, ELC, PLATELET.
IM     LIVER, HEPG2, RBC, LYMPHOMA, ELC, PLATELET>
RN     [1]
RP     MAPPING ON GEL.
RX     MEDLINE; 93162045. [MEDLINE, MEDLARS]
RA     HOCHSTRASSER D.F., FRUTIGER S., PAQUET N., BAIROCH A.,
RA     RAVIER F., PASQUALI C., SANCHEZ J. –C., TISSOT J. –D., ,
RA     BJELLQVIST B., VARGAS R. APPEL R.D., HUGHES G.J.;
RL     ELECTROPHORESIS, 13:992-1001 (1992).
RN     [2]
RP     MAPPING ON GEL.
RA     EDWARDS J., ANDERSON N.G., NANCE S.L., ANDERSON N.L.;
RL     BLOOD, 53:1121 (1979).
RN     [3]
RP     MAPPING ON GEL.
RA     GOLAZ O., WALZER C., HOCHSTRASSER D.F., BJELLQVIST B.,
RA     TURLER H., BALANT L.;
RL     APPL. THEOR. ELECTROPHORESIS, 3:77-82 (1992).
CC     -!- SUBUNIT: HOMOTETRAMER
2D     -!- MASTER: LIVER
2D     -!- MAPPING: IMMUNOBLOTTING [2] AND COMIGRATION [3];
2D     -!- PATHOLOGICAL LEVEL: DECREASED IN ACATALASIA.
DR     SWISS-PORT; P04040; CATA HUMAN.
```

Fig. 2. 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 further link to the entry in the SwissProt database. The white areas indicate spots for which data are available.

and providing a link from these to a database containing fields for the entry of information pertinent to that spot. Several gel analysis and database programs were (and are still evolving) developed such as those from Garrels (1979) PDQuest, Lemkin & Lipkin (1981) GELLAB, Anderson *et al.* (1981) TYCHO (later Kepler), and Appel *et al.* (1991) MELANIE. Anderson & Anderson (1982) put forward the idea of the 'human protein index'. This was to be a database of all the spots resolved by 2D PAGE (from liver) together with structural, functional and clinical information. The main drawback was that although 2D gels were reproducible within a certain laboratory, inter-laboratory gel comparisons were almost impossible. This has changed now with the introduction of immobilized pH gradient gels as the first dimension and the development of powerful software which even allows gels from different labs to be compared interactively over the internet (Lemkin, 1997).

One of the first and to date most complete projects to attempt to map systematically all of the spots on a 2D gel to sequences in a genome was begun in the laboratory of Frederick Neidhardt (Pederson *et al.* 1978) with *E. coli*. Other databases followed, and recently a standard format (see Fig. 2) has been proposed to create a federation of 2D gel databases which are WWW accessible (Appel *et al.* 1996). The following 'Federated' 2D databases are available on the web: SWISS-2DPAGE (Geneva University Hospital); HSC-2DPAGE (Heart Science Centre, Harefield Hospital); HEART-2DPAGE (German Heart Institute, Berlin); HP-2DPAGE (Heart 2-DE Database, MDC, Berlin); PDD Protein Disease Database (NIMH-NCI) as well as a list of partly federated databases including the Yeast

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Protein Database at Proteome, Inc. and the extensive rat liver maps at Large Scale Biology Corp. A list of links to the various sites is maintained at the University of Geneva ExPaSy server at the URL: http://expasy.hcuge.ch/.

With the advent of rapid methods for protein analysis by mass spectrometry progress is being rapidly made (Shevchenko *et al.* 1996; Wasinger *et al.* 1995) and several projects have been announced: *H. influenzae*, Hoffmann–La Roche, Basel, Switzerland; *S. cerevisiae*, Centre for Proteome Analysis, Odensee, Denmark; and *M. genitalium*, Centre for Proteome Research, Eveleigh, Australia. Proteome centres are being established world wide and Biotechnology companies such as Oxford Glycosciences are investing in methods to develop the tools to allow large-scale proteomics.

The major challenge to come from the genome projects, to classify genes into classes and to allocate functions to them, can be approached in a systematic way using proteomics to study the effects of gene mutations. The main strength of proteomics lies not in a massive cataloguing of protein spots, but in allowing proteins to be placed into functional classes by following changes in their expression and post-translational modification levels. An integrated approach of subtractive 2D mapping together with techniques to determine mRNA expression patterns such as SAGE, will allow a cohesive analysis of changes occurring during complex processes to be made, in order to select genes for further study. The group of Hanash (Wimmer *et al.* 1996) have described an approach combining 2D genomic and proteomic approaches to the investigation of changes occurring in cancer. Proteins were identified which show increased expression and were associated with differentiation and proliferation properties of neuroblastoma cells. In parallel, 2D patterns of DNA fragments from genomic digests were analysed and amplified DNA fragments were observed which commonly contain growth related genes. This approach is now being studied at the mRNA level using microarray hybridization. Thus soon a complete picture of all the events involved in a particular disease state will be able to be defined.

3. CHEMICAL ANALYSIS

3.1 *N-terminal*

Currently the most popular method of protein identification is N-terminal sequencing of proteins using the chemistry developed by Per Edman (Edman, 1949). The chemistry is outlined in Fig. 3.

Whilst the chemistry has remained essentially unchanged for over four decades, the sensitivity has increased by many orders of magnitude. The first major advance after its introduction in 1949 was its automation by Edman & Begg in 1967 and the subsequent production of a commercial instrument, the spinning cup sequencer by Beckman which made the technology available outside of specialized laboratories. Since then, various peptide immobilization methods, covalent, entrapment or adsorption and the physical nature of the chemical deliveries, liquid, solid or gas phase have varied but the main advance has come from miniaturization of the plumbing of the instrument which lowers losses and allows

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

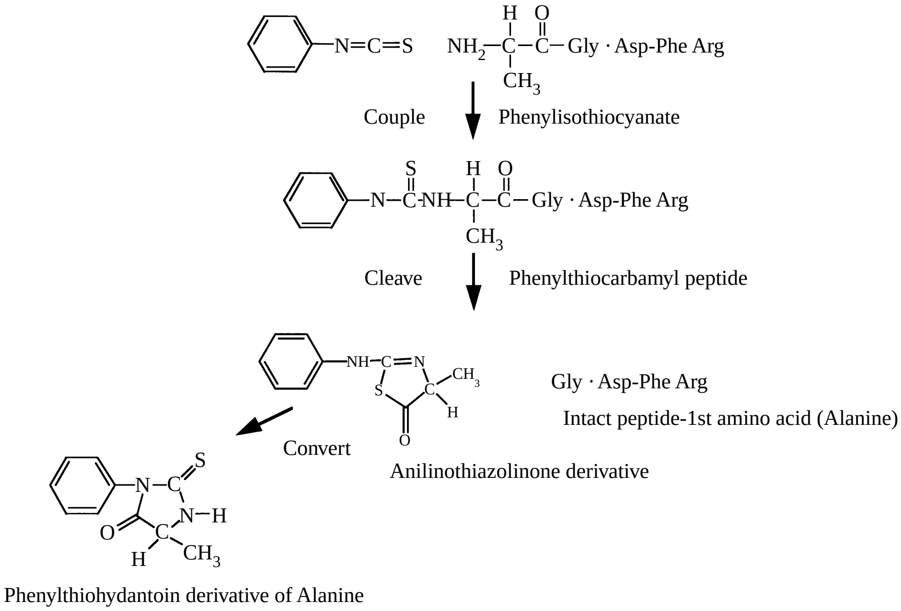

N=C=S
NH2–C–C–Gly · Asp-Phe Arg
H
O
CH3
Couple
Phenylisothiocyanate
S
N–C–NH–C–C–Gly · Asp-Phe Arg
Cleave
Phenylthiocarbamyl peptide
NH–C=N
CH3
S
C
H
O
Gly · Asp-Phe Arg
Intact peptide-1st amino acid (Alanine)
Convert
Anilinothiazolinone derivative
N–C
N–H
O
C
H
CH3
Phenylthiohydantoin derivative of Alanine

Fig. 3. Edman sequencing. The main steps in sequential Edman type N-terminal sequencing are shown here. The derivatization of the N-terminal under basic conditions with phenylisothiocyanate, cleavage of the first amino acid by strong acid and conversion of the anilinothiazolinone to the stable phenylthiohydantoin for subsequent HPLC identification.

coupling to ever more sensitive detection systems (current commercially available instruments can produce sequences at the 200 fmol level). Automation of reagent and solvent deliveries has allowed very high efficiencies of degradation to be achieved ($> 98\%$ at the 10 pmol level) which translates into longer sequences being read. The great advantage of the Edman chemistry is that the phenylthiohydantoin derivatives of the amino acids formed are UV absorbent and can be readily separated by simple reverse phase HPLC methods allowing for ease of automation and data collection.

The development of gas phase sequencing from non-covalent supports (glass fibres and various polymer membranes) revolutionized the sample preparation possibilities. In 1986 Aebersold *et al.* showed that proteins could be directly N-terminally sequenced by electroblotting from analytical SDS–PAGE gels onto activated glass fibre sheets which are stable to the conditions of gas phase Edman chemistry, and this was extended by the group of Vandekerckhove (Bauw *et al.* 1987) to the analysis of proteins electroblotted from 2D PAGE gels. Subsequently Aebersold *et al.* (1987) showed that internal sequence information could be obtained from proteins blotted onto nitrocellulose filters. The main development was the use of a polymer, polyvinylpyrrolidone (PVP-40) to block the filters, preventing the adsorption of the protease, and the use of nitrocellulose as a support which is hydrophobic enough to retain proteins during electroblotting but hydrophilic enough to allow peptide release for subsequent HPLC before

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

sequencing. Much effort has been made since then to improve supports for electroblotting and sequencing or digestion and has been reviewed by Aebersold (1990).

The ability to sequence proteins isolated directly from 2D gels provided the impulse to speed the generation of maps. Standard 2D gels visualized by silver staining show around 1–2000 spots, 75 % of which are in amounts less than 500 fmol. Large format gels allow the visualization of up to 10000 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. Much effort has been put into improving the sensitivity of the method using new coupling reagents. The first attempts were made to introduce fluorescent isothiocyanate derivatives but this never really turned into a viable practical method due to two main difficulties, HPLC separation of the amino acid derivatives is difficult due to the bulky hydrophobic nature of the fluorophore, and secondly, the bulky nature stearically hinders an efficient coupling and hence the efficiency of the sequencing drops. In order to circumvent these problems stearically unhindered isothiocyanates with a cryptic amino group were developed which show normal coupling rates and then subsequently deprotected and fluorescently labelled before cleavage of the thiocarbamyl derivative from the protein (Hood *et al.* 1986) as well as a labelling method using four-aminofluorescein for the azothiazolinone cleavage products (Tsugita *et al.* 1989). Neither of these new reagents (Fig. 4) found widespread application despite sensitivity levels in the low attomole range.

Recently attempts have been made to interface conventional chemical sequencers to mass spectrometers with electrospray ionization interfaces. The group of Aebersold (Aebersold *et al.* 1992; Bures *et al.* 1995) have described the synthesis of the non-stearically hindered Edman like reagent, 4-(3-pyridinylmethylaminocarboxypropyl) phenyl isothiocyanate, the amino acid thiohydantoin derivatives of which show much higher ionization efficiencies than those of the corresponding phenylthiohydantoin. A partial separation is achieved by a rapid HPLC gradient and the eluent is directed into the mass spectrometer for detection. The main drawback of the chemical approach, regardless of the detection used, is the background; both due to impurities in the reagents and solvents and also in the environment such as dust and fingerprints, etc., and the need to isolate individual peptides without great sample losses. Herein lies the great advantage of peptide sequencing by mass spectrometry.

3.2 *C-terminal*

The ability to sequence proteins from the C-terminal would be the ideal complement to the N-terminal approach, and would be especially useful for the many proteins which are N-terminally blocked. In the field of C-terminal sequencing nothing dramatic has changed since Schlack & Kumpf (1926) applied the method of Johnson & Nicolet (1911) for the conversion of acylamino acids to acylthiohydantoins to a small peptide. The method in its current form was

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

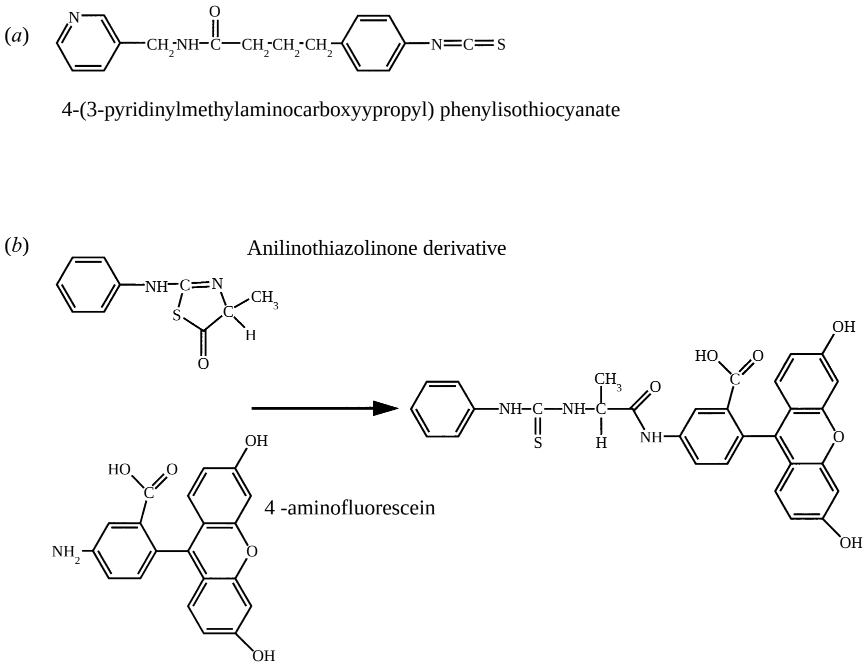


Fig. 4. Increasing the sensitivity of Edman sequencing. (*a*) Shows the alternative Edman based coupling reagent described by Bures *et al.* (1995) which is suitable for detection by an HPLC–MS detector. (*b*) Shows an alternative approach using the normal Edman chemistry but adding a fluorophore to the anilinothiazolinone as proposed by Tsugita *et al.* (1989).

developed by Stark (1968) and involves the activation of the C-terminal carboxyl group with acetic anhydride to form a mixed anhydride. The reaction scheme is outlined in Fig. 5.

This reacts with thiocyanate in the presence of a strong acid to give the thiohydantoin derivative which is cleaved under basic conditions. The thiocyanate method shows a lack of reactivity towards proline and aspartic acid and so the group of John Shively (Bailey & Shively, 1990) introduced a new coupling reagent, trimethylsilylisothiocyanate and trimethylamine as a cleavage reagent. This was further modified to diphenylphosphoroisothiocyanatidate and sodium trimethylsilanolate and incorporated into a fully automated instrument (Bailey *et al.* 1992). This chemistry sequenced through all amino acids except proline at high sensitivity (200 pmol) for up to four cycles. The coupling reagent reacts with the carboxyl group to form an acylthiohydantoin which cyclizes to form the thiohydantoins which loses the amide proton. Proline has no amide proton and thus regenerates the C-terminal proline upon cleavage. Inglis *et al.* (1992) showed that proline thiohydantoin can be synthesized, the key being the protonation of the proline thiohydantoin by acid which is then cleavable by water vapour. This was adapted by Bailey *et al.* (1995) to produce the first automated sequencer which can analyse all 20 amino acids. The only other promising method developed over the past decade was described by Boyd *et al.* in 1992. Again a thiocyanate based

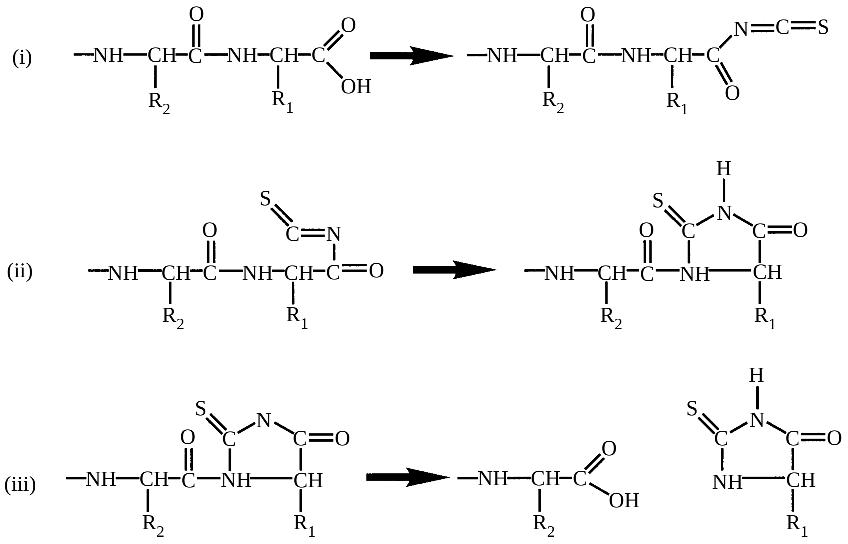


Fig. 5. An outline of C-terminal degradation according to Stark. Ammonium thiocyanate in trifluoroacetic acid and acetic anhydride is added to the peptide, the strong acid protonates the carboxyl group, activating it. It then reacts with the thiocyanic vapour to generate the thiocyanate derivative (i). The C-terminal thiocyanate rearranges spontaneously to give the C-terminal thiohydantoin under the acidic conditions (ii). The C-terminal thiohydantoin is subsequently released by treatment with gaseous trimethylamine (iii).

degradation was used but an S-alkylation of the hydantoin before cleavage was introduced. This improves the leaving group potential of the thiohydantoin and also allows the introduction of fluorescent or mass spectrometric markers to increase sensitivity. Currently commercially available sequencers can at best obtain two to three amino acid long sequences from small (50 pmol) amounts or longer *c.* ten residues from nanomolar amounts, though it has been shown that one can use the same sample for both N- and C-terminal sequencing. The ideal sequencing strategy for subsequent cloning would be to obtain 20–30 residues of sequence from both the N- and C-termini of the protein and to construct oligonucleotide primer to amplify the sequence from a cDNA bank by PCR. However, currently this is not possible and C-terminal sequencing will probably only find use in a quality control setting, showing that a protein product has the correct C-terminal and that no processing has taken place.

4. OTHER IDENTIFICATION METHODS

The major breakthrough in sensitivity and throughput for proteome analysis has come from the development of new techniques in mass spectrometry for the analysis of proteins and peptides. However, there have been many other approaches to protein identification on 2D gels that deserve mention. In order to define a protein, one must measure one or more parameters that are unique to the protein which allow one to search a database. Two of these properties, the

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

isoelectric point (pI) and the molecular weight are exploited by the 2D PAGE technique and so are available directly from the gel.

4.1 *Comigration and HPLC mapping*

One of the first methods of protein identification on 2D gels was comigration with a purified protein; for example, Schubart & Danoff (1987) identified a 19 kDa rat brain protein that comigrates on 2D electrophoresis with a previously sequenced protein p19. A more refined and large scale method for studying protein comigration on 2D gels for identification purposes has been described by Lefkovits *et al*. (1995). They analysed the 2D expression patterns of an ordered library of 4608 cDNA clones from the CEM human leucaemic cell line. The clones were arranged in a matrix array of 24 × 16 × 12, each containing one clone. Pools of clones were then made and co-electrophoresed according to a special concatenation scheme; e.g. pools 1+2+3 and 3+4+5 when compared contain at least those spots in common, which originate from pool 3). This approach was shown to be feasible and matching was completed for one half of the library. Another method of protein identification is by comigration of the tryptic peptides on HPLC with those of a standard protein. The series of closely related spots of P19 mentioned above were shown to be isoelectric variants (probably phosphorylated forms) since they produced almost identical HPLC traces after trypsin digestion. Peptide mapping by HPLC has been used quite often for protein identification, usually by comparison with a standard protein run at the same time. However, more complex mixtures have been analysed and a chromatographic method was developed for identification of protein species through their peptide patterns (Medina & Phillips, 1982). Protein isolates of beef, pork, chicken and soya were heated, enzymatically hydrolysed at optimal conditions and subsequently analysed by sequential application of TLC and HPLC. The chromatographic patterns of the tryptic peptides were then statistically analysed by discriminant analysis. Preliminary studies indicated that an all-beef frankfurter can be discriminated from a standard frankfurter containing 35% pork protein.

4.2 *Accurate pI and MW measurements*

By using the coordinates of a spot on a gel (pH, MW) one can narrow down the number of candidate proteins which can correspond to this spot. The ExPaSy server at Geneva has a program that allows a boundary to be drawn around the potential area on a 2D gel image in which the desired protein from the database may be found. This is a fairly large area and does not serve to define a protein uniquely to a spot, much more accurate determinations of the two parameters are necessary. However, the immobilized pH gradient technique can be used to determine the pI of a protein to within 0·001 pH units on 'zoom gels' (Bjellqvist *et al*. 1982). A zoom gel is a very narrow pH range gel covering say 0·5 pH units over a 20 cm separation range. This can be combined with an accurate molecular weight

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

determined by mass spectrometry of the protein electroblotted from the gel onto a PVDF support (Eckerskorn *et al.* 1992). This combination is accurate enough to allow unequivocal identification of proteins from organisms whose genomes have been determined.

4.3 *Amino acid composition and ratios*

Closely related to an accurate pI is the determination of the amino acid composition of a protein. This can be carried out on the PVDF and glass-fibre supports commonly used for N-terminal sequencing. This was first shown to be effective as a means for identifying proteins isolated by 2D PAGE by Manabe *et al.* in 1982. Human serum proteins were separated by 2D PAGE, the stained spots were punched out and the proteins in each piece of gel were extracted with 0·1 M sodium hydroxide–2 % thiodiglycol. The extracted proteins were hydrolysed and applied to an automated amino acid analyser. The method regained popularity when more sensitive methods for amino acid analysis were developed and have been successfully applied to the identification of mouse brain proteins isolated by 2D PAGE (Eckerskorn *et al.* 1988). The method is still useful for the more abundant proteins (the Coomassie blue stainable spots) and in combination with other techniques is useful for the cross-species identification of proteins from organisms whose genome is unavailable (Wilkins & Williams, 1997).

A potential very sensitive method which would allow thousands of proteins to be monitored simultaneously was put forward by Latter *et al.* in 1983. By using 20 different cultures of a cell line, each grown with a different ^{14}C- or ^{35}S-labelled amino acid, the amino acid composition of each of the proteins could be determined by computer analysis of the autoradiograms of the 2D gels (see Fig. 6).

A preliminary study that we carried out with Professor Gonnet at the ETH in Zurich showed that by determining the ratio of four pairs of amino acids, using ^{14}C (for any amino acid, usually glutamic acid or leucine) *vs.* ^{35}S (methionine or cysteine) the number of gels and cultures could be reduced to four. By analysing gel images scanned using a Phosphorimager every 5 days for 2 months, the radioactive decay curve for a particular spot could be measured and the isotope ratio extrapolated. An accuracy of 1 % is enough to identify the protein in the EMBL database and by restricting the search to a single bacterial genome, an accuracy of 5 % in determining the ratio of a single pair of isotopes, together with an estimate of molecular weight to within 2 % and pI ± 0·2 units would be enough to determine the identity of a protein. The main problem with this method is the metabolic scrambling of the label; if this could be controlled, this method would be one of the most powerful parallel methods for proteome analysis.

4.4 *Identification by antibody recognition*

One of the most commonly used protein identification methods is antibody recognition. The 2D gel is blotted onto a support membrane and blocked by a polymer or milk solution and then incubated with the antibody against the protein

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

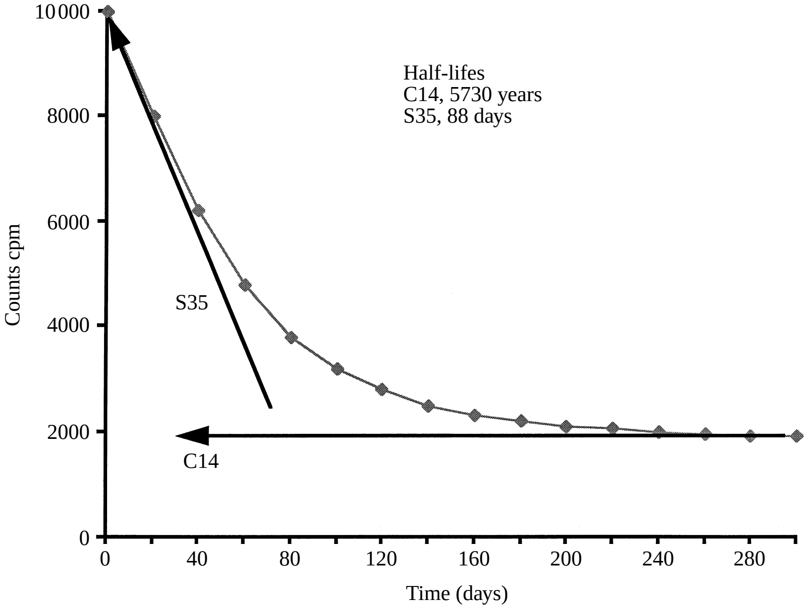


Fig. 6. Determining the amino acid composition of a protein by dual isotope labelling. A protein can be identified in a database search using the ratio of four pairs of amino acids, determined using ^{14}C (for any amino acid, usually glutamic acid or leucine) and ^{35}S (methionine or cysteine). Since the decay times are drastically different the relative amounts of each amino acid can be determined for each spot on the gel.

of interest. The binding of the primary antibody is then visualized by incubation with a secondary antibody directed against the first which carries a marker such as a fluorophore or an activity which can be used in a colour reaction. The main drawback of the method is trying to align the electroblot with the gel to see which spot has been recognized. Since entire genome sequences are available it would be fairly easy to produce a commercial library of monoclonal antibodies produced against each of the open reading frames expressed in another organism. Another possible method of interest would be the generation of antibodies against a single spot on a 2D gel using a phage display system (Clackson *et al.* (1991) and D. Neri, ETH Zürich, personal communication). The antibody produced would be useful in function studies as well as in isolating larger quantities of the target material. The protein could be identified by screening an ordered expression library.

4.5 *Genetic analysis*

A particularly elegant method which complements the DNA and protein based gene identification approaches, is a genetic approach to map the mouse genome based on mapping DNA and protein polymorphisms within the framework of the European Collaborative Interspecific Backcross (Breen *et al.* 1994). Results from interspecies backcross experiments between the two mouse species, *Mus musculus* and *Mus spretus* have been presented, for each of the 982 animals obtained in the

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

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 five 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.

5. MASS SPECTROMETRY

5.1 *Advances in ionization methods*

Up until the late 1970s peptides and proteins were not generally amenable to analysis by mass spectrometry due to low volatility and large size. Extensive chemical modification was necessary to produce either N, O, permethylated peptides (Vilkas & Lederer, 1968; Rose *et al.* 1983*a*) or reduction to polyamino alcohols (Biemann *et al.* 1959). These techniques were used successfully to sequence peptides and in some cases stretches of proteins. However, the major breakthrough came with the development of fast atom bombardment (FAB) by Barber *et al.* in 1981. This method used a neutral beam of argon to ionize peptides in a glycerol matrix and had a mass range of up to around 12000. This ionization procedure although relatively 'soft' can cause enough fragmentation that a set of ions can be identified which define the sequence of the peptide. Other ionization techniques such as plasma desorption and field desorption were developed but never found the widespread acceptance of FAB and its derivatives. Recently two soft ionization methods have been (re)introduced to revolutionize protein and peptide analysis; laser desorption and ionization and electrospray ionization.

5.1.1 *Electrospray ionization (ESI)*

In 1968, Dole *et al.* proposed using a very fine spray of solvent containing the molecule of interest to introduce sample into a mass spectrometer. This was taken up in various forms in the late 1980s as electrospray and ionspray ionization. The technique involves ionization of proteins or peptides at atmospheric pressure by nebulizing a flowing stream of solvent under a potential difference of several thousand volts between the sample exit tip and the MS entrance (Fenn *et al.* 1989; Covey *et al.* 1988).

The ions are desolvated either by passage through a heated capillary or with a counter current of gas before entering the high vacuum region of the MS (Fig. 7). The interface can easily be coupled directly to the effluent of an HPLC or with slight modification to a CZE apparatus. The technique is most often used coupled to triple quadrupole (TSQ) and ion trap (IT) mass spectrometers (as well as a few four sector magnetic and Fourier transform ion cyclotron instruments). These mass spectrometers have a limited mass range up to around 4000 m/z. However, the electrospray ionization process produces multiply positively charged protein and peptide ions when spraying from acidic solutions. Since the TSQ an IT–MS instruments measure m/z (for excellent explanations of the operating principles of these instruments, see Miller & Denton, 1986; Cooks *et al.* 1991, respectively) and

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

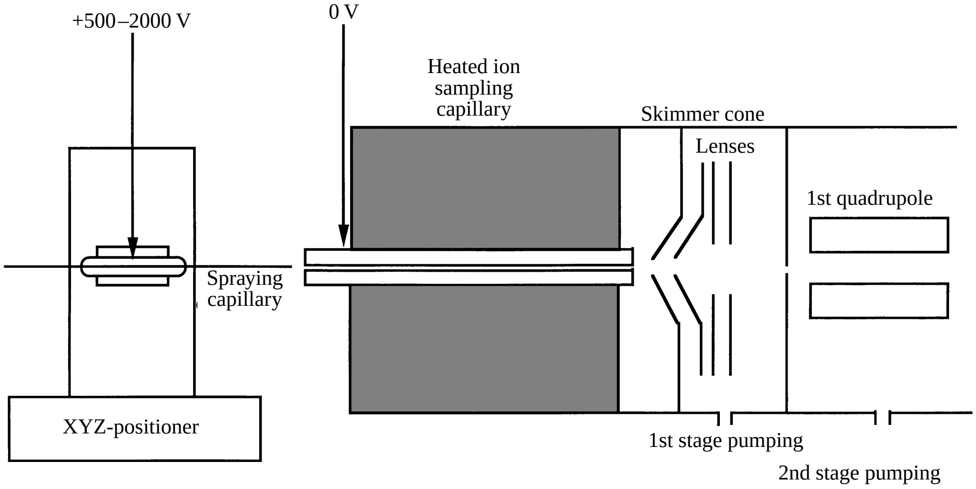


Fig. 7. A schematic diagram of an electrospray source. The scheme shows the set-up of a nanospray electrospray source. The capillary is pulled to a narrow point with an orifice *c*. 1–5 μm in diameter and is covered with a conductive film, usually gold and a potential of around +500 V is placed on it. The capillary tip is aligned 3 mm from the mass spectrometer inlet orifice using an XYZ positioner. The fine spray enters the heated capillary and the fine droplets evaporate leaving a beam of macromolecules to enter the mass spectrometer through two differentially pumped chambers.

on average one charge per kDa is observed (protonation of the N-terminus, and the side chains of lysine, arginine, proline and histidine), protein masses over 150000 have been determined with an accuracy of 0·005%. The electrospray technique is now approaching Dole's original concept of a 'beam of macromolecules' with ultra low flow sample introduction through capillaries drawn to 1–10 μM tips (Gale & Smith, 1993; Andren *et al.* 1994; Wilm & Mann, 1994; Kriger *et al.* 1995; Valaskovic *et al.* 1996). This has allowed an extremely efficient ionization of analytes, virtually all of which enter the mass spectrometer for analysis, allowing sensitivities in the zeptomole to attomole range to be obtained.

5.1.2 *Matrix assisted laser desorption and ionization (MALDI)*

The second ionization technique is laser desorption. This has been used since the late 1960s but it was the introduction of a sample preparation method in which the peptide or protein is embedded in a large excess of a laser light absorbing matrix which allowed the breakthrough to a soft ionization technique for high-molecular-weight measurements (Karas & Hillenkamp, 1988). The sample is acidified and mixed with a matrix such as α-cyano-4-hydroxycinnamic acid (αCN) and 2,5-dihydroxybenzoic acid (DHB) or 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid, SA). These absorp light from the commonly used nitrogen (334 nm) and Nd:YAG (266) lasers.

Most often, the MALDI technique is used in time of flight (TOF) mass spectrometers. A laser pulse is used to desorb the sample from the matrix and to

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

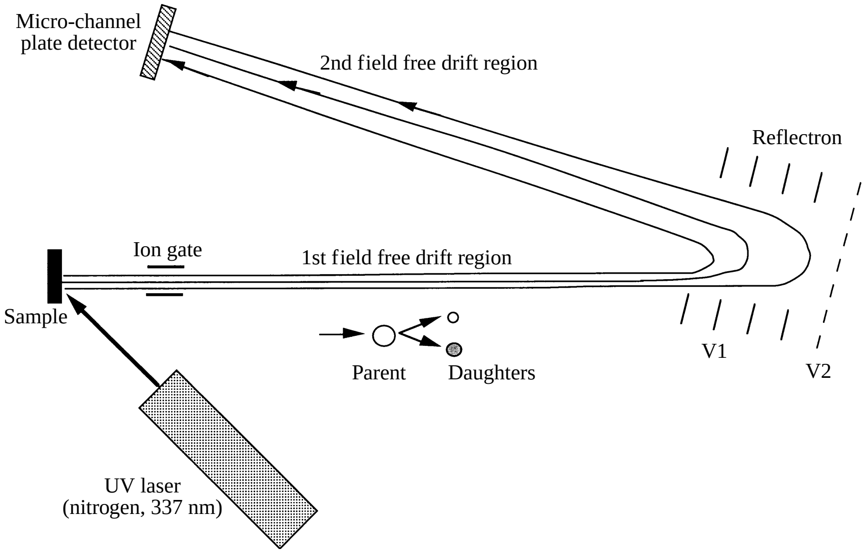

Micro-channel plate detector
2nd field free drift region
Reflectron
Ion gate
1st field free drift region
Sample
Parent
Daughters
V1
V2
UV laser (nitrogen, 337 nm)

Fig. 8. A schematic diagram of a reflectron MALDI-TOF mass spectrometer. The sample is co-crystallized with a UV absorbing matrix. A laser pulse is used to desorb and ionize the sample and to start timing the arrival of ions at the detector. The ions are accelerated by a large potential difference *c.* 30 kV and pass down the evacuated tube to the reflectron. This ion mirror reflects the ions in a mass-dependent manner back to the detector allowing a separation of ions that have decayed (after collisional activation in the initial desorption plume) after acceleration. Parent ions can be selected using an ion gate or deflector which can by application of a potential before and after the arrival of the parent ion remove other ions from the flight path. The mass is determined by measuring the time taken for the ion to travel from the target to the detector.

cause ionization. The ions are accelerated by a high potential (*c.* 30 kV) towards the detector. The principles of construction of such a TOF MS are shown in Fig. 8. The time taken to reach the detector is proportional to the accelerating voltage and the mass. The machine is calibrated by the use of an external standard. The introduction of delayed extraction to a reflectron instrument (Vestal *et al.* 1995), allows very high sensitivity and accuracy mass measurements, with a mass range exceeding 500000.

5.2 *MS/MS sequencing*

MS/MS or tandem mass spectrometry uses two stages of mass analysis coupled together. Various combinations of instruments such as Quadrupole–Time of flight, Magnetic sector–Magnetic sector have been constructed but I will restrict the discussion to the triple quadrupole and ion trap instruments since they have an installed base of tens of thousands of instruments compared to the few dozen four sector magnetic instruments. The great advantage of MS/MS is that individual peptides from complex mixtures can be sequenced without the need for a complete separation which is in contrast to Edman sequencing where the interpretation of mixed sequences is problematic (though if the masses of the

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

peptides being sequenced are known, one can deconvolute the sequences using the MADMAE program described by Johnson & Walsh (1992)). The sensitivity of sequencing by MS is increasing all the time and is now well ahead of that achievable by conventional Edman techniques. Using FAB ionization, sensitivity levels in the hundreds of picomoles could be reached (Hunt *et al.* 1986) and with the use of microcolumn HPLC and electrospray, sequencing became possible in femtomole range (Hunt *et al.* 1992). The development of miniaturized electrospray sources has dropped the sensitivity further into the low femtomole range (Wilm *et al.* 1996*b*) with attomole sensitivity being now recorded using ion cyclotrons (Valaskovic *et al.* 1996).

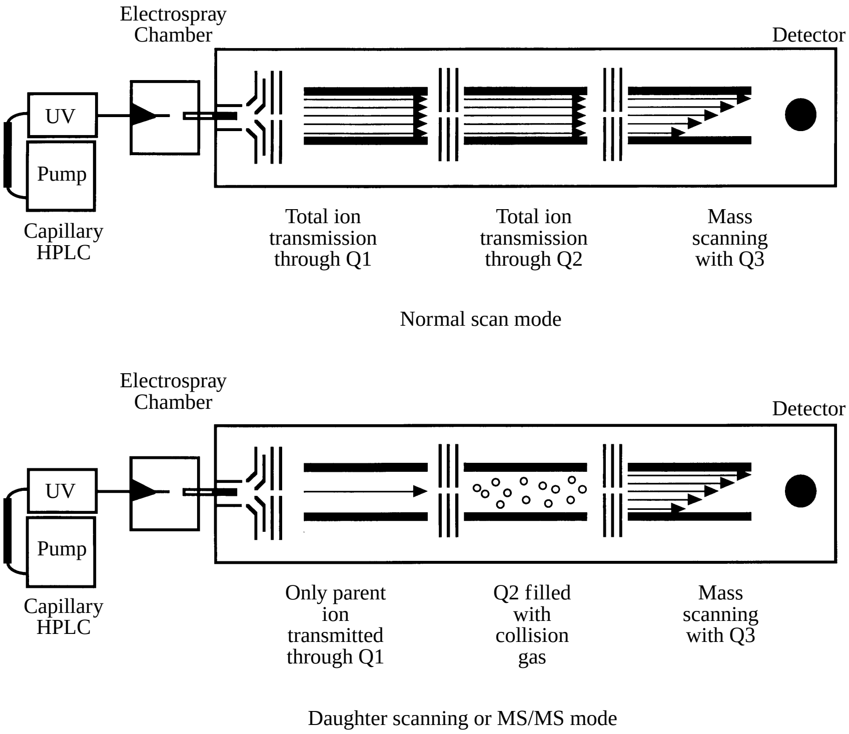
One of the advantages of MS/MS sequencing over chemical methods is the ability to sequence post-translationally modified peptides. N-terminally blocked proteins (this modification is estimated to account for *c.* 70 % of all eucaryotic proteins), phosphorylation sites and unusual modifications such as trimethyllysine can all be sequenced by MS/MS whereas chemical approaches require a standard of the unknown amino acid for comparison or if blocked, extensive chemical deblocking, which is not usually successful. Recently Link *et al.* (1997*b*) have described an extensive analysis of the *H. influenzae* proteome by 2D PAGE and mass spectrometry. Two hundred and thirty-five distinct protein species were identified representing the most abundant proteins in the cell, and of these 25 % were potentially covalently modified (as seen by multiple spots on 2D gels which give almost identical digest patterns). Among the modified proteins, PckA and GlpK showed five different isoforms.

5.2.1 *Triple quadrupole and ion trap mass spectrometers*

The triple quadrupole and ion trap mass spectrometers were invented by Wolfgang Paul (1990) in the late 1950s. The triple quadrupole is essential, a set of three serial mass filters which select ions according to their m/z using a combination of fixed d.c. voltages in combination with an R_f voltage. Ions are allowed to pass through based on their path stability in an alternating electric field as they drift through the central cavity formed by four metal rods. In the first stage a single peptide can be selected (all others are filtered out), the 'parent' ion and allowed to pass into the second stage. This stage is filled with a collision gas at 1–4 mTorr and the parent ions are accelerated into this 'collision cell' where they undergo multiple collisions (*c.* 10) and fragmentation (collisionally induced/activated dissociation CID or CAD). The ions are focused into the third filter where the fragment or 'daughter' ions are analysed (see Fig. 9). The collision energies used are fairly low (10–40 eV) and fragmentation occurs only along the peptide backbone. High energy collisions (1000 eV) as used in four sector magnetic instruments are sufficient to cause side chain fragmentation which sometimes allows the isobaric amino acids leucine and isoleucine to be distinguished.

The ion trap can be considered as a 3D analogue of the quadrupole mass filter. Trapped ions are most often analysed using a mass-selective instability scan in which the amplitude of the R_f voltage applied to the trapping ring is scanned,

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press



Normal scan mode

Daughter scanning or MS/MS mode

Fig. 9. Normal and MS/MS data acquisition modes in a triple quadrupole mass spectrometer. The triple quadrupole may be thought of as three mass filters in series. In normal scanning mode the first two filters are set to allow all ions to pass and the spectrum is accumulated by scanning a fixed width window over the mass range using the third filter. The ions passing through the window are detected and the mass determined by comparison of the RF and d.c. values at that point with a calibration table. In MS/MS mode the first filter is used to select a mass window around an ion allowing it to pass and removing all others. The ion is accelerated into the second quadrupole into the collision gas (usually argon) where it undergoes multiple collisions causing fragmentation. The daughter ions are analysed by scanning the third filter.

causing ions of increasing m/z to adopt unstable trajectories and to be ejected from the trap where they are detected by an external electron multiplier, thus recording a mass spectrum. MS/MS can be carried out in a manner analogous to the triple quadrupole. The parent ion is selected by using d.c. and R_f fields to make all ions except the parent to be ejected. The collisionally induced dissociation is achieved by using resonance excitation to increase the kinetic energy of the trapped ion (though keeping the resonance amplitude low enough to prevent ejection) which undergoes collisions with the helium bath gas which is continually present in the trap. The mass spectrum of the daughter ions is recorded by sequentially ejecting the product ions with a mass-selective instability scan. Since the separation of ions are separated by time and not by multiple mass spectrometer stages, multiple dissociation experiments can be carried out, MS^n. A parent can be isolated, fragmented, a daughter then isolated from the products and further fragmented to granddaughters and so on (up to 10 stages have been carried out in isolated cases).

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

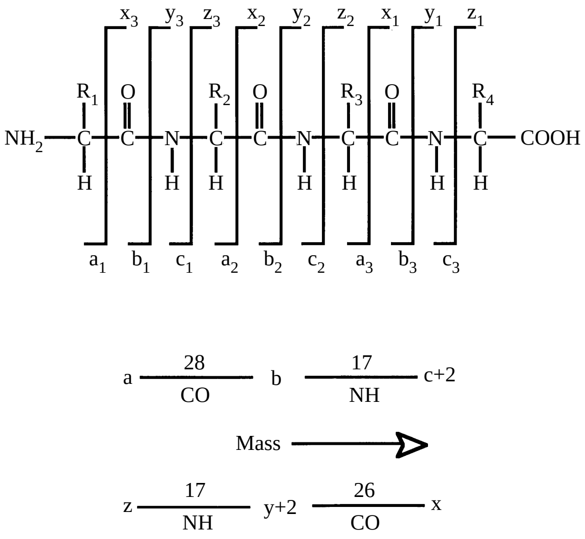

x_3 y_3 z_3 x_2 y_2 z_2 x_1 y_1 z_1
R_1 O R_2 O R_3 O R_4
NH_2 C C N C C N C C N C COOH
H H H H H H H
a_1 b_1 c_1 a_2 b_2 c_2 a_3 b_3 c_3
a 28 CO b 17 NH c+2
Mass
z 17 NH y+2 26 CO x

Fig. 10. Fragmentation nomenclature. The standard nomenclature schemes in use are those proposed by Roepstorff & Fohlman (1984) and modified by Biemann (1988). The low energy collisions in the triple quadrupole and ion trap mass spectrometers mainly involve peptide backbone fragmentation with some neutral losses such as water and ammonia. High energy collisions as achieved in four sector magnetic instruments can produce side chain fragmentation which may allow one to distinguish between the isobaric amino acids leucine and isoleucine.

5.2.2 *Low energy peptide fragmentation*

One of the first people to exploit the triple quadrupole for protein sequencing by tandem mass spectrometry was Don Hunt (Hunt *et al.* 1986). Hunt championed the use of low resolution parent ion selection and daughter ion detection to maximize sensitivity, and chemical modification (peptide methylation and acetylation) to aid the determination of ion series for interpretation. When the set of parent ions (protonated either on basic side chains or along the amide backbone) undergo multiple low energy collisions with the collision gas (usually argon), kinetic energy is converted into vibrational energy and fragmentation at one of the amide bonds occurs (Fig. 10).

The mechanism of gas phase peptide fragmentation is not well understood but a few generalizations can be made (Hunt *et al.* 1986). If the collision energy is high ($>$ 30 eV), fragmentation of a single bond is favoured and b-type ions are formed preferentially (the generally accepted nomenclature for ions is that put forward by Roepstorff & Fohlman in 1984 and modified by Biemann in 1988). If the energy is lower, y-type ions formed by proton transfer to the amide nitrogen and elimination of a ketene will preferentially occur since this involves simultaneous bond formation and cleavage. However, the distribution of product ions between the two species is very much dependent on the number and distribution of residues with high gas phase basicity (lysine, arginine, proline and histidine). Since the neutral and charged fragments produced by low energy fragmentation

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

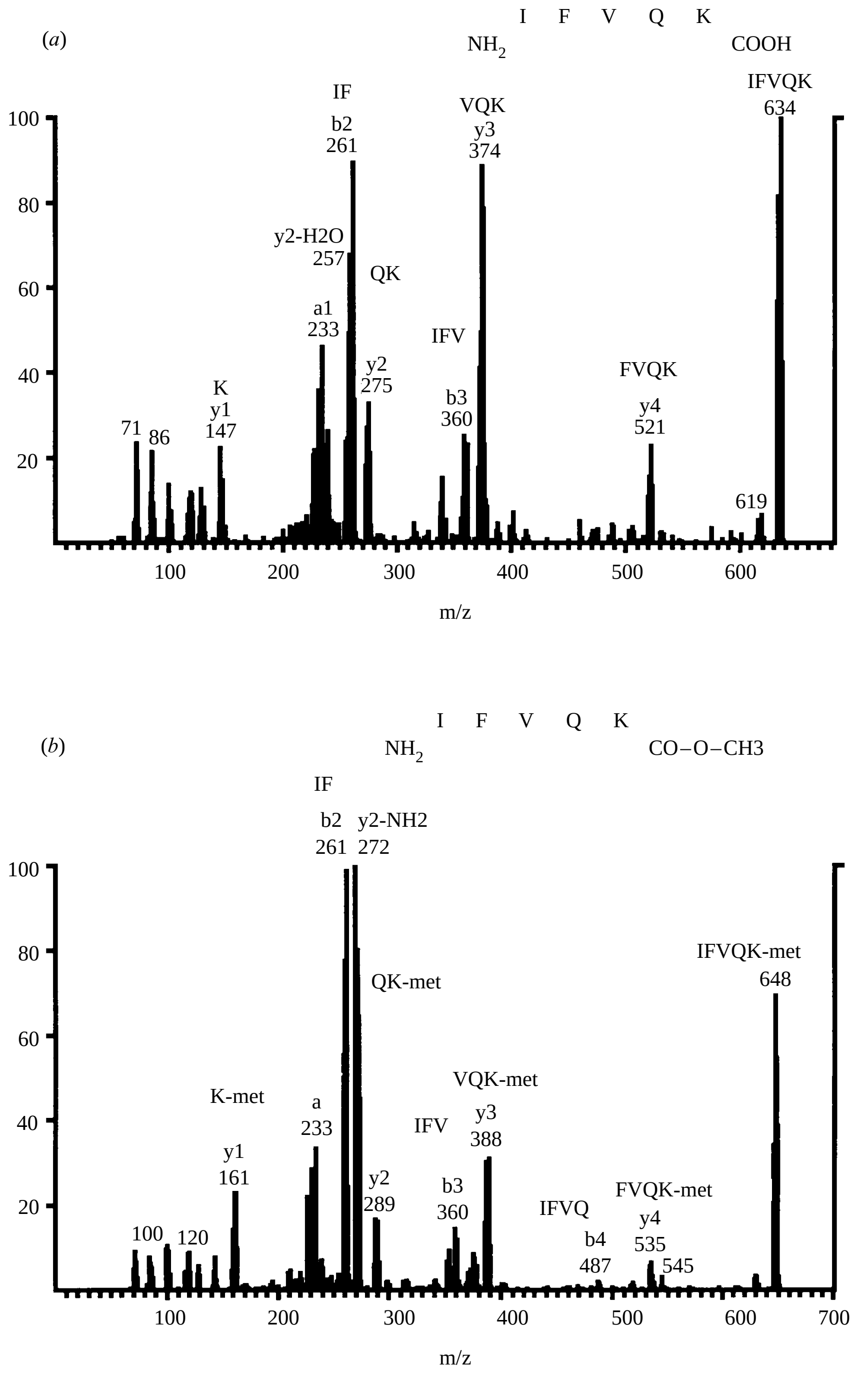


Fig. 11. An example of MS/MS spectra interpretation. The MS/MS spectrum of a simple tryptic peptide, IFVQK is shown ((*a*) is the native peptide, (*b*) is the methylated peptide) as an example of how a spectrum is interpreted. The ions can be assigned to b- and y-ion series by comparing the two spectra, in this case the y-ions can be rapidly identified since they all shift by 14 mass units after methylation. The sequence can be obtained from the mass differences between the ions of the same series since they correspond to the residue mass of the amino acid occupying that position.

do not immediately separate following bond cleavage, but remain associated and undergo further collisions, proton transfer to the fragment of higher basicity may explain the preferential distribution of fragments from a certain ion series (e.g. y-ions from tryptic peptides which have arginine or lysine as the C-terminal amino acid).

Two other major ion types are found in peptide fragmentation spectra which occur as the result of the cleavage of at least two internal peptide chain bonds. The first are the immonium ions which are found in the low mass region and are diagnostic of the amino acid composition of a peptide. The second type occur as a result of two point cleavages in the amide backbone of the peptide and are usually uncommon except on the N-terminal side of proline and histidine residues. These ions are designated $(b_n y_p)_n$, where b_n and y_p indicate the carboxy and amine terminal cleavage points respectively and *n* the number of residues in the fragment. Proline and histidine residues have very high gas phase proton affinities and so protonation and cleavage to form y-type ions which can subsequently lose residues from the C-terminal is highly favoured. All of the ion-types described can eliminate small molecules such as ammonia (−17 U), water (−18 U) and carbon monoxide (−28 U) to give a pair of signals. Other side chain eliminations such as phosphoric acid from phosphoserine and threonine and methylmercaptan from methionine also occur. The drawback of low energy MS/MS is that the isobaric amino acids isoleucine and leucine cannot be distinguished and glutamine and lysine must be differentiated by acetylation. The interpretation of peptide MS/MS spectra has recently been reviewed and the reader is referred to Papayannopoulos (1995) for a more detailed discussion (see Fig. 11 for an example of MS/MS spectra from a native and a methylated peptide).

5.2.3 *Post source decay fragmentation spectra*

Time of flight mass (TOF–MS) spectrometers can be fitted with an ion mirror (a reflectron RE) to increase the effective path length of the ion from target to detector. The longer the flight path, the more temporal and hence mass resolution is gained. Using a laser to desorb peptides from a matrix (MALDI) results in considerable fragmentation if the laser power is increased by a factor of 1·5–2 over the energy needed to obtain a normal signal. Kaufmann *et al.* (1993) demonstrated that in MALDI a large fraction of the desorped ions undergo 'delayed' fragmentation/neutralization reactions during flight and the activation energy for this 'post-source decay' (PSD) comes from multiple collisions of the peptides with the matrix during plume expansion and ion acceleration. Since the accelerating voltage for the peptides is of the order of 30 kV and the gas pressure inside the plume reaches almost atmospheric, very high energy collisions occur (similar to those described for magnetic sector instruments) and the peptides have a relatively long time interval in which to decay. In a conventional linear MALDI TOF–MS, the parent ions reach the detector at the same time as the daughter ions and no PSD spectrum is observed. The parent and daughter ions can be separated using a reflectron since the ions undergo a deceleration and reacceleration which is mass dependent. PSD spectra containing enough information to determine the

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

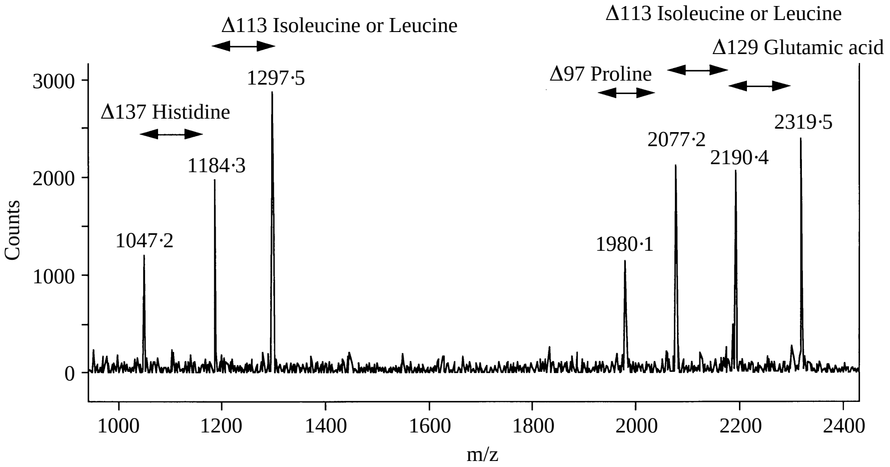


Fig. 12. Ladder sequencing. The figure shows the MALDI spectrum of two peptides (masses 1297 and 2319) after digestion with carboxypeptidase. The mass differences between the digestion products allows one to read a partial C-terminal sequence for each of the peptides.

entire sequence were obtained for bombesin (mass, 1619). With the development of delayed extractions techniques, the mass resolution has been improved dramatically and this will allow the technique to be of general use. However, and this is one caveat, in our experience, only one in five peptides yield good PSD spectra, the others give either no fragmentation or show islands of ions which give only partial sequences (though this is useful for database searching as is discussed later). The spectra are more complex than low energy spectra since a larger variety of ions are found; a, b, c and x, y, z as well as side chain fragmentations d, v and w, thus manual interpretation can be much more time consuming. The advantage is that the isobaric amino acids isoleucine and leucine can be distinguished by their side chain fragmentations since isoleucine has a branched side chain and loses 14 U and 28 U whereas that of leucine is not branched and loses 42 U.

5.4 *Ladder sequencing*

An alternative approach to peptide sequencing by mass spectrometry is the generation of a set of peptides produced by either exopeptidase digestion or by chemical degradation. The mass differences between the adjacent peaks define the residue being removed and hence the sequence. The enzymatic generation of a sequence ladder was first described in 1983 by Aimoto *et al.* during the sequence determination of a heat-stable enterotoxin from *Escherichia coli*. Several groups have used this technique with carboxy- and amino-peptidases to generate N- and C-terminal ladder sequences (see Fig. 12). Recently this method has been revisited using MALDI TOF–MS. The sample is divided amongst a series of sample plate positions in which a series of dilutions of enzyme activity are used for digestion.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

The sequence ladder is then reconstructed by adding together the short digests produced from the various positions (Patterson *et al*. 1995). In 1993 Chait *et al*. demonstrated a chemical method for the generation of N-terminal sequence ladders. The Edman based degradation was slightly modified using a coupling solution of phenylisothiocyanate with 5% phenylisocyanate as a chain terminator. The method is rapid and now with delayed extraction reflectron MALDI TOF instruments, sufficient resolution can be achieved to define long sequences. The main disadvantage is that the degradation is carried out on a polymeric membrane and the peptides have to be recovered from this. In 1994, Bartlet-Jones *et al*. presented a ladder sequencing method using a volatile fluoroisothiocyanate. The ladder was generated not with chain terminators but by dividing the peptide to be sequenced up into aliquots, the number of which define the number of cycles to be determined. After the first cycle a second aliquot is added, after the second cycle a third, etc. so building up a ladder. The main disadvantage is the number of sample handling steps since low quantities of peptides are difficult to store without major losses.

5.5 *Database searching methods*

5.5.1 *Peptide fingerprinting*

Protein identification has taken a quantum leap in speed and sensitivity through the advent of new 'soft' ionization techniques MALDI and ESI, in mass spectrometry. Both allow the accurate measurement (±0·005%) of peptide masses with a sensitivity level for peptide detection by MALDI TOF–MS in the order of tens of femtomoles, whilst triple quadrupole and ion trap mass spectrometers extend the range further with sensitivity levels in the attomole range. It is now possible to rapidly measure the masses of the peptides produced by residue specific enzymatic or chemical digestions, in our laboratory up to 30 samples a day by Auto HPLC–MS/MS and 100 digests by MALDI TOF–MS in a fully unattended mode. 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. A series of papers from the groups of Bill Henzel, Darryl Pappin, Matthias Mann, John Yates and ourselves appeared in the middle of 1993 describing the application of MS based 'peptide mass fingerprinting' for protein identification. Essentially, they described the development of pattern searching algorithms, which use the experimental peptide masses to search for a protein which gives a similar theoretical digest pattern. Table 1 shows the effect of the accuracy of peptide mass determination on the results of such a search. A comparison of the database searching algorithms has recently been carried out by Patterson (1995).

The original idea for protein identification by comparison of peptide masses arising from a digestion to those predicted from a protein database was put forward 10 years earlier by Oliver *et al*. (1983). The masses of peptides generated by partial proteolytic digestion are compared with those predicted for a protein

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Table 1. *The effect of mass accuracy on peptide mass fingerprint searching*

(The table shows the results from two searches carried out with the same data but obtained at different mass accuracies. The higher the peptide mass accuracy, the better the score and the difference between the correct highest scoring hit and the next non-related protein. Currently mass accuracies in the range of ±0·001 are attainable, greatly improving selectivity. AC is the accession number of the database entry.)

Score	AC	DE
Mass accuracy ±0·2 amu. Searching with masses: 932·5, 1064·2, 673·5, 836·5, 915·5		
93·1	P02755	Beta-lactoglobulin, water buffalo
78.7	P02754	Beta-lactoglobulin precursor, bovine
78·7	P02757	Beta-lactoglobulin, sheep
76·0	P02756	Beta-lactoglobulin precursor, goat
51·2	P10834	Pet 54 protein, *S. cerevisiae*
Mass accuracy ±1·5 amu. Searching with masses: 931·3, 1064·5, 674·1, 838, 913·5		
58·2	P02755	Beta-lactoglobulin, buffalo
57·8	P18163	Long-chain-fatty-acid-coaligase, rat
49·4	P02754	Beta-lactoglobulin precursor, bovine
49·4	P02757	Beta-lactoglobulin, sheep
49·3	P05413	Fatty acid-binding protein, human

whose primary amino acid sequence is deduced from a corresponding nucleotide sequence. The proteolytic digestions are accomplished *in situ* in the stacking gel of a 2D PAGE system and the authors were able to show that two variant proteins of the human mitochondrial DNA, MV-1 and MV-2, were allelic and encoded by the unidentified reading frame 3 (URF 3) gene. This approach was possible then due to the small size of the database. As the database increased in size the mass accuracy required increased to a level beyond that which was attainable by SDS–PAGE and the idea had to wait until 1993 to be rediscovered for use with MS. The initial reports of peptide mass fingerprinting were carried out with a mass accuracy of ±0·3–2·0 with 4–5 masses and it was demonstrated that the higher the accuracy, the higher the confidence level was of the result.

As the size of the databases has increased rapidly with the results from the genome sequencing projects, both sample preparation (Vorm & Mann, 1994) and mass spectrometric methods (Jensen *et al.* 1996*a*) have been developed to increase the mass accuracy which greatly improves the search accuracy. 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 (James *et al.* 1994). The

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Table 2. *Increasing search accuracy of DNA databases by using orthogonal data sets*

(When peptide mass fingerprinting is carried out using DNA databases the confidence level drops when using a single set of data. This is due to the larger size of the DNA databases, since for one entry all six reading frame translations must be searched and secondly the number of entries (especially those of expressed sequence tags) are at least 10 times greater. By using orthogonal data sets from the protein, a high confidence level can be restored. ACC is the database accession number of the sequence, Pos, the position of the sequence in the search output and Delta is the difference in score between the correct and next highest scoring non-related sequence.)

			Single digest		Dual digest	
Protein	ACC	Digest	Pos.	Delta	Pos.	Delta
Lambda receptor	P02943	LysC/Tryp	1	+3·9	1	+58·3
Citrate carrier	P31602	Tryp/AspN	2	−2·1	1	+29·8
10 kDa Chaperonin	P15020	V8/Tryp	1	+5·6	1	+119·8
Na/K ATPase Alpha1	P06685	LysC/CNBr	1	+15·8	1	+81·6
Lipid binding protein	P07926	Tryp/V8	2	−7·7	1	+40·9
Apolipoprotein AI	P02647	LysC/Tryp	1	+10·1	1	+102·5
		Average		+4·3		+72·2

difference in the confidence levels obtained using single and orthogonal database searches is shown in Table 2. Peptide mass fingerprinting is emerging and slowly gained acceptance as a reliable and rapid alternative to peptide sequencing by Edman degradation.

5.5.2 *Peptide fragment fingerprinting*

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. A single auto-MS/MS run of a protein digest may contain up to 50 different peptide MS/MS spectra. Manual interpretation of each spectrum to determine the sequence for a database search would take a very long time (30 min per spectrum at the fastest). In order to deal with this problem the group of John Yates developed an algorithm, Sequest, to correlate uninterpreted tandem mass spectral data of peptides to sequences in a protein database in an automated fashion (Eng *et al*. 1994). The MS/MS spectra are automatically stripped from an auto-MS/MS HPLC run datafile and the mass of each parent ion is used to search for isobaric amino acid stretches in the database. The experimentally determined spectrum from that parent ion is compared with the theoretically predicted spectrum for that sequence and the 500 best matches are subject to cross

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Database searching
with MS/MS spectra

Raw data handling

1. Extract MS/MS Spectra from HPLC run
2. Convert to ascii files
3. Filter spikes from spectra
4. Collapse isotopic peaks to one mass
5. Remove 10 amu window around parent
6. Output file: *c*.200 masses and intensities

Database search

1. Translate DNA in all six reading frames
2. Find all peptides ± 3 Da in database
3. Predict MS/MS spectra of peptides
4. Find best 500 matches to the spectrum
5. Cross-correlate 'found' *vs*. 'real' spectra

Fig. 13. Sequest MS/MS database searching outline. The figure outlines the data processing and searching procedure (Sequest) developed by the group of John Yates to allow fully automated protein identification using uninterpreted MS/MS data.

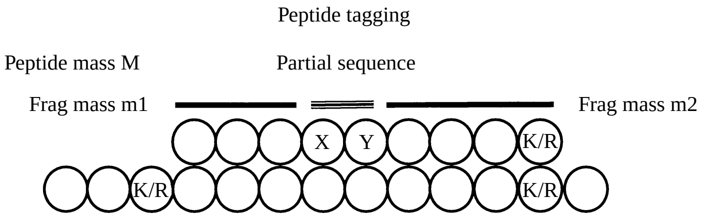


Fig. 14. Peptide tag searching. The figure shows the data parameters used by the PeptideSearch program developed in the group of Matthias Mann. Each MS/MS spectrum is manually inspected and four search parameters are extracted: parent mass, mass at start of sequence, a small sequence found by the inspection and the mass at the end of the sequence. This can be used to define a search in which one or more of the parameters can be relaxed allowing for modifications of the peptides.

correlation analysis. Fig. 13 outlines the steps in database searching using MS/MS data according to the Sequest approach.

The group then extended the algorithm to deal with tandem mass spectra of modified peptides (Yates *et al.* 1995*b*) and then to search DNA databases (Yates *et al.* 1995*a*). This technique is very useful when inaccurate or small DNA sequence stretches such as ESTs (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. Matthias Mann took a different approach to peptide identification using peptide sequence tags (Mann & Wilm, 1994). The algorithm PeptideSearch, 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. This is outlined in Fig. 14.

This approach is complementary to that of Sequest. The power of both of these methods is that they can be used for the identification of proteins in complex

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

mixtures (McCormack *et al.* 1997). The ion series obtained in an MS/MS spectrum depends on the collision energy being used and so versions of Sequest have been written to take this into account for database searching with MALDI–PSD spectra of peptides (Griffin *et al.* 1995) as well as tandem mass spectra from four sector magnetic instruments (Yates *et al.* 1996).

5.5.3 *Ragged termini*

Alternative approaches to generating sequence tags for database searching without the use of MS/MS has been developed. Jensen *et al.* (1996*b*) observed that many peptides in a mass spectrum arise from incomplete cleavages of the type KK, RK, KR and RR at the C-terminal in the case of trypsin due to the low exopeptidase activity of the enzyme (this is true for most other specific endopeptidases). On average, for a protein > 20000 four of these sequences will occur. The pairs of peptides that arise from cuts at either site can be automatically identified and used with the peptide masses for a modified tag search. Alternatively a single step of Edman degradation can be carried out on the unseparated peptide mixture and single amino acid tags extracted. We have recently extended this ragged termini approach by developing an algorithm for the identification of proteins in sequence databases using peptide masses and N- or C-terminal sequence tags generated by sequential endo- and exopeptidase digestions (Korostensky *et al.* submitted). MALDI–TOF spectra of unseparated protein digests before and after carboxypeptidase or aminopeptidase digestion are used to define a series of peptide tags. Unusually on average only two or three peptides will be digested whilst the others remain intact. The approach is useful for the identification of difficult to digest proteins where a less specific protease such as chymotrypsin or pepsin must be used and so the database search must be carried out against all possible peptides and not a subset defined by a digestion.

6. THE GENOME–PROTEOME INTERFACE

6.1 *Analysis of gene function in the post-genome era*

One of the biggest problems to arise from the genome projects is how to assign functions to the large numbers of open reading frames that cannot be assigned on the basis of homology searches. Even those genes for which a function can be assigned must be reappraised and the function defined when acting in concert with the cellular patterns. The proteome approach allows one to approach complex problems in a global fashion by subtractive 2D analysis, and in combination with a microarray method for mRNA analysis, will allow a systematic study of the correlation of gene activity and protein expression in a cell at defined time. Since the proteome and gene expression analysis methods are quantitative, new systematic approaches to define protein functions in terms of the interaction networks built up in the cell. N. L. Anderson has defined three major areas for the analysis of gene function and regulation: molecular anatomy (protein composition of cells and tissues); molecular pathology (analysis of disease in terms of changes

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

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 cell's micro- or macro-environment. Furthermore proteins must be assigned to a position in the cell. The biotechnology company, Large Scale Biology is developing preparative ultracentrifuges to allow cell organelles to be purified in bulk, so that a systematic analysis of the proteins present can be undertaken. Link *et al.* (1997*a*) have already described a strategy for the identification of proteins localized to subcellular spaces and applied it to *E. coli* periplasmic proteins.

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 *c.* 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; proteome analysis will be one of the key technologies in helping understand the changes occurring as a result of mutations.

6.2 *Genes looking for functions and functions looking for genes*

We have been working for the past four years on methods to allocate functions to genes (proteome methods) and have defined two working approaches; a function looking for a gene (using either gene in/activation by a fixed stimulus or random gene knockout and screening for loss of a specific function), and a gene looking for a function (by directed gene knockout or directed gene activation). Using the first approach of a function in search of gene we searched for the genes responsible for the adaptation of *E. coli* to growth in the absence of inorganic sulphate. 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 (Fig. 15). All the proteins were identified by MS analysis (Quadroni *et al.* 1996; Dainese *et al.* 1997*c*). Of the four known proteins, three, sulphate binding protein, cysteine synthase A, and cystine binding protein, belong to the cys regulon whereas the fourth was found to be a general stress protein, alkylhydroperoxide reductase. Two of the proteins found on the gel were identified in the 8·5′ region of the genome, which contains four previously unidentified open reading frames with a single promoter region; the other two proteins were found in the 21·5′ region. Both 8·5′ and 21·5′ regions show many features common to ABC-type transport operons; a periplasmic substrate binding protein, a channel forming membrane protein and a cytoplasmic nucleotide binding protein.

These two operons were also found in a parallel study using random phage

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

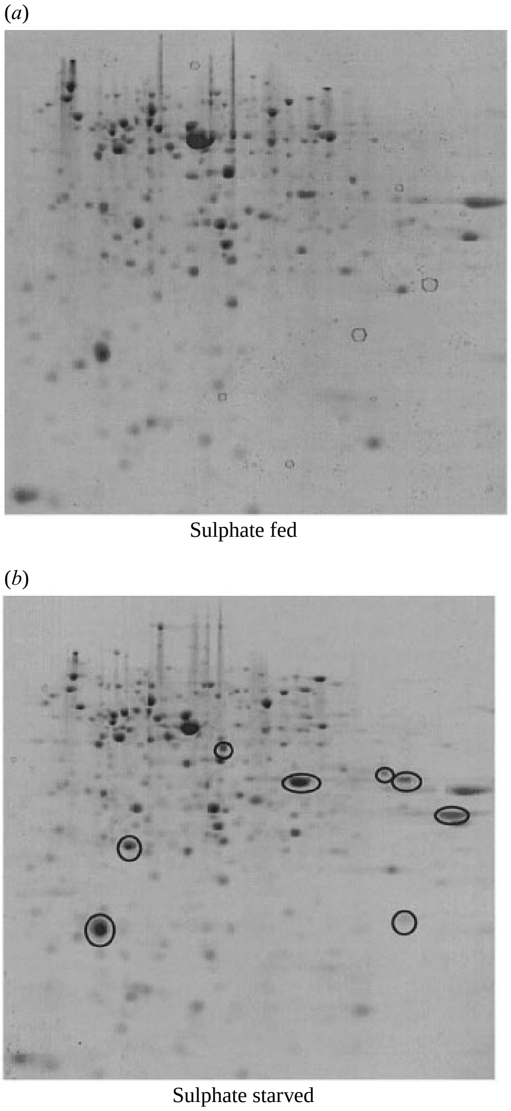


Fig. 15. 2D PAGE mapping of *E. coli* grown in the presence or absence of inorganic sulphate. *E. coli* was grown with $500\ \mu\text{M}$ sulphate (*a*) or $500\ \mu\text{M}$ ethanesulphonate (*b*) as the sole sulphur source. All the proteins which were induced during growth with ethanesulphonate by more than a factor of × 2 are labelled. These spots were analysed by mass spectrometry and the data used to find the corresponding genes in the genome sequence.

insertion mutagenesis and screening for mutants that cannot grow in the absence of inorganic sulphate. Promoter analysis and growth experiments using various organic sulphur sources confirmed that the four 8·5′ ORFs are a sulphate regulated, taurine uptake operon, whereas the 21·5′ ORFs are a sulphate regulated, alkylsulphonate uptake operon. We have used the second approach of directed gene activation and inactivation to study transcriptional activators responsible for inducing nitrogen fixing genes in the symbiotic bacterium, *Bradyrhizobium japonicum*. Several transcriptional activators are responsible for bringing about the switch from the free living to the symbiotic nitrogen fixing lifestyle in the host legume. The transcriptional activator responsible for the induction of the nitrogen fixing genes had been identified by the first approach of random gene knockout (Dixon *et al*. 1980) and we then used a transposon insertion mutant to define which genes were not being activated in the mutant. As a positive control we constructed a mutant in which the transcriptional activator was permanently active to see which genes were being switched back on (Dainese *et al*. 1997*a*, submitted).

6.3 *Entering the post-genome era*

The post-genome era has already been heralded in Nowak, 1995. The focus of biological problem solving must now be rethought in the new surroundings of the genome era and its accompanying flood of data. Previously, complex processes such as development or differentiation were studied in a reductionist fashion, concentrating on individual proteins or at most small sets of proteins. In the case of an organism for which the entire genome is known, a global approach can now be taken by simultaneously monitoring the changes in the expression of all possible genes under various conditions. The tools to carry out such analyses have been recently developed, at least for monitoring the levels of mRNA expression. The initial attempts at a global determination of gene expression have been carried using *E. coli*, using an ordered set of overlapping lambda clones which span the entire genome (Kohara *et al*. 1987). Chuang *et al*. (1993) analysed the levels of all mRNAs by hybridization to the lambda clone inserts which had been immobilized on a nylon membrane in a 21 × 21 grid. The response of *E. coli* to various stimuli or genetic defects were studied allowing both the extent of the response and the position of the responding genes in the genome to be determined. The lambda inserts used were too large, sometimes carrying two or three genes, making interpretation difficult. Recently Freiberg *et al*. (1997) have reported monitoring the expression levels of all the genes encoded on the symbiotic plasmid of the bacterium, *Rhizobium* sp. NGR234, during *in vitro* growth and *in situ* within the host plant. The plasmid was completely sequenced and fragments of each potential open reading frame were constructed and immobilized to a membrane for hybridization. This allows both qualitative and quantitative information on gene expression to be obtained.

The situation is somewhat different for organisms with very large genomes, such as the human, which have not been sequenced. A brute force approach has recently been described by Lee *et al*. (1995) in which a comparative EST tag profile analysis was carried out on rat adrenal chromaffin PC-12 cell before and after nerve growth factor treatment. Six thousand EST were sequenced, and 600 differentially expressed mRNAs were found, many of which encode proteins belonging to cellular pathways not previously known to be regulated by nerve

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

growth factor. Alternatively, expression patterns can be approximately determined using subtractive cloning, the construction of cDNA libraries by removal of all mRNA common to cells sampled under both the two conditions of interest (Adams *et al*. 1991).

Recently two new approaches have been described for the quantitative analysis of gene expression. Both are open to automation and can be used for very large scale gene expression studies and are even capable of dealing with mRNA with low expression levels (5–10 copies per cell). Schena *et al*. (1995) described the construction of complementary DNA microarrays, prepared by robotic printing of gene or EST specific cDNAs onto glass plates in a defined manner. The mRNA from a cell is extracted and subjected to a single round of amplification and fluorescence tagging. The mRNA is then hybridized with the microarray which can then be scanned for fluorescence which shows hybridization has occurred, the intensity of which is proportional to the amount of the mRNA binding. An alternative method proposed by Velculescu *et al*. (1995) is called Serial amplification of Gene expression (SAGE). A short (9 bp) sequence tag is created at a defined position in each of the mRNAs in the pool isolated from the cells of interest. The tag is created by tagging all the mRNAs in the pool with biotin at the 3′ end. A defined restriction endonuclease is used to digest all the mRNAs and those carrying a biotin labelled 3′ end are isolated using streptavidin beads. A second restriction enzyme is used to clip out a nine base pair fragment (the tag) from the mRNAs. This tag is then ligated to a primer, amplified and joined to a second tag to form a ditag which is then isolated, concatenated and cloned. The short tags are all aligned on a single strand and are separated by sequences defined by the specificity of the endonucleases used in the preparation procedure. Clones containing 10–50 tags are amplified by PCR and the sequence determined. Matching of the tag to the sequence database together with the frequency of its occurrence defines the gene expression pattern for the cell line or tissue.

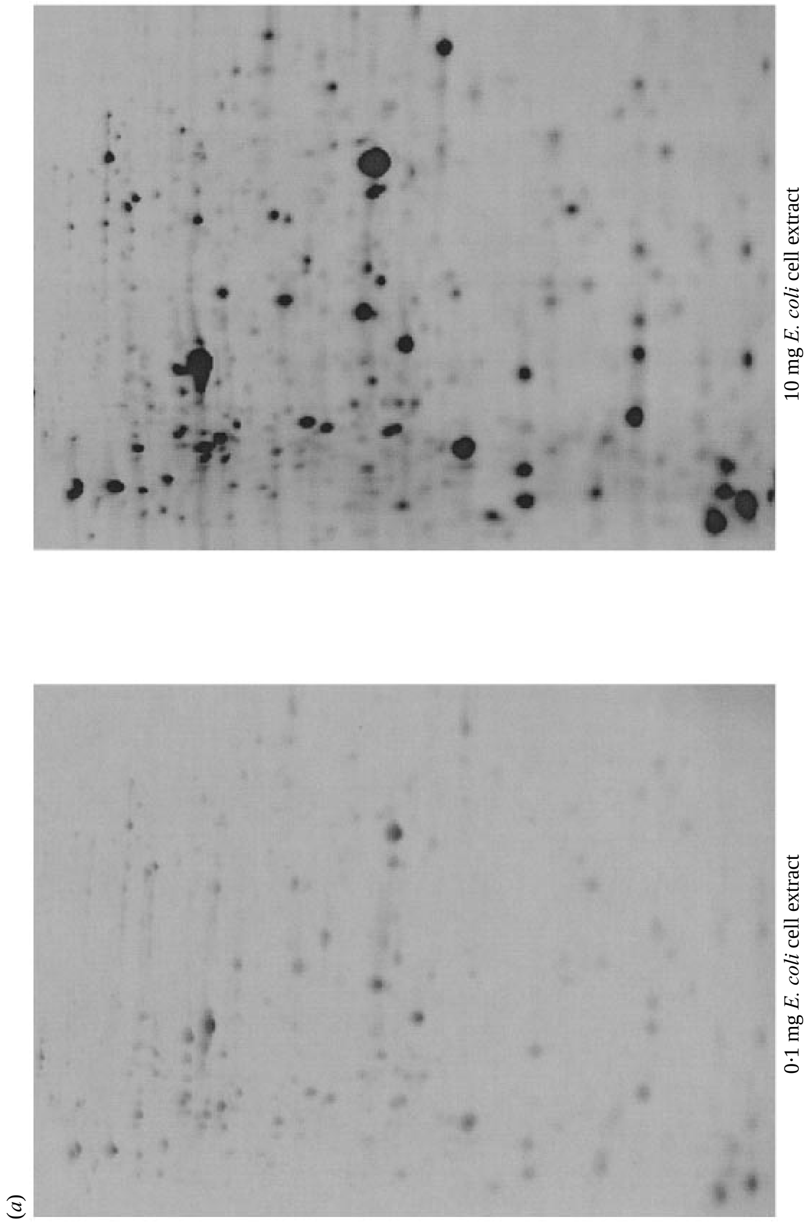
The trends in both DNA and protein analysis are clear: (i) innovation giving a new analytical method, (ii) miniaturization to give high sensitivity and the possibility of large scale operation and (iii) automation, so that sample preparation to data collection and analysis is free from as many variables as possible. These requirements have largely been fulfilled for DNA/RNA analysis, though refinements to improve speed, sensitivity and reliability (and of course, cost) are still being actively studied. The main challenge lies in the field of study of protein expression and modification. Can similar large scale and high sensitivity proteome methods be developed to allow a comparison of the results obtained by two methods (James, 1997; Kahn, 1995)?

7. TRENDS

7.1 *Protein isolation*

Unlike DNA sequencing, there is no amplification method analogous to PCR for proteins. As the amount of protein or peptide needed for identification analysis has

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press



0·1 mg *E. coli* cell extract

10 mg *E. coli* cell extract

Fig. 16. For legend see facing page.

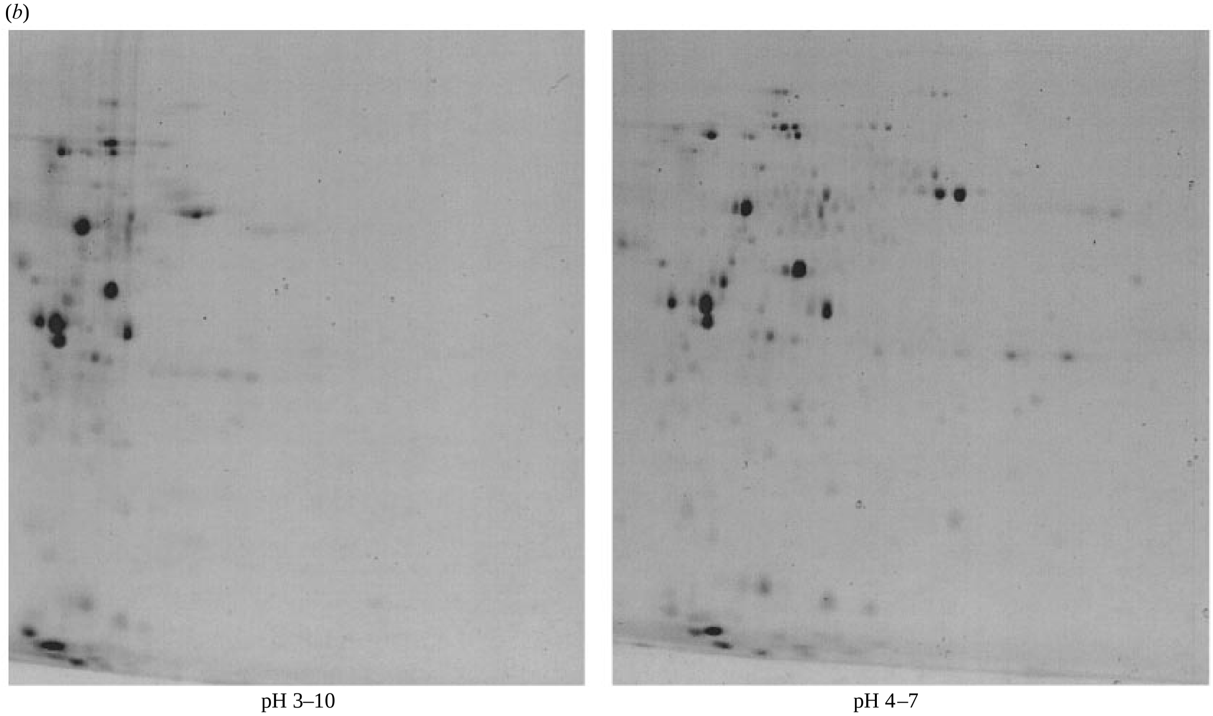


Fig. 16. Sample application methods and zoom gels. (*a*) A gel loaded in the 'traditional' point loading method (left) as compared to one loaded by gel rehydration (right). This allows a hundred-fold increase in protein loading without loss of resolution. (*b*) The rehydration method in combination with a narrowing of pI range to allow even higher loading and increased resolution.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

dropped from the milligram to microgram and now to the picogram level, losses due to irreversible binding to surfaces play an ever more important role. Thus strategies must be developed for handling polypeptides on a microscale (for a recent review, see Smith & Tempst, 1997). One limiting factor in proteome analysis is detection of protein on the 2D gel. Coomassie blue staining can visualize around 500 spots (more using colloidal staining) down to the subpicomole range with silver staining increasing this by one or two orders of magnitude, visualizing up to 10000 spot and new fluorescent methods may reach the attomole level. This is based on 2D gel sample loading of a few hundred micrograms. Recently, with the introduction of immobilized pH gradient gels which are made in the form of strips which are dehydrated for storage, sample applications of tens of milligrams can be carried out by gel rehydration (Rabilloud *et al.* 1994), as shown in Fig. 16*a*. Secondly, instead of using one 20 cm strip to cover the range from 2·5–12, a set of 20 cm 'zoom gels' with overlapping pH ranges (2·5–4·0, 3·5–5·0 etc.) can be used allowing even high loadings (Fig. 16*b*).

This can be combined with some form of protein prepurification, preferably orthogonal to pI and MW (Corthals *et al.* 1997). Finally the spots isolated from multiple 2D gels (> 50) can be combined and concentrated down to a single spot using gel concentration systems which also allows an effective substitution of a MALDI compatible detergent for SDS (Dainese *et al.* 1997*b*; *Rider et al.* 1995).

One of the most tedious aspects of proteome analysis, and a potential source of irreproducibility is the collection of spots from the gels. What is needed is a robotic system with automated detection and spot identification which allows spots to be cut out according to user defined criteria (perhaps based on changes with respect to a master gel). Prototype instruments which can carry these functions out have already been constructed by Oxford Glycosciences. In order for large scale collaborative proteome projects to be possible, the gels must be absolutely reproducible. This has taken a step forward with the introduction of the immobilized pH gradient gels in the first dimension but the construction of an automated 2D gel apparatus (Nokihara *et al.* 1992; Harrington *et al.* 1993), in which the gels are cast, samples loaded, run, stained and scanned must be one of the highest priority goals of proteomics over the next few years.

7.2 *Protein digestion*

Protein digestion is usually carried out either *in situ* in the gel or on membrane supports after electroblotting. There is a fair amount of disagreement as to which is more efficient. Electroblotting is preferred by many since it is closely associated with N-terminal sequence determination, though the efficiency of blotting is extremely variable and conditions should be optimized for each protein. Lui *et al.* (1996) undertook a methodical analysis of protein–nitrocellulose interactions to design a refined digestion protocol to maximize peptide recovery from blot digests by optimizing the pH and detergent used. A detergent is needed to prevent protease loss by adsorption to the membrane and to aid peptide extraction but can make the subsequent analysis difficult, though MALDI is a fairly forgiving

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

method and tolerates the presence of low amounts of salt and sometimes detergent (Gharahdaghi *et al.* 1996).

The advantage of 2D gels and the various concentration gel systems is that the protein is recovered in a *c.* 20 μl gel slice, ideal for digestion, since the surface area available for losses is small as is the amount of protease required. Usually the gel piece is partially dehydrated and the protease is added in the rehydration solution. Peptides are recovered after digestion by repeated extractions using organic solvents which are then reduced in volume under vacuum before dilution in acidified water and injection onto a reversed phase HPLC column. The extraction and subsequent volume reduction steps are probably where most losses occur. If proteins are to be analysed at attomole levels a high efficiency extraction and digestion method must be developed, preferably one which can be interfaced directly to the mass spectrometer for analysis. Davis *et al.* (1995 *a*) have developed microscale immobilized protease reactor columns for protein identification by HPLC-MS. This can be easily adapted to the extraction and digestion of proteins in gel slices and the autoproteolytic products can be eliminated if polymer based column material is used instead of silica (James, unpublished data). The protein is eluted from the gel and pumped slowly through the protease column and collected directly on a capillary reverse phase column.

7.3 *Miniaturization*

The sensitivity of electrospray MS has taken a dramatic leap forward with the development of ultra low flow rates (nl min^{-1}) and small spraying orifi (< 10 μM) for sample introduction and is approaching Dole's concept of a beam of macromolecules being directed into the mass spectrometer without loss (Gale & Smith, 1993; Andren *et al.* 1994; Wilm & Mann, 1994; Kriger *et al.* 1995; Valaskovic *et al.* 1996). In order to match this, separation methods must be developed which can couple on-line separation directly with a 'nano' or 'pico-electrospray' source. Capillary zone electrophoresis (CZE) already operates at these low flow rates but suffers from a low sample volume loading capacity, whereas HPLC can deal with very high sample volumes but it was not possible to form gradients for elution at such levels.

7.3.1 *HPLC-MS*

When reverse phase HPLC columns were first marketed, analytical columns were usually 4·6 mm in diameter and 25 cm long and were run at flow rates of 1 ml min^{-1}. As the sensitivity of Edman sequencing increased (for a review see Simpson *et al.* 1989), smaller sample volumes and lower amounts of material lead to the successive introduction of narrow bore (2·1 mm), microbore (1 mm) and finally capillary bore (< 500 μm), with a concomitant reduction in flow rates, from 500 μl to 10 nl min^{-1} (Shelly *et al.* 1984). Conventional HPLC pumps can only produce reproducible gradients down to 1 μl min^{-1} without flow splitting before the column. Therefore Davis *et al.* (1995 *b*) developed a low flow solvent delivery

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

system for capillary HPLC-tandem mass spectrometry. They used a rheodyne injection loop to store a preformed gradient made by two syringe pumps, once the sample has been injected at a high flow rate the loop is switched in-line and the gradient pumped out using a constant pressure syringe pump and the reversed phase capillary column was attached directly to a microscale electrospray interface developed for low flow rates (Davis *et al.* 1995*c*).

7.3.2 *Capillary zone electrophoresis-MS*

CZE is ideally suited for coupling to a mass spectrometer (Deterding *et al.* 1991) combining extreme sensitivity with very high resolution. The main drawback has been the loading of sample. Long capillaries with small internal diameters (100 cm by 50 μm) have a small volume and only 2% of this can be loaded with sample without adversely affecting the separation. Several approaches to loading dilute samples have been developed which fall into two categories, either pre-concentration by solid phase extraction or by using special injection techniques. The latter include isotachophoresis, stacking, and field amplification. These depend on the stacking or focusing of analytes zones according to the variations in ion mobilities with field strength or chemical environment. The maximum loading volume is then determined by the total capillary volume (< 1 μl). The alternative method involves pre-concentration by binding to a solid phase, such as C-18 derivatized silica beads (Figeys *et al.* 1996; Figeys *et al.* 1997*a*) or an impregnated porous membrane (Tomlinson *et al.* 1996). Both procedures can be carried out on-line, allow large (> 500 μl) volumes to be loaded within a few minutes and the sample can be extensively washed to remove salts and other impurities coming from the gel pieces. Peptides obtained from the capillary columns are very clean (with respect to non-peptidic compounds) and concentrated and can be directly coupled to a nanospray interface, allowing detection limits in the low attomolar/μl range to be obtained with nanogram amounts of protein isolated by 2D gel electrophoresis.

7.4 *Automation of data collection and evaluation*

7.4.1 *Improving data quality*

The disadvantage of the super low flow electrospray interfaces is that at nl min^{-1} flow rates even the smallest dead volume in the system can lead to huge delays. One approach to dealing with this problem in capillary HPLC is the use of variable flow rates (Stahl *et al.* 1996). They describe a modification of their 'nanoHPLC' in which a preformed gradient is pushed out through the capillary RP column by a syringe pump. By operating this pump in pressure control mode, several flow rates can be pre-programmed, one high flow rate around μl min^{-1} for sample loading and washing, a medium flow of 300 nl min^{-1} for elution and a very low flow (parking) of 20 nl min^{-1} to carry out extensive MS/MS data accumulation on an eluting peptide. Instrument control language programs have been written that automate the accumulation of MS/MS data from peptides as they elute from

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

an HPLC column into the mass spectrometer (Fig. 17). Our program picks the most intense ion, checks to see if this has already been sequenced, and if not sets up MS/MS parameters according to the mass, assuming it is a 2+ ion. The program then looks at the distribution of the daughter ions above the parent to determine the charge state (3+/2+/1+) and the collision energy is scanned from low- to high-energy to ensure a good coverage of ions across the daughter spectrum. The sequences of four metallothioneins (Piccinni *et al.* 1994) were obtained using on-line automated collision activated dissociation controlled by tandem quadrupole mass spectrometry by an instrument control language procedure.

Typically peptides were assumed to be doubly charged and were sequenced using a collision energy of between 10–30 eV and 1·8 mTorr Argon with a parent ion window of 2–3 mass units (50% half height) and a daughter ion resolution of 1–2 mass units. Using the parked flow method of Stahl *et al.* (1996), peptides elute within 30 s at the elution flow rate, but by dropping the flow by a factor of 10, one has 5 min in which to optimize the MS/MS conditions for all parts of the spectrum and to increase the S/N by averaging over a long time. A commercial variation on this strategy has been introduced in an ion trap mass spectrometer produced by Finnigan MAT, the so-called 'triple play'. The MS carries out full range scans until a signal greater than a defined S/N ratio appears, a slow high resolution scan is then carried out over the parent ion to resolve the isotopes and determine the charge state before MS/MS is carried out with the appropriate energy settings. The higher quality of data greatly improves the results of automated database searching by Sequest.

7.4.2 *Optimizing data collection*

This data control approach is useful for filtering out unwanted species such as autoproteolysis products and non-peptide impurities from the gel. A list of commonly found autolysis products of the enzyme used can be entered in a user defined list together with the strongest ion in their MS/MS spectrum. The program upon detecting an ion for MS/MS would check the list and ignore the ion if present in the list. A similar voltage operated system should be possible to design for the CZE separations. Another variation is the use of parent scanning to pick out peptides when their signal drops below the chemical noise in the mass spectrometer (Wilm *et al.* 1996*a*).

The mass spectrometer is set to scan the masses of all species entering the collision chamber whilst monitoring one or two daughter ion masses (shown schematically in Fig. 18). The system can be programmed to monitor the low daughter ion masses of arginine (175) or lysine (147) and under data control, switch to normal MS/MS mode when the S/N of an eluting mass rises over a set value (or 86 m/z for peptides containing leucine or isoleucine from other types of digest).

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

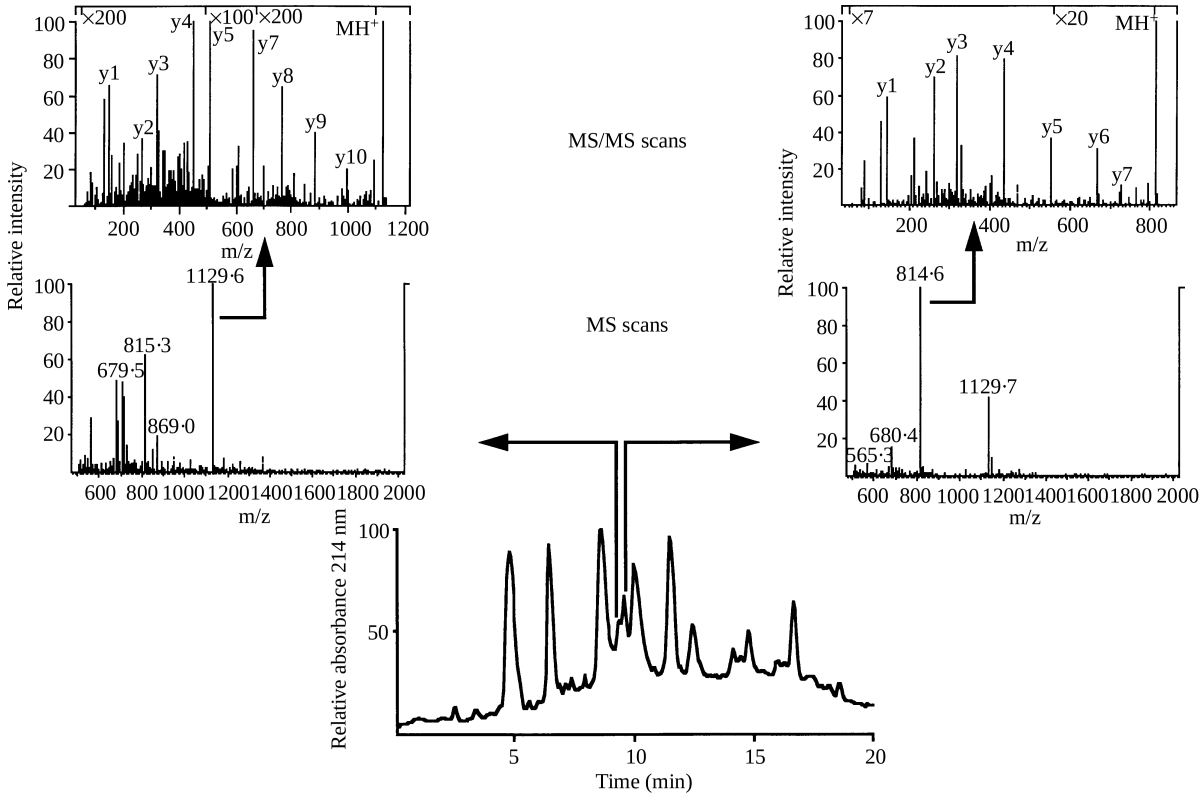


Fig. 17. Auto-HPLC-MS/MS data accumulation. A tryptic digest of a protein was separated on a C_{18} reverse phase capillary column and the eluting peptides were subject to on-line automated collision activated dissociation (CAD) tandem quadrupole mass spectrometry (MS/MS) using a program written in the Finnigan MAT instrument control language (ICL). The mass spectrometer is operated in normal mode scanning with Q3 (in the presence of 1·8 mTorr gas but no collision offset voltage) and the program picks out peptides showing a signal/noise ratio $>$ 5 for MS/MS analysis. The middle left hand panel shows the peptide of mass 1129 being observed in normal mode and then being subject to MS/MS analysis (upper left panel). After collecting six spectra the program reverts to normal scanning mode and picks the next closely eluting peptide, mass 814 for MS/MS analysis (right hand panels).

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

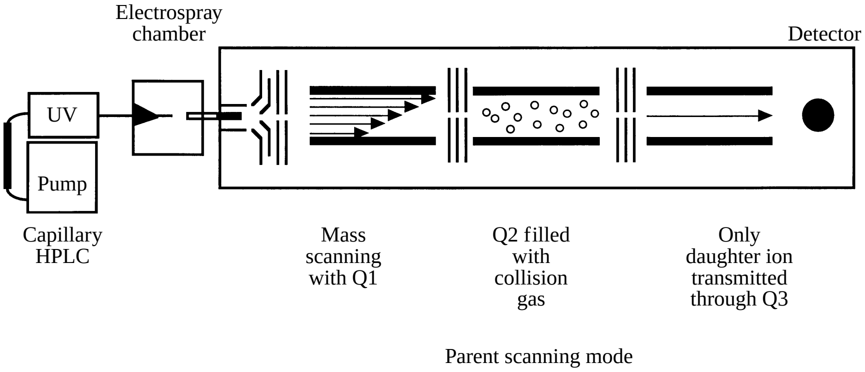


Fig. 18. Parent mode scanning. This is almost the reverse of MS/MS mode. The first mass filter is set to scan in normal mode. As peptides of defined mass pass the filter they are subject to collisionally induced dissociation in the second quadrupole and the third filter is set to detect one or two masses only. These mass windows can be set to ions diagnostic of peptides such as 86 for isoleucine and leucine or 147 for lysine allowing peptides to be picked out even when their intensity is the same as the chemical noise in the system.

7.4.3 de novo *peptide sequencing*

Despite the vast amounts of sequence data being generated from the EST and genome projects there will still be a need to obtain sequence information by interpreting MS/MS spectra. This is a tedious task that requires a fair amount of experience. Several attempts have been made at automatic interpretation using computer algorithms (Johnson & Biemann, 1989; Bartels, 1990; Yates *et al.* 1991; Hines *et al.* 1992) but none was very successful, except to a limited extent for tryptic peptides. The main problem in interpretation is allocating the peaks to an N- or C-terminal (or internal) ion series. This can be facilitated by chemical means such as esterification and/or acetylation as practised by Hunt. Another approach which also has the disadvantage of having to sequence the sample twice (in the case of FAB or ESI) is Hydrogen–Deuterium exchange (Sepetov *et al.* 1993). Since each amino acid exchanges different numbers of protons, from zero for proline to five for arginine, this can be combined with the 'normal spectrum' to reduce the number of possibilities greatly. This has also been applied to PSD sequencing on MALDI TOF instruments (Spengler *et al.* 1993). Here the sample is analysed and then the same sample deuterated on target without the need for extra material. Recently Taylor & Johnson (1997) have taken another approach to sequence determination. The output from a *de novo* sequencing program is used to create a list of twenty or so candidate sequences which are used as a query for a subsequent homology based database search routine. This allows several of the problems of MS/MS interpretation to be overcome, such as sequence reversal of pairs of amino acids, or isobaric substitutions, Asn for Gly–Gly etc.

A more refined approach which has been used in mass spectrometry for many years and was adapted by Rose *et al.* (1983) for peptide sequencing is the use of isotopic marking of one of the termini. Rose carried out the tryptic digestion of the protein in a 50:50 mixture of H_2O^{16}:H_2O^{18} thus the C-terminus of all the peptides

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

produced (except that coming from the C-terminal of the protein) appear as peak doublets separated by two mass units. This makes the interpretation of the MS/MS spectrum easier since all y-ions will show a doublet whilst all b-ions will be singlets. The alternative approach of labelling the N-terminal by acetylation with a mixture of trideuteroacetic and acetic anhydride allows all N-terminal ions to be detected by their appearance as a doublet separated by three mass units.

7.5 *Outlook*

The development of analytical techniques for protein identification and sequencing from 2D gels has advanced rapidly in terms of speed and sensitivity. Attention now must be turned to the development of suitable methods for the structure analysis of the oligosaccharide modifications of proteins that play a very important biological role in protein activity and targeting. Sugar analysis is a difficult undertaking due to the nonlinear and conformational aspects of sugar structure. Nonetheless, advances have been made in increasing the sensitivity of sequencing (Bigge *et al.* 1995) with the development of non-selective and efficient fluorescent labelling for glycans. This labelling allows high sensitivity detection for sequencing based on exo- and endoglycosidase digestion and liquid chromatographic analysis of the products (Prime *et al.* 1996). This can be coupled with the reagent array method for fast sequencing of oligosaccharides put forward by Edge *et al.* (1992). These methods are being adapted for use in sequencing oligosaccharides isolated from proteins in 2D gels. One obvious extension of the method would be an automated release of sugars by hydrazinolysis after protein digestion within a 2D gel and the use of labelling methods to increase the sensitivity of detection of the oligosaccharides by HPLC-MS analysis.

Clearly two trends that will continue into the future are miniaturization and increased throughput. Microchips have been developed for high speed separation methods using very low amounts of material, almost down to single molecule detection by fluorescence (see Manz, 1997 for an overview). Recently the coupling of microchips to mass spectrometers has been reported (Ramsey & Ramsey, 1997; Figeys *et al.* 1997*b*). Ideally the ultimate method would be the direct coupling of a 2D gel to a mass spectrometer in an array format that would allow proteins to be directly eluted into an ion trap mass spectrometer for analysis. Recently it has been shown that it is possible to obtain fragmentation of intact proteins. Loo *et al.* (1990) showed that it was possible to obtain sequence information from MS/MS of intact proteins using a triple quadrupole mass spectrometer for proteins as large as 66 000. Mortz *et al.* (1996) showed that by using a very high sensitivity Fourier transform ion cyclotron instrument which has tremendous resolving power, sets of two or three sequence tags could be obtained for a variety of proteins which was enough to identify them uniquely in a sequence database. Perhaps in the future it will be possible to ionize protein directly from an ultra thin gel for identification by sequence tagging (Eckerskorn *et al.* 1997).

With the appearance of low cost (< $200 000), sensitive (MS/MS requiring 10 fmol) and simple mass spectrometers (fully automated, including database

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

search), the time has come for every serious protein chemistry lab to come to terms with the MS revolution. The task of DNA sequencing is now being carried out by robots and soon protein and peptide mass fingerprinting will be fully automated, leaving the study of proteins to turn back to the basics, to study how they function, especially in complex processes involving the coordinated regulation of hundreds of proteins. The protein is becoming the centre of attention again. After all, the word is derived from the Greek, 'Proteios' meaning 'of the first rank' – a position it fully deserves.

8. ACKNOWLEDGEMENTS

The author thanks Paola Dainese, Martin Münchbach and Sjouke Hoving for useful discussions and critical reading of the manuscript, and John Yates for sharing results prior to publication.

9. REFERENCES

ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMEROPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. & VENTER, J. C. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science, N.Y.* **252**, 1651–1656.

AEBERSOLD, R. (1990). High sensitivity sequence analysis of proteins separated by polyacrylamide gel electrophoresis. In *Advances in Electrophoresis* (eds A. Chrambach, M. J. Dunn and B. J. Ladula), vol. 4, pp. 81–168. Germany: VCH-Verlag, Weinheim.

AEBERSOLD, R. H., BURES, E. J., NAMCHUK, M., GOGHARI, M., SHUSHAN, B. & COVEY, T. (1992). Design, synthesis, and characterization of a protein sequencing reagent yielding amino acid derivatives with enhanced detectability by mass spectrometry. *Protein Sci.* **1**, 494–503.

AEBERSOLD, R. H., LEAVITT, J., SAAVEDRA, R. A., HOOD, L. E. & KENT, S. B. (1987). Internal amino acid sequence analysis of proteins separated by 1- or 2D gel electrophoresis after *in situ* digestion on nitrocellulose. *Proc. natn. Acad. Sci. U.S.A.* **84**, 6970–6974.

AEBERSOLD, R. H., TEPLOW, D. B., HOOD, L. E. & KENT, S. B. (1986). Electroblotting onto activated glass. High efficiency preparation of proteins from analytical SDS–PAGE gels for direct sequence analysis. *J. Biol. Chem.* **261**, 4229–4238.

AIMOTO, S., TAKAO, T., SHIMONISHI, Y., HARA, S., TAKEDA, T., TAKEDA, Y. & MIWATANI, T. (1983). Amino acid sequence of a heat-stable enterotoxin produced by human enterotoxigenic *Escherichia coli*. *Eur. J. Biochem.* **129**, 257–264.

ANDERSON, N. G. & ANDERSON, N. L. (1982). The human protein index. *Clin. Chem.* **28**, 739–748.

ANDERSON, N. L., SWANSON, M., GIERE, F. A., TOLLAKSEN, S., GEMMELL, A., NANCE, S. & ANDERSON, N. G. (1986). Effects of Aroclor 1254 on proteins of mouse liver: application of two-dimensional electrophoretic protein mapping. *Electrophoresis* **7**, 44–48.

ANDERSON, N. L., TAYLOR, J., SCANDORA, A. E., COULTER, B. P. & ANDERSON, N. G. (1981). The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin. Chem.* **27**, 1807–1820.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Andren, P. E., Emmet, M. R. & Caprioli, R. M. (1994). Micro-electrospray: Zeptomole/Attomole per microliter sensitivity for peptides. *J. Am. Mass Spectrom.* **5**, 867–869.

Appel, R. D., Bairoch, A., Sanchez, J.-C., Vargas, J. R., Golaz, O., Pasquali, C. & Hochstrasser, D. F. (1996). Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. *Electrophoresis* **17**, 540–546.

Appel, R. D., Hochstrasser, D. F., Funk, M., Vargas, J. R., Pellegrini, C., Muller, A. F. & Scherrer, J. R. (1991). The melanie project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis* **12**, 722–735.

Bailey, J. M., Nikfarjam, F., Shenoy, N. R. & Shively, J. E. (1992). Automated carboxy-terminal sequence analysis of peptides and proteins using diphenylphosphoroisothiocyanatidate. *Protein Sci.* **1**, 1622–1633.

Bailey, J. M., Tu, O., Issai, G., Ha, A. & Shively, J. E. (1995). Automated carboxy-terminal sequence analysis of polypeptides containing C-terminal proline. *Anal. Biochem.* **000**, 588–596.

Bailey, J. M. & Shively, J. E. (1990). Carboxy-terminal sequencing: formation and hydrolysis of carboxy-terminal peptidylthiohydantoins. *Biochem. J.* **29**, 3145–3156.

Barber, M., Bordoli, R. S., Sedgwick, R. D. & Taylor, A. N. (1981). Fast atom bombardment of solids: a new ion source for mass spectrometry. *J. Chem. Soc. Chem. Comm.* **000**, 325–327.

Bartels, C. (1990). Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environ. Mass Spec.* **19**, 363–368.

Bartlet-Jones, M., Jeffery, W. A., Hansen, H. F. & Pappin, D. J. (1994). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Comm. Mass Spec.* **9**, 737–742.

Bauw, G., De Loose, M., Inze, D., Van Montagu, M. & Vandekerckhove, J. (1987). Alterations in the phenotype of plant cells studied by amino terminal sequence analysis of proteins electroblotted from 2D gels. *Proc. natn. Acad. Sci. U.S.A.* **84**, 4806–4810.

Biemann, K. (1988). Contributions of mass spectrometry to peptide and protein structure. *Biomed. Environ. Mass Spec.* **16**, 99–100.

Biemann, K., Gapp, F. & Seibl, J. (1959). Application of mass spectrometry to structure problems. I. Amino acid sequence in peptides. *J. Am. Chem. Soc.* **81**, 2274–2275.

Bigge, J. C., Patel, T. P., Bruce, J. A., Goulding, P. N., Charles, S. M. & Parekh, R. B. (1995). Nonselective and efficient fluorescent labeling of glycans using 2-amino benzamide and anthranilic acid. *Analyt. Biochem.* **230**, 229–238.

Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Westermeier, R. & Postel, W. (1982). Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J. Biochem. Biophys. Meth.* **6**, 317–339.

Boyd, V. L., Bozzini, M., Zon, G., Noble, R. L. & Mattaliano, R. J. (1992). Sequencing of peptides and proteins from the carboxy terminus. *Analyt. Biochem.* **206**, 344–352.

Breen, M., Deakin, L., MacDonald, B., Miller, S., Sibson, R., Tarttelin, E., Avner, P., Bourgade, F., Guenet, J. L., Montagutelli, P., Guenet, X., Poirier, C., Simon, D., Tailor, D., Bishop, M., Kelly, M., Rysavy, F., Rastan, S., Norris, D., Sheperd, D., Abbot, G., Pilz, A., Hodge, S., Jackson, I., Boyd, Y., Blair, H.,

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Maslen, G., Todd, J. A., Reed, P. W., Stoye, J., Ashworth, A., McCarthy, C., Cox, R., Schwalkwyk, L., Lehrach, H., Klose, J., Gangadharan, U. & Brown, S. (1994). Towards high resolution maps of the mouse and human genomes – a facility for ordering markers to 0·1 cM resolution. *Human Mol. Gen.* **3**, 621–627.

Bures, E. J., Nika, H., Chow, D. T., Morrison, H. D., Hess, D. & Aebersold, R. (1995). Synthesis of the protein-sequencing reagent 4-(3-pyridinylmethylaminocarboxypropyl) phenyl isothiocyanate and characterization of 4-(3-pyridinylmethylaminocarboxypropyl) phenylthiohydantoins. *Analyt. Biochem.* **224**, 364–372.

Cash, P. (1991). The application of two-dimensional polyacrylamide gel electrophoresis to medical microbiology: molecular epidemiology of viruses and bacteria. *Electrophoresis* **12**, 592–604.

Celis, J. E., Gesser, B., Rasmussen, H. H., Madsen, P., Leffers, H., Dejgaard, K., Honore, B., Olsen, E., Ratz, G. & Lauridsen, J. B. (1990). Comprehensive two-dimensional gel protein databases offer a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis* **11**, 989–1071.

Chait, B. T., Wang, R., Beavis, R. C. & Kent, S. B. H. (1993). Protein ladder sequencing. *Science, N.Y.* **262**, 89–92.

Chuang, S. E., Daniels, D. L. & Blattner, F. R. (1993). Global regulation of gene expression in *Escherichia coli*. *J. Bacteriol.* **175**, 2026–2036.

Clackson, T., Hoogenboom, H. R., Griffiths, A. D. & Winter, G. (1991). Making antibody fragments using phage display libraries. *Nature, Lond.* **352**, 624–628.

Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. (1973). Construction of biologically functional bacterial plasmids *in vitro*. *Proc. natn. Acad. Sci. U.S.A.* **70**, 3240–3244.

Cooks, R. G., Glish, G. L., McLuckey, S. A. & Kaiser, R. E. (1991). Ion trap mass spectrometry. *Chem. Eng. News.* (March) **25**, 26–41.

Corthals, G. L., Molloy, M. P., Herbert, B. R., Williams, K. L. & Gooley, A. A. (1997). Prefractionation of protein samples prior to two dimensional electrophoresis. *Electrophoresis* **000**, 317–323.

Covey, T. R., Bonner, R. F., Shushan, B. I. & Henion, J. (1988). The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. *Rapid Comm. Mass Spectrom.* **2**, 249–256.

Dainese, P., Fischer, H.-M., Hennecke, H. & James, P. (1997*a*). Applying proteome analysis to the assignment of gene function in symbiotic nitrogen fixation in *Bradyrhizobium japonicum*. (Submitted.)

Dainese, P., Quadroni, M., Staudenmann, W. & James, P. (1997*b*). Concentration of and SDS removal from proteins isolated from multiple silver stained two dimensional electrophoresis gels. *Eur. J. Biochem.* **246**, 336–343.

Dainese, P., Staudenmann, W., Quadroni, M., Korostensky, C., Gonnet, G., Kertesz, M. & James, P. (1997*c*). Probing protein function by gene knockout and proteome analysis. *Electrophoresis* **18**, 432–443.

Danna, K. J., Sack, G. H. Jr & Nathans, D. (1973). Studies of simian virus 40 DNA. VII. A cleavage map of the SV40 genome. *J. molec. Biol.* **78**, 363–376.

Davis, M. T., Lee, T. D., Ronk, M. & Hefta, S. A. (1995*a*). Microscale immobilized protease reactor columns for peptide mapping by liquid chromatography/mass spectral analyses. *Anal. Biochem.* **224**, 235–244.

Davis, M. T., Stahl, D. C. & Lee, T. D. (1995*b*). Low flow solvent delivery system for capillary HPLC-tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **6**, 571–577.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Davis, M. T., Stahl, D. C., Hefta, S. A. & Lee, T. D. (1995*c*). A microscale electrospray interface for online capillary HPLC-tandem mass spectrometry. *Analyt. Chem.* **67**, 4549–4556.

Deterding, L. J., Moseley, M. A., Tomer, K. B. & Jorgenson, J. W. (1991). Nanoscale separations combined with tandem mass spectrometry. *J. Chrom.* **554**, 73–82.

Dixon, R., Eady, R. R., Espin, G., Hill, S., Iaccarino, M., Kahn, D. & Merrick, M. (1980). Analysis of regulation of Klebsiella pneumoniae nitrogen fixation (nif) gene cluster with gene fusions. *Nature, Lond.* **286**, 128–132.

Dole, M., Mack, L. L., Hines, R. L., Mobley, R. C., Ferguson, L. D. & Alice, M. B. (1968). Molecular beams of macroions. *J. Chem. Phys.* **49**, 2240.

Eckerskorn, C., Jungblut, P., Mewes, W., Klose, J. & Lottspeich, F. (1988). Identification of mouse brain proteins after two-dimensional electrophoresis and electroblotting by microsequence analysis and amino acid composition analysis. *Electrophoresis* **9**, 830–838.

Eckerskorn, C., Strupat, K., Karas, M., Hillenkamp, F. & Lottspeich, F. (1992). Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis* **13**, 664–665.

Eckerskorn, C., Strupat, K., Schleuder, D., Hochstrasser, D., Sanchez, J.-C., Lottspeich, F. & Hillenkamp, F. (1997). Analysis of proteins by direct scanning infrared MALDI mass spectrometry after 2D-PAGE separation and electroblotting. *Anal. Chem.* **69**, 2888–2892.

Edge, C. J., Rademacher, T. W., Wormald, M. R., Parekh, R. B., Butters, T. D., Wing, D. R. & Dwek, R. A. (1992). Fast sequencing of oligosaccharides: the reagent-array analysis method. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6338–6342.

Edman, P. (1949). A method for the determination of the amino acid sequence of peptides. *Arch. Biochem. Biophys.* **22**, 475–483.

Edman, P. & Begg, G. (1967). A protein sequenator. *Eur. J. Biochem.* **1**, 80–91.

Eng, J. K., McCormack, A. L. & Yates, J. R. (1994). Correlating tandem mass spectral data of peptides to sequences in a protein database. *J. Am. Soc. Mass Spec.* **5**, 976–989.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science, N.Y.* **246**, 64–71.

Figeys, D., Ducret, A. & Aebersold, R. (1997*a*). Identification of proteins by capillary electrophoresis-tandem mass spectrometry. Evaluation of an on-line solid-phase extraction device. *J. Chrom.* **763**, 295–306.

Figeys, D., Ducret, A., Yates, J. R. III & Aebersold, R. (1996). Protein identification by solid phase microextraction–capillary zone electrophoresis–microelectrospray–tandem mass spectrometry. *Nature, Biotechnology* **14**, 1579–1583.

Figeys, D., Ning, Y. & Aebersold, R. (1997*b*). A microfabricated device for rapid protein identification by microelectrospray ion trap mass spectrometry. *Anal. Chem.* **69**, 3153–3160.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L.,

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, *N.Y.* **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchinson, C. A. & Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, *N.Y.* **270**, 397–403.

Freiberg, C, Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A. & Perret, X. (1997). Molecular basis of symbiosis between *Rhizobium* and legumes. *Nature*, *Lond.* **387**, 394–401.

Gale, D. C. & Smith, R. D. (1993). Small volume and low flow rate electrospray ionization mass spectrometry of aqueous samples. *Rapid Comm. in Mass Spectrom.* **7**, 1017–1021.

Garrels, J. I. (1979). Two dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* **254**, 7961–7977.

Gharahdaghi, F., Kirchner, M., Fernandez, J. & Mische, S. (1996). Peptide mass profiles from PVDF bound protein digests in the presence of detergents. *Anal. Biochem.* **233**, 94–99.

Gilman, A. (1987). G proteins: Transducers of receptor-generated signals. *Ann. Rev. Biochem.* **56**, 615–649.

Griffin, P. R., MacCoss, M. J., Eng, J. K., Blevins, R. A., Aaronson, J. S. & Yates, J. R. III. (1995). Direct database searching with MALDI-PSD spectra of peptides. *Rap. Comm. Mass Spec.* **9**, 1546–1551.

Harrington, M. G., Lee, K. H., Yun, M., Zewert, T., Bailey, J. E. & Hood, L. (1993). Mechanical precision in two-dimensional electrophoresis can improve protein spot positional reproducibility. *Appl. Theor. Electrophor.* **3**, 347–353.

Henzel, W., Billeci, T., Stults, J., Wong, S., Grimley, C. & Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. natn. Acad. Sci. U.S.A.* **90**, 5011–5015.

Hewick, R. M., Hunkapiller, M. W., Hood, L. E. & Dreyer, W. J. (1981). A gas/liquid/solid phase peptide and protein sequenator. *J. Biol. Chem.* **256**, 7990–7997.

Hines, W. M., Falick, A. L., Burlingame, A. L. & Gibson, B. W. (1992). Pattern based algorithm for peptide sequence for tandem high energy collision induced energy mass spectra. *J. Am. Soc. Mass Spec.* **3**, 326–336.

Hood, L., Kent, S., Smith, L., Aebersold, R., Teplow, D., Kaiser, R., Clark-Lewis, I., Woo, D., Hines, W. & Sanders, J. (1986). The development of a facility to analyse and synthesise genes and proteins. In *Methods in Protein Sequence Analysis*, pp. 21–41 (ed. K. Walsh). Clifton, New Jersey: Humana Press.

Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E. & Engelhard, V. H. (1992). Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science*, *N.Y.* **255**, 1261–1263.

Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. (1986). Protein sequencing by tandem mass spectrometry. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6233–6237.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Inglis, A. S., Duncan, M. W., Adams, P. & Tseng, A. (1992). Formation of proline thiohydantoin with ammonium thiocyanate: progress towards a viable C-terminal amino acid sequencing procedure. *J. Biochem. Biophys. Meth.* **25**, 163–171.

James, P. (1997). Of genomes and proteomes. *Biochem. Biophys. Res. Comm.* **231**, 1–6.

James, P., Quadroni, M., Carafoli, E. & Gonnet, G. (1993). Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Comm.* **195**, 58–64.

James, P., Quadroni, M., Carafoli, E. & Gonnet, G. (1994). Protein identification in DNA databases by peptide mass fingerprinting. *Protein Sci.* **3**, 1347–1350.

Jensen, O. N., Podtelejnikov, V. & Mann, M. (1996*a*). Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rap. Comm. Mass Spec.* **10**, 1371–1378.

Jensen, O. N., Vorm, O. & Mann, M. (1996*b*). Sequence patterns produced by incomplete enzymatic digestion or one step Edman degradation of peptide mixtures as probes for protein database searches. *Electrophoresis* **17**, 938–944.

Johnson, R. S. & Biemann, K. (1989). Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed. Environ. Mass Spec.* **18**, 945–957.

Johnson, R. S. & Walsh, K. A. (1992). Sequence analysis of peptide mixtures by automated integration of Edman and mass spectrometric data. *Protein Sci.* **1**, 1083–1091.

Johnson, T. B. & Nicolet, B. H. (1911). Hydantoins: The synthesis of 2-thiohydantoin. *J. Am. Chem. Soc.* **33**, 1973–1978.

Kahn, P. (1995). From genome to proteome: looking at a cell's proteins. *Science, N.Y.* **270**, 369–370.

Karas, M. & Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60**, 2299–2301.

Kaufmann, R., Spengler, B. & Lützenkirchen, F. (1993). Mass spectrometric sequencing of linear peptides by product ion analysis using MALDI-TOF-MS. *Rapid Comm. in Mass Spect.* **7**, 902–910.

Kenrick, K. G. & Margolis, J. (1970). Isoelectric focusing and gradient gel electrophoresis: A two-dimensional technique. *Anal. Biochem.* **33**, 204–207.

Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–243.

Klose, J. & Kobalz, U. (1995). Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis* **16**, 1034–1059.

Kohara, Y., Akiyama, K. & Isono, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**, 495–508.

Korostensky, C., Staudenmann, W., Dainese, P., Gonnet, G. & James, P. (1998). An algorithm for the identification of proteins in sequence databases using peptide masses and N- or C-terminal sequence tags generated by sequential endo- and exopeptidase digestions. (Submitted).

Kovarova, H., Stulik, J., Macela, A., Lefkovits, I. & Skrabkova, Z. (1992). Using two-dimensional gel electrophoresis to study immune response against intracellular bacterial infection. *Electrophoresis* **13**, 741–742.

Kriger, M. S., Cook, K. D. & Ramsey, R. S. (1995). Durable gold coated fused silica capillaries for use in electrospray mass spectrometry. *Anal. Chem.* **67**, 385–389.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Latter, G. I., Metz, E., Burbeck, S. & Leavitt, J. (1983). Measurement of amino acid composition of proteins by computerised microdensitometry of two-dimensional electrophoresis gels. *Electrophoresis* **4**, 122–126.

Lee, N. H., Weinstock, K. G., Kirkness, E. F., Earle-Hughes, J. A., Fuldner, R. A., Marmaros, S., Glodek, A., Gocayne, J. D., Adams, M. D., Kerlavage, A. R., Fraser, C. M. & Venter, J. C. (1995). Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. natn. Acad. Sci. U.S.A.* **92**, 8303–8307.

Lefkovits, I., Frey, J. R., Kuhn, L., Kettman, J. R., Behar, G., Auffray, C., Hoffmann, J. P. & Coleclough, C. (1995). Human lymphocyte cDNA ordered library analyzed by 2D gel electrophoresis. Pooling strategy and matching of gel patterns. *Appl. Theor. Electro.* **5**, 35–42.

Lemkin, P. F. (1997). Comparing two-dimensional electrophoretic gel images across the internet. *Electrophoresis* **18**, 461–470.

Lemkin, P. F. & Lipkin, L. E. (1981). gellab: a computer system for two-dimensional gel electrophoresis analysis. III. Multiple two-dimensional gel analysis. *Comp. Biomed. Res.* **14**, 407–446.

Link, A. J., Carmack, E. & Yates, J. R. (1997*a*). A strategy for the identification of proteins localized to subcellular spaces: Application to *E. coli* periplasmic proteins. *Int. J. Mass Spec. Ion Proc.* **160**, 303–316.

Link, A. J., Hays, L. G., Carmack, E. B. & Yates, J. R. III. (1997*b*). Identifying the abundant proteins in the Haemophilus influenzae proteome. *Electrophoresis* (in press).

Loo, J. A., Edmunds, C. G. & Smith, R. D. (1990). Primary sequence information from intact proteins by electrospray ionisation tandem mass spectrometry. *Science, N.Y.* **248**, 201–204.

Lui, M., Tempst, P. & Erdjument-Bromage, H. (1996). Methodical analysis of protein–nitrocellulose interactions to design a refined digestion protocol. *Anal. Biochem.* **241**, 156–166.

Malcolm, A. D. (1978). The decline and fall of protein chemistry? *Nature, Lond.* **275**, 90–91.

Manabe, T., Oda, O. & Okuyama, T. (1982). Amino acid microanalysis of proteins extracted from spots of fixed, stained 2-dimensional gels. *J. Chrom.* **241**, 361–370.

Mann, M., Hojrup, P. & Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in databases. *Biol. Mass Spec.* **22**, 338–345.

Mann, M. & Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399.

Manz, A. (1997). Ultimate speed and sample volumes in electrophoresis. *Biochem. Soc. Trans.* **25**, 278–281.

Maxam, A. M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560–564.

McCormack, A. L., Schieltz, D. M., Goode, B., Yang, S., Barnes, G., Drubin, X. & Yates, J. R. III. (1997). Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767–776.

Medina, M. B. & Phillips, J. G. (1982). Investigations on trypsin-hydrolyzed peptides for protein identification. *J. Agricult. Food Chem.* **30**, 1250–1253.

Miller, P. E. & Denton, M. B. (1986). The quadrupole mass filter: basic operating concepts. *J. Chem. Ed.* **63**, 617–622.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Mortz, E., O'Connor, P. B., Roepstorff, P., Kelleher, N. L., Wood, T. D., McLafferty, W. & Mann, M. (1996). Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. natn. Acad. Sci. U.S.A.* **93**, 8264–8267.

Nokihara, K., Morita, N. & Kuriki, T. (1992). Applications of an automated apparatus for two-dimensional gel electrophoresis, Model TEP-01, for microsequence analysis of proteins. *Electrophoresis* **13**, 710–717.

Nowak, R. (1995). Entering the postgenome era. *Science, N.Y.* **270**, 368–369.

Numa, S. A. (1989). A molecular view of neurotransmitter receptors and ionic channels. In *The Harvey Lectures*. New York: Alan Liss Inc.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.

Oliver, N. A., Greenberg, B. D. & Wallace, D. C. (1983). Assignment of a polymorphic polypeptide to the human mitochondrial DNA unidentified reading frame 3 gene by a new peptide mapping strategy. *J. Biol. Chem.* **258**, 5834–5839.

Papayannopoulus, I. A. (1995). The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spec. Rev.* **14**, 49–73.

Pappin, D., Hojrup, P. & Bleasby, A. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **3**, 327–332.

Patterson, D. H., Tarr, G. E., Regnier, F. E. & Martin, S. A. (1995). C-terminal ladder sequencing via matrix-assisted laser desorption mass spectrometry coupled with carboxypeptidase Y time-dependent and concentration-dependent digestions. *Anal. Chem.* **67**, 3971–3978.

Patterson, S. D. (1995). Matrix-assisted laser-desorption/ionization mass spectrometric approach for the identification of gel-separated proteins in the 5–50 pmol range. *Electrophoresis* **16**, 1104–1111.

Paul, W. (1990). Electromagnetic traps for charged and neutral particles (Nobel lecture). *Agnew. Chem. Int. Ed. Engl.* **29**, 739–746.

Pederson, S., Bloch, P. L., Reeh, S. & Neidhart, F. C. (1978). Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. *Cell* **14**, 179–190.

Piccinni, E., Staudenmann, W., Albergoni, V., De Gabrieli, R. & James, P. (1994). Purification and primary structure of metallothioneins induced by cadmium in the protists *Tetrahymena pigmentosa* and *Tetrahymena pyriformis*. *Eur. J. Biochem.* **226**, 853–859.

Praxmayer, C., Murach, K. F., Baumgartner, B., Aberger, F., Schlegel, E. & Illmensee, K. (1992). Protein synthesis in murine organs during post-implantation development detected by two-dimensional gel electrophoresis. *Electrophoresis* **13**, 720–722.

Prime, S., Dearnley, J., Ventom, A. M., Parekh, R. B. & Edge, C. J. (1996). Oligosaccharide sequencing based on exo- and endoglycosidase digestion and liquid chromatographic analysis of the products. *J. Chrom.* **720**, 263–274.

Quadroni, M., Staudenmann, W., Kertesz, M. & James, P. (1996). Analysis of global responses by protein and peptide fingerprinting of proteins isolated by two-dimensional gel electrophoresis: application to the sulphate starvation response of *Escherichia coli*. *Eur. J. Biochem.* **239**, 773–781.

Rabilloud, T., Valette, C. & Lawrence, J. J. (1994). Sample application by in gel rehydration improves the resolution of two dimensional electrophoresis with immobilised pH gradients in the first dimension. *Electrophoresis* **15**, 1552–1558.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Ramsey, R. S. & Ramsey, J. M. (1997). Generating electrospray from microchip devices using electroosmotic pumping. *Anal. Chem.* **69**, 1174–1178.

Rider, M. H., Puype, M., Van Damme, J., Gevaert, K., De Boeck, S., D'Alayer, J., Rasmussen, H. H., Celis, J. & Vanderkerckhove, J. (1995). An agarose-based gel-concentration system for microsequence and mass spectrometric characterisation of proteins previously purified in polyacrylamide gels starting at low picomole levels. *Eur. J. Biochem.* **230**, 258–265.

Roepstorff, P. & Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Environ. Mass Spectrom.* **11**, 601–603.

Rose, K., Simona, M. & Offord, R. (1983*a*). Amino acid sequence determination by g.l.c. – mass spectrometry of permethylated peptides. Optimization of the formation of chemical derivatives at the 2–10 nmol level. *Biochem. J.* **215**, 261–272.

Rose, K., Simona, M., Offord, R., Prior, C. P., Otto, B. & Thatcher, D. R. (1983*b*). A new mass-spectrometric C-terminal sequencing technique finds a similarity between gamma-interferon and alpha 2-interferon and identifies a proteolytically clipped gamma-interferon that retains full antiviral activity. *Biochem. J.* **215**, 273–277.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science, N.Y.* **239**, 487–491.

Sanger, F. (1945). The free amino groups of insulin. *Biochem. J.* **39**, 507–512.

Sanger, F. (1981). Determination of nucleotide sequences in DNA. *Science, N.Y.* **214**, 1205–1210.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. & Smith, M. (1977). Nucleotide sequence of bacteriophage *Φ*174 DNA. *Nature, Lond.* **265**, 687–695.

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, N.Y.* **270**, 467–470.

Schlack, P. & Kumpf, W. (1926). Über eine neue Methode zur Ermittlung der Konstitution von Peptiden. *Z. Physiol. Chem.* **154**, 125–130.

Schubart, U. K. & Danoff, A. (1987). Identification in rat brain of a 19-kDa protein that comigrates on two-dimensional electrophoresis with p19, a hormonally regulated phosphoprotein of insulinoma cells. *Biochem. Biophys. Res. Comm.* **146**, 410–415.

Sepetov, N. F., Issakova, O., Lebl, M., Swiderek, K., Stahl, D. C. & Lee, T. (1993). The use of Hydrogen–Deuterium exchange to facilitate peptide sequencing by electrospray tandem mass spectrometry. *Rapid Comm. in Mass Spectrom.* **7**, 58–62.

Shelly, D. C., Gluckman, J. C. & Novotny, M. V. (1984). Dead-volume free termination for packed columns in microcapillary liquid chromatography. *Anal. Chem.* **56**, 2990–2292.

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. 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. natn. Acad. Sci. U.S.A.* **93**, 14440–14445.

Simpson, R. J., Moritz, R. L., Begg, G. S., Rubira, M. R. & Nice, E. C. (1989). Micropreparative procedures for high sensitivity sequencing of peptides and proteins. *Anal. Biochem.* **177**, 221–236.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

Smith, B. J. & Tempst, P. (1997). Strategies for handling polypeptides on a microscale. *Meth. Molec. Biol.* **64**, 1–16.

Spengler, B., Lützenkirchen, F. & Kaufmann, R. (1993). On target deuteration for peptide sequencing by laser mass spectrometry. *Organic Mass Spectrometry* **28**, 1482–1490.

Stahl, D. C., Swiderek, K. M., Davis, M. & Lee, T. D. (1996). Data controlled automation of liquid chromatography/tandem mass spectrometry (LC/MC/MC) analysis of peptide mixtures. *J. Am. Soc. Mass Spectrom.* **6**, 532–540.

Stark, G. R. (1968). Sequential degradation of peptides from their carboxyl termini with ammonium thiocyanate and acetic anhydride. *Biochemistry* **7**, 1796–1807.

Steinbeck, R., Steinbeck, C., Postel, E., Busse, H., Havsteen, B. & Auer, G. (1984). High-resolution 2-dimensional protein mapping of 'diploid' and 'aneuploid' mammary adenocarcinomas. *Histochemistry* **84**, 338–341.

Taylor, J., Anderson, N. L., Scandora, A. E., Villard, K. E. & Anderson, N. G. (1982). The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin. Chem.* **28**, 861–866.

Taylor, J. A. & Johnson, R. S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Comm. Mass Spec.* **11**, 1067–1075.

Tomlinson, A. J., Benson, L. M., Jameson, S. & Naylor, S. (1996). Rapid loading of large sample volumes, analyte cleanup, and modified moving boundary transient isotachophoresis conditions for membrane preconcentration-capillary electrophoresis in small diameter capillaries. *Electrophoresis* **17**, 1801–1807.

Tsugita, A., Kamo, M., Jone, C. S. & Shikama, N. (1989). Sensitization of Edman amino acid derivatives using the fluorescent reagent, 4-aminofluorescein. *J. Biochem.* **106**, 60–65.

Valaskovic, G. A., Kelleher, N. L. & McLafferty, F. W. (1996). Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science, N.Y.* **273**, 1199–1202.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science, N.Y.* **270**, 484–487.

Vestal, M. L., Juhasz, P. & Martin, S. A. (1995). Delayed extraction matrix assisted laser desorption time of flight mass spectrometry. *Rapid Comm. Mass Spec.* **9**, 1044–1050.

Vilkas, E. & Lederer, E. (1968). N-methylation of peptides by the method of Hakomori: structure of mycoside Cbl. *Tetrahedron Lett.* **000**, 3089–3092.

Vorm, O. & Mann, M. (1994). Improved mass accuracy in MALDI-TOF MS of peptides. *J. Am. Soc. Mass Spec.* **5**, 955–958.

Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R, Duncan, M. W., Harris, R., Williams, K. L. & Humphery-Smith, I. (1995). Progress with gene-product mapping of *Mycoplasma genitalium*. *Electrophoresis* **16**, 1090–1094.

Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F. & Williams, K. L. (1995). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotech. Gen. Eng. Rev.* **13**, 19–50.

Wilkins, M. R. & Williams, K. L. (1997). Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J. Theor. Biol.* **186**, 7–15.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press

WILM, M. S. & MANN, M. (1994). Electrospray and Taylor–Cone theory, Dole's beam of macromolecules at last? *Int. J. Mass Spectrom. Ion Proc.* **136**, 167–180.

WILM, M. S., NEUBAUER, G. & MANN, M. (1996*a*). Parent ion scans of unseparated peptide mixtures. *Anal. Chem.* **68**, 527–533.

WILM, M., SHEVCHENKO, A., HOUTHAEVE, T., BREIT, S., SCHWEIGERER, L., FOTSIS, T. & MANN, M. (1996*b*). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature, N.Y.* **379**, 466–469.

WIMMER, K., KUICK, R., THORAVAL, D. & HANASH, S. M. (1996). Two-dimensional separations of the genome and proteome of neuroblastoma cells. *Electrophoresis* **17**, 1741–1751.

YATES, J. R., ENG, J., CLAUSER, K. R. & BURLINGAME, A. L. (1996). Searching sequence databases with uninterpreted high-energy CID spectra of peptides. *J. Am. Soc. Mass Spectrom.* **7**, 1089–1098.

YATES, J. R., ENG, J. K. & MCCORMACK, A. L. (1995*a*). Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67**, 3202–3210.

YATES, J. R., ENG, J. K., MCCORMACK, A. L. & SCHIELTZ, D. (1995*b*). Correlating tandem mass spectra of modified peptides to sequences in a database. *Anal. Chem.* **67**, 1426–1436.

YATES, J. R., GRIFFIN, P. R. & HOOD, L. E., (1991). In *Techniques in Protein Chemistry* (ed. J. Villafranca), vol. 2, pp. 477–486. San Diego: Academic Press.

YATES, J. R., SPEICHER, S., GRIFFIN, P. R. & HUNKAPILLAR, T. (1993). Peptide mass maps – a highly informative approach to protein identification. *Anal. Biochem.* **214**, 397–408.

https://doi.org/10.1017/S0033583597003399 Published online by Cambridge University Press