Protein Identification: The Origins of Peptide Mass Fingerprinting

William J. Henzel and Colin Watanabe
Protein Chemistry Department and Bioinformatics Department, Genentech, Inc., South San Francisco, California, USA

John T. Stults
Analytical Sciences Department, Biospect, Inc., South San Francisco, California, USA

Peptide mass fingerprinting (PMF) grew from a need for a faster, more efficient method to identify frequently observed proteins in electrophoresis gels. We describe the genesis of the idea in 1989, and show the first demonstration with fast atom bombardment mass spectrometry. Despite its promise, the method was seldom used until 1992, with the coming of significantly more sensitive commercial instrumentation based on MALDI-TOF-MS. We recount the evolution of the method and its dependence on a number of technical breakthroughs, both in mass spectrometry and in other areas. We show how it laid the foundation for high-throughput, high-sensitivity methods of protein analysis, now known as proteomics. We conclude with recommendations for further improvements, and speculation of the role of PMF in the future. (J Am Soc Mass Spectrom 2003, 14, 931–942) © 2003 American Society for Mass Spectrometry

A protein may be defined as a set of amino acids arranged in a specific sequence to yield a defined activity or property. Although some proteins may have a high degree of homology—sequence similarity—with other proteins, some, if not many portions of any one protein's sequence are unique. If a protein could be cut in a predictable manner, the sizes of the pieces should form a fingerprint for that protein. Further, if each entry in a database of protein sequences could be cut in the same manner *in silico*, the fingerprint would serve to identify the protein. This hypothesis was the basis for the first experiments in what would become commonly known as peptide mass fingerprinting.

The Quest for Faster Protein Sequence Analysis, 1989

The primary driving force in developing peptide mass fingerprinting was to increase the speed of protein analysis. In 1989, automated Edman degradation had a cycle time of nearly one hour per amino acid residue. Samples of interest often contained complex mixtures of proteins, which usually required separation by SDS-PAGE followed by electroblotting onto a PVDF membrane [1]. Proteins were detected by a variety of staining methods, most commonly Coomassie blue. Bands were excised from the membrane and directly sequenced in an automated protein sequencer. Proteins that co-purified with the protein of interest required significant instrument time and expense to determine their identity. Many investigators observed similar contaminants, such as serum albumin, often on a frequent basis, and the sequences of a number of these proteins began to appear in the Dayhoff database [2]. The ability to identify proteins rapidly would enable more efficient use of protein sequencer time for the analysis of novel proteins. The concept of peptide mass fingerprinting for protein identification was based on the assumption that commonly encountered protein contaminants were generally abundant proteins that would have known sequences. At the time, few protein sequences were known, relative to today's database size.

In 1989, fast atom bombardment ionization (FAB) [3] was the most widely utilized method for the characterization of peptides by mass spectrometry. We used a JEOL tandem high-resolution sector instrument, operated in single MS mode, to analyze the ions produced by the FAB source. Mass accuracy was approximately 0.2 Da for most tryptic peptides containing <30 residues. On average 0.1–1 nanomole of peptide was re-

Published online July 24, 2003

Address reprint requests to Dr. J. T. Stults, Analytical Sciences Department, Biospect, Inc., 951 Gateway Blvd. Suite 3B, South San Francisco, CA 94080, USA. E-mail: jstults@bioscpect.com

1044-0305/03/$30.00
doi:10.1016/S1044-0305(03)00214-9

Received March 4, 2003
Revised March 31, 2003
Accepted March 31, 2003

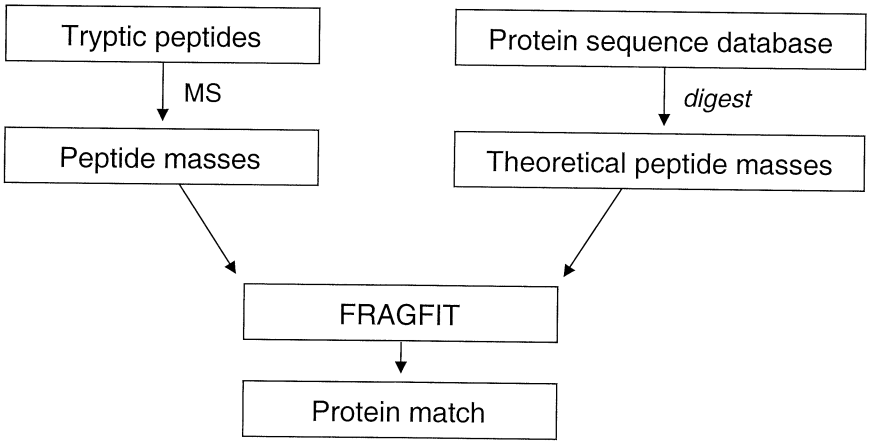


Figure 1. A flowchart illustrating the concept of peptide mass fingerprinting. The masses of peptides from a database are compared with experimentally determined masses using the program FRAGFIT.

quired to obtain a high quality signal compared to 10 picomoles for the Applied Biosystems 470A gas-phase protein sequencer, which in 1989 had become the most widely utilized instrument for protein sequence determination. We designed a computer program, called FRAGFIT, to test our hypothesis that peptide masses derived from using an endoprotease or chemical cleavage method that cleaved a protein at specific sites would be sufficient to identify a protein (see Figure 1). The program used a protein sequence database, from which each protein was cleaved into hypothetical peptides based on the specificity of the enzyme or chemical cleavage reagent (Figure 2). The masses of the hypothetical peptides were calculated and compared with the experimental masses measured by FAB mass spectrometry obtained from the unseparated peptide fragments. The protein that contained the largest number of masses that matched the measured masses obtained the highest ranking. The FRAGFIT program performed calculations "on the fly", not requiring a precompiled database of peptide fragments. This allowed the use of a current database, which in our research facility was updated on a daily basis.

We utilized several commercially available proteins to test our concept of peptide mass fingerprinting. Figure 3a shows the FAB mass spectrum of an Asp-N endoprotease digest of lysozyme. The three peptide masses that were detected in the FAB spectrum were input to the FRAGFIT program and the output obtained is shown in Figure 3b. The program matched all three masses to chicken lysozyme. We also analyzed a CNBr cleavage of horse cytochrome *c* by FAB mass spectro-

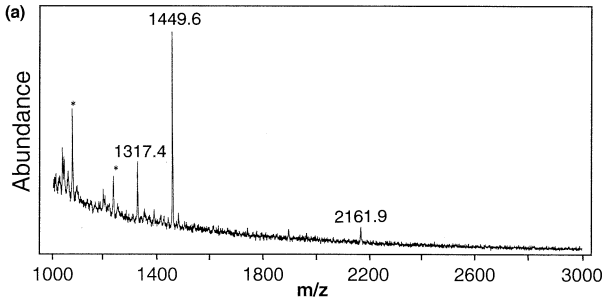
```
FRAGFIT -- find database proteins containing specified protonated molecules

Mass of fragment(s) MH+:  ____________________________________

Options: _
e enzyme: CNBr____________     w average or monoisotopic mol. weights [a,m]: a
t tolerance [.001-5.]: .3      o output file: fragfit.out   b background [y/n]: y
================================================================================
   e  enzyme or proteolytic agent.  Type the number next to the
      name, e.g., "1" for trypsin.

          1.   Trypsin I (C-side of Lys, Arg)
          2.   Trypsin II (C-side of Arg)
          3.   Chymotrypsin (C-side of Phe, Tyr, Trp)
          4.   S. AureusV8 I (C-side of Glu)
          5.   S. AureusV8 II (C-side of Glu, Asp)
          6.   CNBr (C-side of Met)
          7.   Trypsin + V8 II (C-side of Lys, Arg, Glu, Asp)
          8.   Hydroxylamine (Asn'Gly)
          9.   Pro endopeptidase (C-side of P)
         10.   Trp cleavage (C-side of Trp)
         11.   Lys-C (C-side of Lys)
         12.   Asp-N (N-side of Asp)
```

Figure 2. The input page from the original FRAGFIT program. The input consisted of a list of masses, the selection of a specific enzyme or chemical cleavage reagent, mass tolerance, and use of either monoisotopic or average mass.

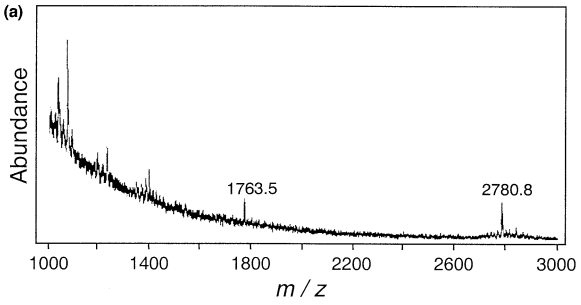


(b) enzyme: Asp-N (N-side of Asp)
Mass of MH+: 1317.400 1449.600 2161.900 (tol: 1.000)
LZCH Lysozyme c (EC 3.2.1.17) precursor - Chicken
2162.444 84: DGRTPGSRNLCNIPCSALLSS
1449.706 105: DITASVNCAKIVS
1317.552 137: DVQAWIRGCRL

Figure 3. The FAB spectrum (**a**) of a 250 pmol tryptic digest of Asp-N digest of lysozyme and (**b**) the FRAGFIT output page showing a match with chicken egg white lysozyme obtained using the masses from the FAB spectrum.

metry and obtained two masses (Figure 4a). The two masses observed were sufficient to identify the protein as cytochrome c and permitted the identification of the species (Figure 4b). At the time this search was performed, the database contained nearly a hundred different species of cytochrome *c* and more than 20,000 proteins. It was remarkable that a fingerprint comprised of only two peptides was sufficient to both identify the protein and the species.

We investigated the composition of our database and found that some amino acids including methionine, tryptophan, and cysteine were less abundant. Thus, peptides resulting from dilute acid and hydroxylamine cleavages were on average very long with masses



(b) enzyme: CNBr (C-side of Met)
Mass of MH+: 1763.500 2780.800 (tol: 0.600)
CCHO Cytochrome C - Horse
1764.031 66: EYLENPKKYIPGTKM
2781.268 81: IFAGIKKKTEREDLIAYLKKATNE
CCHOD Cytochrome C - Donkey and common zebra
(tentative sequences)
1764.031 66: EYLENPKKYIPGTKM
2781.268 81: IFAGIKKKTEREDLIAYLKKATNE

Figure 4. The FAB spectrum (**a**) of a 500 pmol CNBr cleavage of horse heart cytochrome *c* and (**b**) the FRAGFIT output page showing a match with cytochrome *c* obtained using the masses from the FAB spectrum. The output included all proteins that matched the mass list, based on the search criteria.

Table 1. Distrubution of peptide fragment length from 20,639 proteins

Enzyme/reagent	Residues cleaved	Total fragments	Avg. fragment length
Trypsin	K/R	662,981	8
Lys-C	K	359,140	16
Asp-N	D	321,655	18
CNBr	M	150,605	38
Hydroxylamine	N-G	36,643	152
Dilute acid	D-P	35,574	166

exceeding 10,000 Da (see Table 1). Longer peptides provided more specificity in a database search thus requiring fewer peptides for a match. However, longer peptides also increased the possibility that a peptide would contain a post-translational modification, or that the sequence would contain an error, both of which would preclude a match in the database.

We also performed some of these experiments with plasma desorption mass spectrometry (PDMS) [4]. Although the sensitivity was as good or better than FAB, particularly for larger peptides, both the mass accuracy and the resolving power of the time-of-flight instrument was substantially lower than the sector instrument used for FAB-MS. Thus, our early demonstrations mainly utilized FAB spectra.

The results from these and other experiments demonstrated that peptide mass fingerprinting was a rapid method that was useful in identifying known proteins. This work was presented at the 1989 meeting of the Protein Society in Seattle [5] and it received considerable attention. Although we had demonstrated the utility of this approach, the lack of sufficient sensitivity of FAB mass spectrometry prevented us from implementing this method as a routine tool. It should be noted that Laemmli and coworkers first demonstrated by SDS-PAGE the concept of a proteolytic peptide pattern that is characteristic of a protein [6].

At the same time, we also developed a program, called "SEQSORT", that was able to sort mixture sequences obtained from automated Edman sequencing data and was later published [7]. The program generated a matrix of sequences using all possible combinations of adjacent amino acids observed in the Edman sequence data. This sequence list was used to search a protein sequence database. With this program we were able to identify proteins that co-migrated in a single band on a SDS-PAGE gel, in the absence of mass spectral data. Figure 5 shows the results obtained using the SEQSORT program with data obtained from automated sequencing of a fraction that contained two components. Automated Edman sequencing data can be used to sort sequences only when the proteins in a mixture are present in different amounts. In the example shown in Figure 5 not all the residues could be sorted based on differences in amino acid concentration. When all the amino acid residues found by Edman sequencing was used as input into the SEQSORT pro-

(a)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	X	L	M	P	F	S	P	Y	L	P	L	(M)	F	N	A	M	X	Q	P
		9.1	16.5	8.5	9.3	3.0	6.5	5.4	5.4	5.8	5.7		3.6	2.5	3.8	4.2		4.4	3.0
	X	Q	T	V	S	D	N	E	E	Q	E		(S)	H	Q	G		K	Y
		6.4	5.5	6.6	2.1		3.8		5.0	4.4	2.9				2.0	1.8		2.6	1.1

(b) Output

```
Tue Oct 10 15:39:33 1989
Motif in file test2: [LQ][MT][VP][FS][SD][PN][YE][EL][PQ][LE]
Mismatches: 0
March 31, 1989, Release 20.0, 22105 sequences
GEN2289 Complement-associated protein SP-40,40 - human, mw: 52494
    24 QTVSDNELQE
GEN2289 Complement-associated protein SP-40,40 - human, mw: 52494
   229 LMPFSPYEPL
```

Figure 5. Protein mixture sequencing in 1989 using the SEQSORT program. (**a**) The results obtained by automated Edman sequencing showing the presence of two components at similar concentrations. The amino acid(s) observed during each cycle (1–19) of Edman degradation are shown. An "X" is an uncertain identification. Amino acids in parentheses are tentative calls. The value below each amino acid is the measured quantity in pmols. (**b**) The output of the SEQSORT program from a protein database showing that the two sequences matched to two different regions of the same protein.

gram, two sequences were found that matched to a complement-associated protein. This program, and the relative high-sensitivity of automated Edman protein sequence analysis diminished the importance of PMF for protein identification at the time.

Mass Spectrometry Evolves, 1992

At nearly the same time as the first demonstration of peptide mass fingerprinting, a rapid evolution began in mass spectrometry. Fast atom bombardment, and to a lesser extent PDMS, were the primary techniques for the production of ions from large, non-volatile molecules such as peptides and small proteins. However, these techniques required heroic effort to produce ions larger than about 20 kDa, and required fairly large amounts of material, often on the order of one nanomole. More importantly for peptide mass fingerprinting, hundreds of picomoles of peptides were often needed to produce a mass spectrum, even for smaller peptides. As a result, we only occasionally used peptide mass fingerprinting after the initial demonstration of the method.

Two new ionization techniques, electrospray ionization (ESI) [8] and matrix assisted laser desorption/ionization (MALDI) [9, 10], quickly eclipsed the performance of FAB and PDMS. Each of these new techniques provided subpicomole limits of detection and a mass range in excess of 100 kDa. As Figure 6 shows, these techniques provided not just an incremental improvement in performance, but a dramatic leap. The improvements in the limits of detection, in particular, quickly made a new set of biological problems accessible by mass spectrometry.

Although both ionization techniques were introduced prior to our initial demonstration of protein identification, it was not until about 1991 that the first commercial instruments made them widely available. Furthermore, these new techniques could be implemented on instruments such as quadrupoles and time-of-flight mass analyzers, which were relatively inexpensive compared with the sector instruments that had been routinely used for FAB measurements.

The impact of these new techniques for peptide mass fingerprinting soon proved to be significant. The amount of protein required for proteolytic digestion and mass measurement was substantially reduced. More importantly, the quantity of protein required was

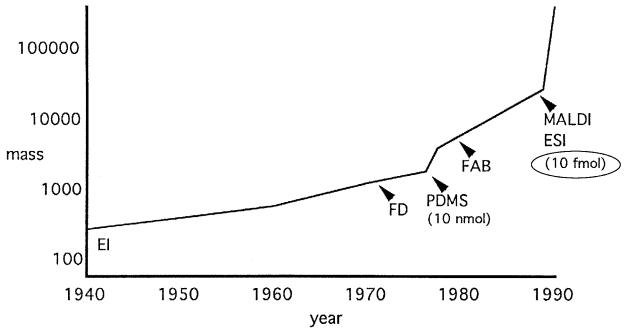


Figure 6. The chronology of mass spectrometry ionization techniques, showing typical mass ranges and detection limits over the last 60 years. EI is electron ionization, FD is field desorption.

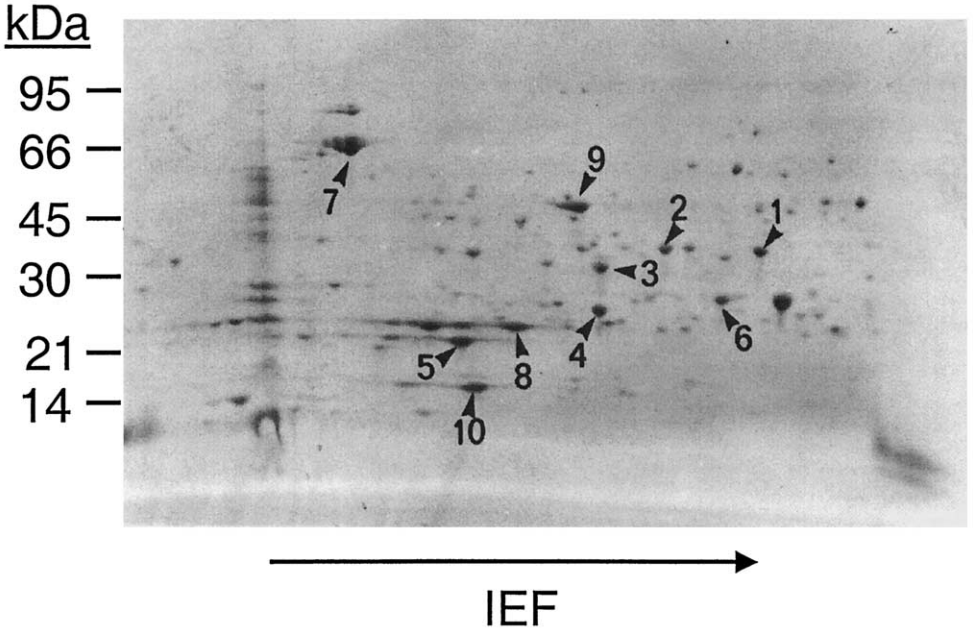
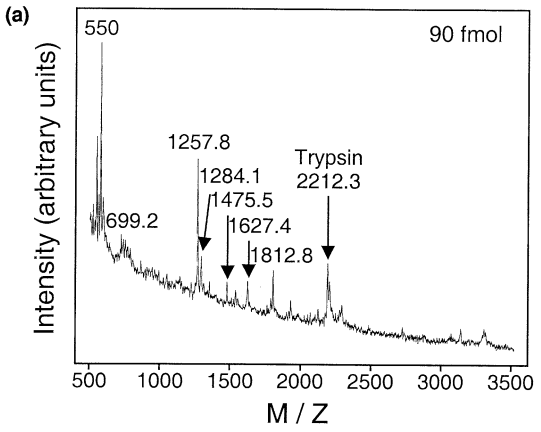


Figure 7. A two-dimensional gel of an *E. coli* cell lysate showing spots that were used for mass spectrometry analysis. The proteins were stained with Coomassie blue after electroblotting on a PVDF membrane.

now in an amount that one might encounter in proteins derived from cell lysis and separated by gel electrophoresis (low picomoles). In our early experiments with MALDI, we showed that solution digests of recombinantly expressed proteins could be measured, at that time with a binary matrix of 2,5-dihydroxybenzoic acid (DHB) combined with fucose [11]. These measurements were performed with the first version of the Vestec time-of-flight instrument, which utilized a single sample target introduced manually through a vacuum lock, and a manually adjustable laser position and fluence.

The dramatic increase in sensitivity of MALDI compared to FAB mass spectrometry encouraged us to reinvestigate peptide mass fingerprinting as a tool for low-level protein identification. We chose for a demonstration the 2-dimensional gel electrophoresis separation of a cell lysate of *E. coli*, a complex mixture of proteins (Figure 7) [12]. These cells were readily available as the by-product of protein expression experiments at Genentech.

The proteins were blotted to poly(vinylidene difluoride) (PVDF) and stained with Coomassie Blue. Ten spots were excised with a razor blade, reduced and alkylated with iodoacetic acid, and then digested with trypsin. This blotting-digestion procedure was identical to that used in experiments routinely performed at the time to generate peptides for Edman sequencing from blocked proteins [13, 14]. A 10% aliquot of the digests was analyzed by MALDI mass spectrometry, again using the DHB/fucose binary matrix. Figure 8a shows the MALDI mass spectrum of a 10% aliquot of spot number 1 that was estimated to be approximately 90 fmol. The mass accuracy of linear MALDI-TOF-MS without internal standards was relatively poor by current standards; however, limiting the search by speci-



(b)

Enzyme: Trypsin I (C-side of Lys, Arg; next not Pro)
Mass of MH+: **1257.800 1284.100 1475.500 1627.400 1812.800**
Tolerance: 4.000
Number of misses allowed:
Cys -> CarboxymethylCys
Molecular weight range: 500-60000

CYSK_ECOLI Cysteine synthase a - *E. coli* (34358.13 Da)

1813.064	0.26	3: IFEDNSLTIGHTPLVR
1257.473	0.33	105: ALGANLVLTEGAK
1627.924	0.52	226: IQGIGAGFIPANLDLK
1474.654	0.85	246: VIGITNEEAISTAR
1284.498	0.40	293: NIVVILPSSGER

Figure 8. (**a**) MALDI spectrum of peptides from an on-membrane tryptic digestion of spot number 1 obtained from the 2-D gel shown in Figure 7. (**b**) The FRAGFIT output showing that all five masses used for the search matched with cysteine synthase from *E. coli*.

fying the species still allowed accurate identification. Note also that the resolving power was insufficient to resolve the carbon isotope peaks. Duplicate 2-D gels were generated and spots were analyzed independently by Edman degradation to estimate the amount of protein present in each spot as well as to confirm independently the identity of the protein predicted by the FRAGFIT program. The result of the FRAGFIT database search using the masses obtained from spot number 1 is shown in Figure 8b. Only one protein matched the five masses input to the FRAGFIT program using a protein database of over 100,000 proteins.

These experiments illustrated a number of advantages of using PMF for protein identification. Unlike automated Edman degradation, there is no requirement for a free amino terminus. Because a protein frequently can be identified from a subset of the peptide masses, the correct protein may be identified even if some of the peptides contain post-translational modifications, or if the sequence database entry contains errors. Mass differences between the measured and predicted peptides may provide evidence for the identity of peptide modifications. The individual proteins in a mixture of up to three or four components can be identified. The sensitivity for the technique is high, with modern instrumentation able to measure peptide mixtures at the subfemtomole level. Efficient sample handling is the key—detection is generally limited by the ability to digest and recover small quantities of protein that can now be performed with tens of femtomoles in the best laboratories. The high sensitivity for mass measurement means that only an aliquot of the digested protein is required for mass analysis, and the remainder can be used for alternative measurements, for instance, to improve the confidence in the identification or to learn more about post-translational modifications. The method is easily automated for high throughput operation, and robotic workstations are now available for all of the steps.

There are limitations to protein identification by peptide mass mapping. The protein sequence, obviously, must be present in a database. Even for organisms with "complete" genome sequences, particularly eukaryotic, the entire list of all actual protein sequences is not yet available because gene and alternative splicing prediction, and other mechanisms that affect the amino acid sequence, are still far from infallible. This problem may continue for some years to come. Cross-species identification is only possible for proteins with large amounts of sequence *identity*; homology is not sufficient. Protein isoforms and alternatively spliced proteins may not be distinguished if the unique sequence regions are not observed in the peptide mixture. Proteins that have extensive post-translational modifications may fail to yield good matches. The individual components from mixtures of more than three or four proteins are difficult to identify.

At the time our experiments for protein identification were in progress, four other research groups were independently pursuing similar approaches. The results for these efforts in the laboratories of Peter Roepstorff [15], Darryl Pappin [16], Peter James [17], and John Yates [18] were all published in 1993. It is worth noting that the term "peptide mass fingerprinting" was first used in the paper by Pappin and coworkers [16].

The initial demonstrations of peptide mass fingerprinting utilized MALDI-MS spectra, because of the ease with which spectra could be obtained from small amounts of unseparated digest mixtures. Demonstrations of PMF with electrospray ionization appeared soon thereafter. We showed in 1994 that some of the same protein digests from the 2-D gel of the *E. coli* lysate, described above, could also be identified by capillary LC-MS [19]. The separation step, in fact, significantly improved the number of peptides observed, and hence yielded greater sequence coverage and improved protein identification.

Electrospray ionization also was the platform for the first demonstrations that a protein could be identified from a single MS/MS fragment ion spectrum of a peptide from proteolytic digestion. Matthias Mann and coworkers [20] showed that a short sequence, along with the fragment ion masses that denoted the beginning and end of the sequence, constituted a "sequence tag" that was useful for protein identification. A short sequence of two to four amino acid residues was often easily found in the fragment ion spectra of doubly-charged tryptic peptides produced by triple quadrupole mass spectrometers; a set of prominent y-ions were commonly observed above the m/z of the precursor mass. The sequence tag approach, importantly, allowed for error tolerance in the sequence database. At the same time, John Yates and colleagues [21] used a very different approach, based on cross-correlation of a predicted spectrum with the actual fragment ion spectrum, to identify the protein. Their program, named SEQUEST, allowed completely automated protein identifications from a set of tandem MS/MS spectra. Both the Mann and Yates approaches led, under some conditions, to the identification of a protein from a single peptide. This ability became the basis for "shotgun" proteomics years later [22].

By 1995 mass spectrometry had become an integral tool for determining protein identity. Automated Edman degradation continued to be the main methodology utilized for sequence determination of proteins not present in a protein database. Obtaining sufficient sequence for cDNA cloning on proteins at the low picomole amounts was still a challenge. To meet this need, we developed another program called "MOLWFIT" which utilized masses obtained from MALDI mass analysis and Edman sequence data of capillary reversed-phase separated fractions after endoprotease digestion [7]. The masses and sequences obtained were used as input for this program. All sequences that were derived from combinations of adjacent amino acids that fit within a given mass tolerance to the measured mass are generated. Figure 9a shows the automated Edman

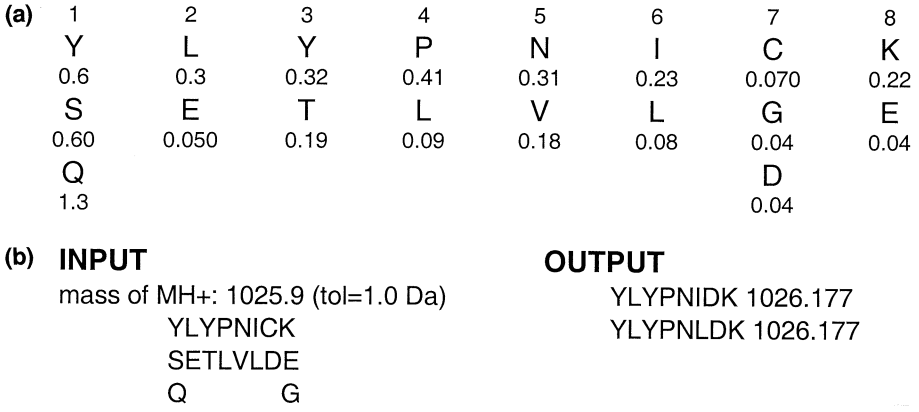


Figure 9. The MOLWFIT program for sorting peptide sequences with mass. (**a**) The automated Edman data obtained from a Lys-C peptide fraction of IL STAT (96 kDa). All values are in picomoles. (**b**) The output of the MOLWFIT program using as input a measured mass and the mixture of amino acids found at each cycle in the automated Edman degradation.

sequence analysis of a tryptic fraction containing 2 peptides. The fraction was analyzed by MALDI-TOF mass spectrometry and only one mass was found. Using the mass observed and the amino acids detected in the Edman sequence data, the "MOLWFIT" program found two possible sequences that fit the measured mass. These two sequences were identical except for residue 7 which could be an isoleucine or a leucine since these amino acids are isobaric. This program enabled us to extend the interpretation of Edman sequence data. We were able to generate cDNA probes from the use of this program, which was not possible using the Edman data alone. This further demonstrated the importance of mass spectrometry as a tool for the characterization of novel proteins.

Protein Analysis Evolves—Proteomics, 1996

By the mid-1990s, a variety of approaches were in common use for protein identification, using one or more methods for protein separation, protein digestion, peptide separation, mass analysis, and database searching [23]. The Human Genome Project was well under way by the mid-1990s, and it attracted considerable attention in both scientific circles and in the popular press. The concept of measuring all the proteins produced in an organism had been proposed in the early 1980s by Anderson and coworkers [24], yet the idea lay dormant for years while the technological capabilities necessary for such an endeavor matured. The concept of analyzing all the proteins (gene products) produced by a genome gained momentum, in part, with a new name. The term "proteome" (the PROTEin complement of a genOME) was coined by Marc Wilkins in 1995 [25]. The term, and the concept of its complete analysis, initially was not embraced enthusiastically. It was a few years later that the analysis of a proteome became commonly known as "proteomics". As interest in proteomics increased, so did the interest, particularly among suppliers who saw it as a profit center, in including nearly all protein analytical work under the umbrella of proteomics.

Proteome analysis is an integrated approach that includes not just qualitative identification of components, but also quantitative measurements. Measurements of protein differences that correspond to a biological change (e.g., receptor signaling, cell cycle progression, onset of disease) can yield insight to the underlying biological processes. We undertook a study of proteins that changed with the onset of cardiac hypertrophy, associated with congestive heart failure, in a model system based on drug-induced non-mitotic growth of cardiac myocytes in culture [26]. We used multiple 2-D gels to find proteins that showed over- or under-expression in the hypertrophied cells, relative to controls (Figure 10). This was one of the first proteomics studies to show statistically significant differences in protein abundance patterns. The protein identifications were based on a combination of PMF and tandem mass spectrometry.

Subsequent quantitative reverse transcript-polymerase chain reaction (RT-PCR) studies for the same myocyte cell culture system (unpublished data) revealed advantages and disadvantages of protein-based measurements relative to cDNA-based experiments. The protein experiments revealed changes due to post-translational modification (two myosin light chain spots changed in pI, presumably due to a change in phosphorylation state), as well as protein level changes that did not correlate with the message level. Conversely, transcript measurements revealed changes below the detection limit of gel staining, as well as changes for transcripts whose gene products are predicted to lie outside the mass and pI ranges of conventional 2-D gels.

Many technological innovations in the mid-1990s laid the way for considerable improvement in the practice of PMF, in its application to ever-increasingly

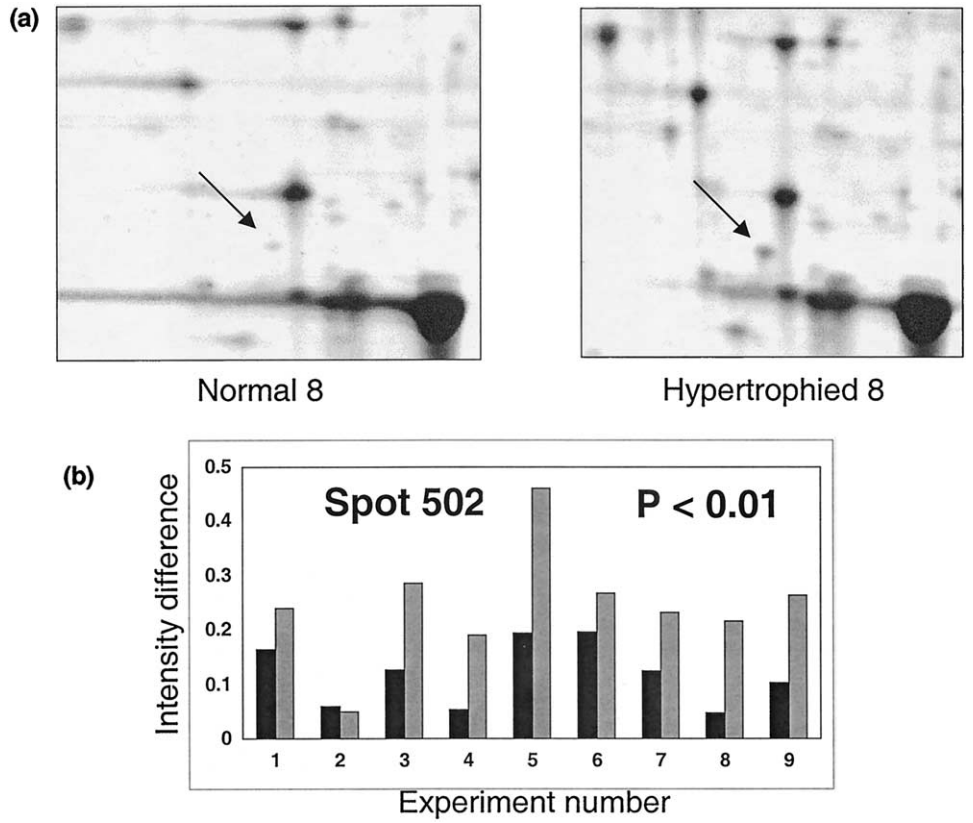


Figure 10. Enlarged sections of 2-D gels (**a**) comparing normal and hypertrophied cardiac myocytes, and (**b**) the pair-wise comparison of staining intensity for the highlighted spot in the gel. Graphical results are shown from nine experiments: normal (solid bars) and hypertrophied (hash bars) cells. The arrows in (**a**) point to spots for which staining intensities are shown graphically for experiment 8 in (**b**). The experiments together showed a 2.1-fold increase ($p < 0.01$) in the protein, identified as myosin light chain 2, atrial isoform.

important problems in biological research, and to proteomics in general.

Computation tools were becoming widespread, and the improvements in performance, in speed, and a reduction in the cost of computers made ubiquitous the "desktop" PC. Communication changed dramatically as email became the primary method of communication in scientific circles. The World Wide Web provided, initially, widespread access to collections of information and data, and soon thereafter, access to many tools and programs. Search engines for protein identification could be accessed on the web, with early sites including Protein Prospector (UCSF), ProFound (Rockefeller University), Mascot (Matrix Science), and others.

The size of sequence databases was growing enormously. The sequencing of short pieces of cDNA, expressed sequence tags (EST's), progressed rapidly and the amount of DNA sequence in Genebank grew from 217 million bases to 3.8 billion bases between 1994 and 1999. The first complete genomes were determined for the micro-organisms *H. influenzae* [27], and *E. coli* [28]. To make use of the increasing amount of genomic sequence, mass spectral-based identification algorithms were adapted to search DNA databases. The adaptation included 6-frame translation, since it could not be certain that the correct reading frame was known (or stayed in-frame due to sequencing errors) or the correct strand was used. Peptide mass fingerprinting has been used successfully with DNA databases for microorganisms. However, due to the much larger amount of sequence for eukaryotic genomes, it was impractical to search these databases successfully with peptide masses alone. Fragment ion spectra were necessary for each peptide mass. Early demonstrations of protein identification from DNA databases, using mass spectral data, came from Yates [29] and Mann [30].

The adoption of protein identification also relied on advances in sample preparation. The early use of protein identification relied on digestion of gel-separated proteins that were stained by Coomassie blue, zinc, and other reversible stains, with digestion performed either directly in a gel piece or after transfer to a membrane. The demonstration of efficient digestion from silver stained gel pieces by Mann in 1996 [31] laid the groundwork for more sensitive protein detection and digestions. Simple methods for sample concentration and desalting, using reversed-phase resins in micropipette tips greatly improved sample processing [32–34]. Improved capillary LC made the separation of minute quantities of peptides possible.

Dramatic improvement in the performance of mass spectrometers was likewise an important factor in the increasing acceptance and use of protein identification. MALDI-TOF-MS was the primary instrumentation used for PMF from the beginning, and it continues to be to this day. The commercial time-of-flight instruments in use in 1992 were generally of limited resolution ($m/\Delta m$ <1000) and mass accuracy (>500 ppm). The development of high performance TOF instruments in the intervening years has been tremendous, with the combination of reflectors and delayed extraction now yielding resolution of >10,000 and mass accuracy <50 ppm. Automation now allows unattended measurement of 100 or more samples. Tandem MS combined with MALDI was possible in a limited way with post-source decay (PSD) on reflector TOF instruments [35] but the spectra were frequently of limited information content and were, in practice, difficult to measure. Genuine MALDI-MS/MS became possible with the quadrupole-TOF [36] and TOF-TOF configurations [37].

Improvements in electrospray have been no less impressive since the beginning of PMF. Low flow nano-electrospray [38] provided a means to generate ions for extended time from a microliter-sized sample. Micro-electrospray interface designs for on-line, low-flow HPLC separations [39–42] helped to extend the limits of detection for electrospray from picomole to femotomole to attomole levels. The commercial availability of new analyzer configurations such as the quadrupole ion trap [43] and quadrupole-TOF [44] provided researchers with improved sensitivity and resolving power. Just as importantly, instrument vendors developed user-friendly software for data dependent experiments.

Protein Identification Matures, 2003

The technology improvements mentioned above are part of an evolution that continues, apparently with no end in sight. Sequence databases are still growing at an exponential rate. Figure 11 shows the growth of GenBank [45] during the last twenty years. It is likely that the growth will continue, following the completion of the human genome, as many additional genomes are sequenced. There are similar sets of data available for the protein sequence databases, (e.g., the NCBI non-redundant protein sequence database, Swiss-Prot) and the growth in protein sequences is no less spectacular. The protein database at Genentech used in the initial PMF studies was 22,105 entries in 1989. It grew to over 91,000 in 1993 and was over 2.5 million in 2002.

The growth of the nascent field of proteomics has been nothing less than astounding. Figure 12 shows the number of papers published in proteomics since its conception. Furthermore, at least three new journals are devoted solely to topics in proteomics. More important than numbers is the growing impact of proteomics on biology. Numerous studies that utilize mass spectrometry-based approaches have produced valuable, new

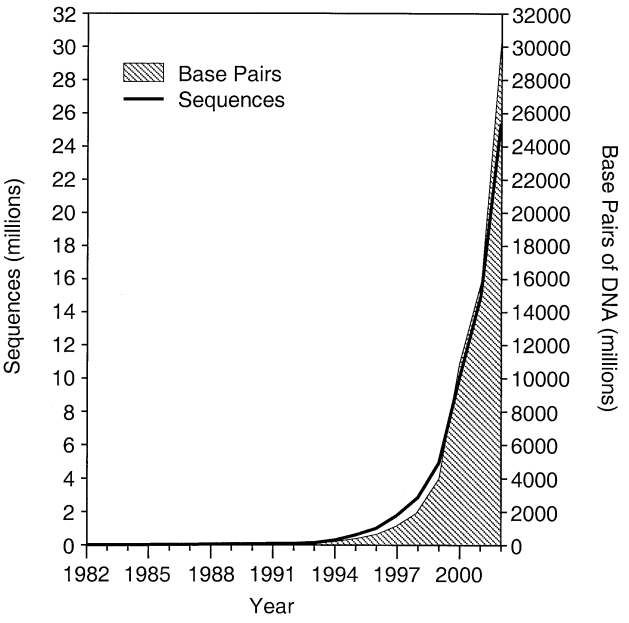


Figure 11. The growth of the GenBank DNA sequence database was particularly dramatic during the last five years (adapted from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

data through measurements of protein expression levels [46, 47], protein modifications [48, 49], multiprotein complexes [50–52] protein-protein interactions [53, 54], and subcellular localization [55, 56]. (We are able to cite here only a few, important applications of proteomics in biological research, from hundreds of published reports.) It is often the changes in these characteristics that provide insight into biological function, such as changes that may occur during receptor signaling, cell cycle progression, pathogen insult, and cancer progression. Proteomic studies that are functionally targeted have had the greatest impact on biological understanding [57].

What does the future hold? The major growth in protein identification will continue to be in the use of tandem mass spectral data for identification of proteins from one or more peptides. The concept of shotgun

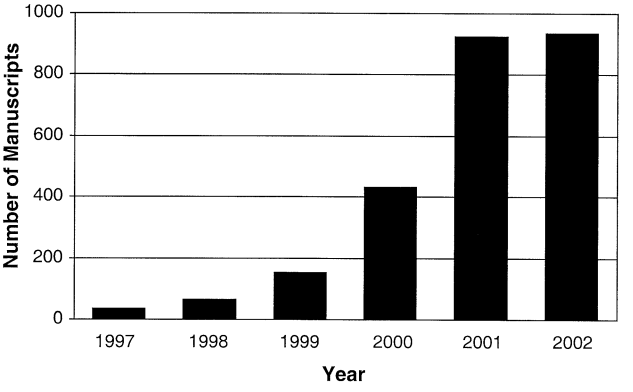


Figure 12. Proteomics papers published from 1997–2002, based on Medline entries for proteome and proteomics.

sequencing for proteolytic digests of complex mixtures [58] will continue to grow in importance.

Peptide mass fingerprinting itself relies on peptide masses derived from one or a few proteins. As such, increased use of PMF will depend on improvements in high resolution protein separations, most importantly multidimensional separations such as 2-D gels and 2-D protein chromatography [59]. These protein separation methods will continue to have advantages over shotgun approaches for some applications. Among the advantages are the ability to distinguish different forms of the same protein more readily (isoforms, alternativly spliced, proteolytically processed, post-translationally modified), as well as a larger dynamic range for quantitation.

Further refinements to PMF, including better matching algorithms and improved scoring schemes, will increase its utility. Of vital importance are approaches that give a statistical basis with which to evaluate the validity of protein identification. Early results often relied on manual inspection of the data and incorporation of additional data (e.g., knowledge of protein mass or pI, additional data from alternative protoeolytic digests). With the generation of data for very large numbers of proteins, manual verification has become inadequate and unfeasible. Early statistically based scoring schemes [60, 61] are being supplemented with newer, more global approaches [62, 63]. Yet a universally accepted scoring scheme with clear confidence values remains to be widely implemented. Furthermore, a move toward common representations of data [64] including mass spectral data formats, will be necessary for integration of proteomics results from many different experiments.

Peptide mass fingerprinting is one of the foundational technologies driving the growth of proteomics. We believe that PMF will continue to grow in importance for protein identification. However, the creation of new, innovative approaches to the design of proteomics experiments will be a key factor for future applications. With such improvements, mass spectrometry-based approaches for solving important problems in biology will continue to grow.

Acknowledgments

The authors gratefully acknowledge the generous support of Genentech, Inc., and the encouragement of many colleagues at Genentech. In particular, Dick Vandlen provided the authors with the freedom and resources to pursue this work. WH wishes to acknowledge the many members of his laboratory, in particular Susan Wong and Wendy Sandoval, and the encouragement and support of his late wife Bonnie. JS acknowledges those who helped stimulate his early interest—in analytical chemistry, T. R. Williams, and in mass spectrometry, Roland Gohlke, Chris Enke, Jack Holland, and Jack Watson. JS thanks his wife Kathy for her love and support, as well as collaboration over the years.

References

1. Matsudaira, P. Sequence from Picomole Quantities of Proteins Electroblotted onto Polyvinylidene Difluoride Membranes. *J. Biol. Chem.* **1987,** *262,* 10035–10038.
2. Orcutt, B. C.; George, D. G.; Dayhoff, M. O. Protein and Nucleic Acid Sequence Database Systems. *Annu. Rev. Biophys. Bioeng.* **1983,** *12,* 419–441.
3. Barber, M.; Bordoli, R. S.; Sedgewick, R. D.; Tyler, A. N. Fast Atom Bombardment of Solids (FAB). A New Ion Source for Mass spectrometry. *J. Chem Soc. Chem. Commun.* **1981,** *11,* 325–347.
4. Torgerson, D. F.; Skowronski, R. P.; Macfarlane, R. D. New Approach to the Mass Spectroscopy of Non-Volatile Compounds. *Biochem. Biosphys. Res. Commun.* **1974,** *60,* 616–621.
5. Henzel, W. J.; Stults, J. T.; Watanabe, C. *Proceedings of the Third Symposium of the Protein Society*; Seattle, WA, 1989.
6. Cleveland, D. W.; Fischer, S. G.; Kirschner, M. W.; Laemmli, U. K. Peptide Mapping by Limited Proteolysis in Sodium Dodecyl Sulfate and Analysis by Gel Electrophoresis. *J. Biol. Chem.* **1977,** *252,* 1102–1106.
7. Henzel, W. J.; Stults, J. T. ; Wong, S. C.; Namenuk, A.; Yashio, J.; Watanabe, C. In *Techniques in Protein Chemistry*; Marshak, D. R., Ed.; Academic Press: San Diego, 1995; Vol. VII, pp 341–346.
8. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecule. *Science* **1989,** *246,* 64–67.
9. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60,* 2299–2301.
10. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. Protein and Polymer Analyses up to *m/z* 100,000 by Laser Ionization Time-of-flight Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **1988,** *2,* 151–153.
11. Billeci, T. M.; Stults, J. T. Tryptic Mapping of Recombinant Proteins by Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Anal. Chem.* **1993,** *65,* 1709–1716.
12. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases. *Proceedings of the National Academy of Sciences*; **1993,** *90,* 5011–5015.
13. Aebersold, R.; Leavitt, J.; Saavedra, R. A.; Hood, L. E.; Kent, S. B. Internal Amino Acid Sequence Analysis of Proteins Separated by One- or Two-Dimensional Gel Electrophoresis After in Situ Protease Digestion on Nitrocellulose. *Proceedings of the National Academy of Sciences*; **1987,** *84,* 6970–6974.
14. Wong, S. C.; Grimley, C.; Padua, A.; Bourell, J. H.; Henzel, W. J. In *Techniques in Protein Chemistry IV*; Angeletti, R. H., Ed.; Academic Press: San Diego, 1993, pp 371–378.
15. Mann, M.; Hojrup, P.; Roepstorff, P. Use of Mass Spectrometric Molecular Weight Information to Identify Proteins in Sequence Databases. *Biol. Mass Spectrom.* **1993,** *22,* 338–345.
16. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. Rapid Identification of Proteins by Peptide Mass Fingerprinting. *Current Biol.* **1993,** *3,* 327–332.
17. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein identification by Mass Profile Fingerprinting. *Biochem. Biophys. Res. Commun.* **1993,** *195,* 58–64.
18. Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide Mass Maps: A Highly Informative Approach to Protein Identification. *Anal. Biochem.* **1993,** *214,* 397–408.
19. Henzel, W. J.; Grimley, C.; Bourell, J. H.; Billeci, T. M.; Wong, S. C.; Stults, J. T. Analysis of Two-Dimensional Gel Proteins by Mass Spectrometry and Microsequencing. *Methods Enzymol.* **1994,** *6,* 239–247.

20. Mann, M.; Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994,** *66,* 4390–4399.
21. Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
22. Wolters, D. A.; Washburn, M. P.; Yates, J. R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **2001,** *73,* 5683–5690.
23. Jungblut, P.; Thiede, B.; Zimny-Arndt, U.; Müller, E. C.; Scheler, C.; Wittmann-Liebold, B.; Otto, A. Resolution Power of Two-Dimensional Electrophoresis and Identification of Proteins from Gels. *Electrophoresis* **1996,** *17,* 839–847.
24. Taylor, J.; Anderson, N. L.; Scandora, A. E., Jr.; Willard, K. E.; Anderson, N. G. Design and Implementation of a Prototype Human Protein Index. *Clin. Chem.* **1982,** *28,* 861–866.
25. Wilkins, M. R.; Sanchez, J.-C.; Gooley, A. A.; Appel, R. D.; Humphrey-Smith, I.; Hochstrasser, D. F.; Williams, K. L. Progress with Proteome Projects: Why All Proteins Expressed by a Genome Should be Identified and How to Do It. *Biotech. Gen. Eng. Rev.* **1995,** *13,* 19–50.
26. Arnott, D.; O'Connell, K. L.; King, K. L.; Stults, J. T. An Integrated Approach to Proteome Analysis: Identification of Proteins Associated with Cardiac Hypertrophy. *Anal. Biochem.* **1998,** *258,* 1–18.
27. 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. G. ; Kelley, J. M.; Fritchman, J. L.; Weidman, J. F.; Small, K. V.; Sandusky, M.; Fuhrmann, J. L.; Nguyen, D. T.; Utterback, T.; Saudek, D. M.; Philips, C. A.; Merrick, J. M.; Tomb, J.; Dougherty, B. A.; Bott, K. F.; Hu, P. C.; Lucier, T. S.; Paterson, S. N.; Smith, H. O.; Venter, J. C. The Minimal Gene Complement of Mycoplasma genitalium. *Science* **1995,** *270,* 397–404.
28. Blattner, F. R.; Plunkett III, G.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **1997,** *277,* 1453–1462.
29. Yates, J. R., III; Eng, J. K.; McCormack, A. L. Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. *Anal. Chem.* **1995,** *67,* 3202–3210.
30. Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Boucherie, H.; Mann, M. Linking Genome and Proteome by Mass Spectrometry: Large-Scale Identification of Yeast Proteins from Two-Dimensional Gels. *Proc. Natl. Acad. Sci. U.S.A.* **1996,** *93,* 14440–14445.
31. Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass Spectrometric Sequencing of Proteins from Silver Stained Polyacrylamide Gels. *Anal. Chem.* **1996,** *68,* 850–858.
32. Otto, A.; Thiede, B.; Müller, E. C.; Scheler, C.; Wittmann-Liebold, B.; Jungblut, P. Identification of Human Myocardial Proteins Separated by Two-Dimensional Electrophoresis Using an Effective Sample Preparation for Mass Spectrometry. *Electrophoresis* **1996,** *17,* 1643–1650.
33. Erdjument-Bromage, H.; Lui, M.; Lacomis, L.; Grewal, A.; Annan, R.; McNulty, D.; Carr, S.; Tempst, P. Examination of Micro-Tip Reversed-Phase Liquid Chromatographic Extraction of Peptide Pools for Mass Spectrometry Analysis. *J. Chromatogr. A* **1998,** *826,* 167–181.
34. Gobom, J.; Nordhoff, E.; Mirgorodskaya, E.; Ekman, R.; Roepstorff, P. Sample Purification and Preparation Technique Based on Nano-Scale Reversed-Phase Columns for the Sensitive Analysis of Complex Peptide Mixtures by Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *J. Mass Spectrom.* **1999,** *34,* 105–116.
35. Kaufmann, R.; Spengler, B.; Luetzenkirchen, F. Mass Spectrometric Sequencing of Linear Peptides by Product-Ion Analysis in a Reflectron Time-of-Flight Mass Spectrometer Using Matrix-Assisted Laser Desorption Ionization. *Rapid Commun. Mass Spectrom.* **1993,** *7,* 902–910.
36. Shevchenko, A.; Loboda, A.; Ens, W.; Standing, K. G. MALDI Quadrupole Time-of-Flight Mass Spectrometry: A Powerful Tool for Proteomic Research. *Anal. Chem.* **2000,** *72,* 2132–2141.
37. Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. The Characteristics of Peptide Collision-Induced Dissociation Using a High-Performance MALDI-TOF/TOF Tandem Mass Spectrometer. *Anal. Chem.* **2000,** *72,* 552–558.
38. Wilm, M.; Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Anal. Chem.* **1996,** *68,* 1–8.
39. Wahl, J. H.; Gale, D. C.; Smith, R. D. Sheathless Capillary Electrophoresis-Electrospray Ionization Mass Spectrometry Using 10 m i.d. Capillaries: Analyses of Tryptic Digests of Cytochrome c. *J. Chromatogr. A* **1994,** *659,* 217–222.
40. Davis, M.; Lee, T. Rapid Protein Identification Using a Microscale Electrospray LC/MS System on an Ion Trap Mass Spectrometer. *J. Am. Soc. Mass Spectrom.* **1998,** *9,* 194–201.
41. Martin, S. E.; Shabanowitz, J.; Hunt, D. F.; Marto, J. A. Subfemtomole MS and MS/MS Peptide Sequence Analysis Using Nano-HPLC Micro-ESI Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2000,** *72,* 4266–4274.
42. Gatlin, C.; Kleemann, G.; Hays, L.; Link, A.; Yates, J. Protein Identification at the Low Femtomole Level from Silver-Stained Gels Using a New Fritless Electrospray Interface for Liquid Chromatrography-Microspray and Nanospray Mass Spectrometry. *Anal. Biochem.* **1998,** *263,* 93–101.
43. Stafford, G. C.; Kelley, P. E.; Syka, J. E. P.; Reynolds, W. E.; Todd, J. F. J. Recent Improvements in Analytical Applications of Ion Trap Technology. *Int. J. Mass Spectrom. Ion Processes* **1984,** *60,* 85–98.
44. Morris, H. R.; Paxton, T.; Dell, A.; Langhorne, J.; Berg, M.; Bordoli, R. S.; Hoyes, J.; Bateman, R. H. High Sensitivity Collisionally-Activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-Acceleration Time-of-flight Mass Spectrometer. *Rapid Commun. Mass Spectrom.* **1996,** *10,* 889–896.
45. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J. A.; Rapp, B. A.; Wheeler, D. L. GenBank. *Nucleic Acids Res.* **2002,** *30,* 17–20.
46. Ideker, T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlett, D. R.; Aebersold, R.; Hood, L. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* **2001,** *292,* 929–934.
47. Celis, J. E.; Palsdottir, H.; Ostergaard, M.; Gromov, P.; Primdahl, H.; Orntoft, T. F.; Wolf, H.; Celis, A.; Gromova, I. Proteomic Strategies to Reveal Tumor Heterogeneity among Urothelial Papillomas. *Mol. Cell Proteomics* **2002,** *1,* 269–279.
48. Steen, H, K. B.; Fernandez, M.; Pandey, A.; Mann, M. Tyrosine Phosphorylation Mapping of the Epidermal Growth Factor Receptor Signaling Pathway. *J. Biol. Chem.* **2002,** *277,* 1031–1039.
49. Chen, L. S.; Shou, W.; Deshaies, R. J.; Annan, R. S.; Carr, S. A. Mass Spectrometry-Based Methods for Phosphorylation Site Mapping of Hyperphosphorylated Proteins Applied to Net1, a Regulator of Exit from Mitosis in Yeast. *Mol. Cell Proteomics* **2002**, *1,* 204–212.
50. Rout, M. P.; Aitchison, J. D.; Suprapto, A.; Hjertaas, K.; Zhao, Y.; Chait, B. T. The Yeast Nuclear Pore Complex: Composition,

Architecture, and Transport Mechanism. *J. Cell Biol.* **2000,** *148,* 635–652.
51. Allen, N. P.; Patel, S. S.; Huang, L.; Chalkley, R. J.; Burlingame, A.; Lutzmann, M.; Hurt, E. C.; Rexach, M. Deciphering Networks of Protein Interactions at the Nuclear Pore Complex. *Mol. Cell Proteomics* **2002,** *1,* 930–946.
52. Deshaies, R. J.; Seol, J. H.; McDonald, W. H.; Cope, G.; Lyapina, S.; Shevchenko, A.; Verma, R.; Yates, J. R., III. Charting the Protein Complexome in Yeast by Mass Spectrometry. *Mol. Cell Proteomics* **2002,** *1,* 3–10.
53. Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L. Y.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Figeys, D.; Tyers, M. Systematic Identification of Protein Complexes in *Saccharomyces cerevisiae* by Mass Spectrometry. *Nature* **2002,** *415,* 180–183.
54. Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature* **2002,** *415,* 141–147.
55. Garin, J. D. R.; Kieffer, S.; Dermine, J. F.; Duclos, S.; Gagnon, E.; Sadoul, R.; Rondeau, C.; Desjardins, M. The Phagosome Proteome: Insight into Phagosome Functions. *J. Cell Biol.* **2001,** *152,* 165–180.
56. Andersen, J. S.; Lyon, C. E.; Fox, A. H.; Leung, A. K.; Lam, Y. W.; Steen, H.; Mann, M.; Lamond, A. I. Directed Proteomic Analysis of the Human Nucleolus. *Curr. Biol.* **2002,** *12,* 1–11.
57. Huber, L. A. Is Proteomics Heading in the Wrong Direction? *Nature Rev. Mol. Cell Biol.* **2003,** 74–80.
58. MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. Shotgun Identification of Protein Modifications from Protein Complexes and Lens Tissue. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99,* 7900–7905.
59. Lubman, D. A.; Kachman, M. T.; Wang, H. X.; Gong, S. Y.; Yan, F.; Hamler, R. L.; O'Neil, K. A.; Zhu, K.; Buchanan, N. S.; Barder, T. J. Two-Dimensional Liquid Separations-Mass Mapping of Proteins from Human Cancer Cell Lysates. *J. Chromatogr. B* **2002,** *782,* 183–196.
60. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999,** *20,* 3551–3567.
61. Zhang, W. Z.; Chait, B. T. Profound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information. *Anal. Chem.* **2000,** *72,* 2482–2489.
62. Eriksson, J.; Fenyö, D. A Model of Random Mass-Matching and Its Use for Automated Significance Testing in Mass Spectrometric Proteome Analysis. *Proteomics* **2002,** *2,* 1615–9861.
63. Fenyö, D.; Beavis, R. C. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Anal. Chem.* **2003,** *75,* 768–774.
64. Taylor, C. F.; Paton, N. W.; Garwood, K. L.; Kirb, P. D.; Stead, D. A.; Yin, Z.; Deutsch, E. W.; Selway, L.; Walker, J.; Riba-Garcia, I.; Mohammed, S.; Deery, M. J.; Howard, J. A.; Dunkley, T.; Aebersold, R.; Kell, D. B.; Lilley, K. S.; Roepstorff, P.; Yates, J. R., III; Brass, A.; Brown, A. J. P.; Cash, P.; Gaskell, S. J.; Hubbard, S. J.; Oliver, S. G. A Systematic Approach to Modeling, Capturing, and Disseminating Proteomics Experimental Data. *Nat. Biotech.* **2003,** *21,* 247–254.