Amino Acid Sequencing of Proteins

KLAUS BIEMANN* AND IOANNIS A. PAPAYANNOPOULOS†

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received July 27, 1994

Introduction

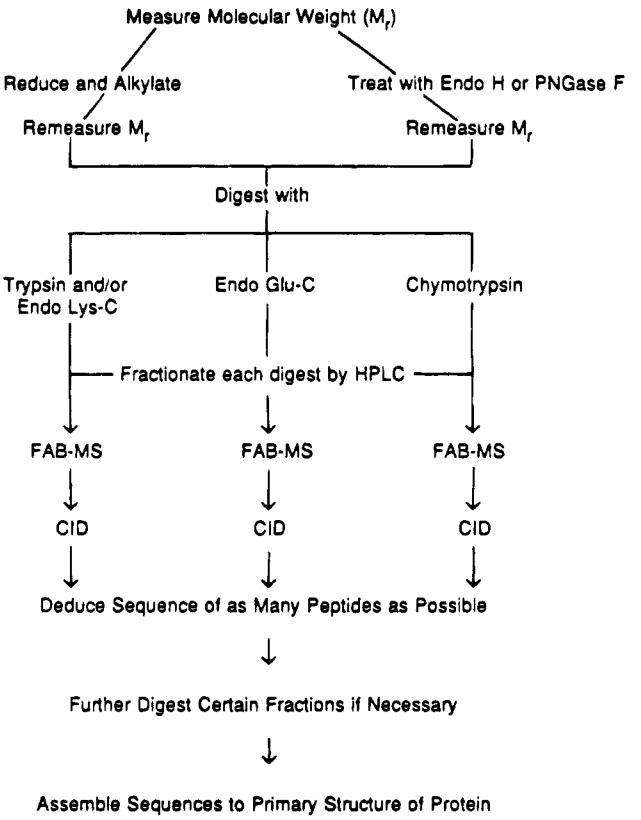
During the 1980s, various developments occurred in the field of mass spectrometry (MS) that caused an almost explosive expansion of its applicability to certain areas of biochemistry and biology. Fast atom (or ion) bombardment (FAB or LSIMS)[1] first made it possible to ionize relatively large, polar molecules ranging to at least 10 kDa in molecular mass. This range was greatly extended to the 10^5 Da level and beyond by the development of matrix-assisted laser desorption ionization (MALDI)[2] and electrospray ionization (ESI).[3] All these methods produce abundant $(M + H)^+$ and/or $(M + nH)^{n+}$ ions but cause little fragmentation, and tandem mass spectrometry[4] is the approach of choice for obtaining detailed structural information. All three ionization techniques were popularized using peptides or proteins as examples, and most of the applications published since have dealt with various aspects of this field.[5–8] Other areas, such as carbohydrates, glycoconjugates, nucleotides, etc., soon followed.[7,8]

Since the mid-1980s, we have developed a strategy for the determination of the primary structure of proteins by FAB and tandem mass spectrometry using collision-induced dissociation (CID) at relatively high collision energies (2–10 keV). The $(M + H)^+$ ions produced by FAB in the first mass spectrometer (MS-1) of the tandem instrument do not have sufficient excess energy to fragment spontaneously, but the small fraction of kinetic energy converted into vibrational energy upon collision with a helium or xenon atom suffices to cause fragmentation of the peptide and side chain bonds. The resulting structure-specific fragment ions are then mass analyzed in the second mass spectrometer (MS-2). The sequencing methodology was improved and applied first to the thioredoxins isolated from the photosynthetic green sulfur bacterium *Chlorobium thiosulfatophilum*[9] and the purple sulfur bacterium from *Chromatium vinosum*.[10] The sequencing of other members of this protein family and of the somewhat related glutaredoxin class followed.

Klaus Biemann was born on November 2, 1926, in Innsbruck, Austria. He obtained his Ph.D. degree in organic chemistry from the University of Innsbruck in 1951. He moved to the Massachusetts Institute of Technology in 1955 as a postdoctoral research associate with George Büchi, joined the faculty in 1957, and has been a professor of chemistry there since 1963. Since 1958, he has carried out research devoted to the development and application of mass spectrometry for the determination of the structure of natural products, beginning with alkaloids, amino acids, and peptides. Presently, his major interest is in the use of this methodology for the determination of the primary structure of peptides and proteins, as well as related areas.

Ioannis A. Papayannopoulos was born in Greece in 1958. He came to the United States in 1977 and received his A.B. from Bowdoin College and his Ph.D. in organic chemistry under Professor Klaus Biemann from the Massachusetts Institute of Technology. He remained at MIT for several years after graduation, first as a postdoctoral associate and, subsequently, as a research scientist and assistant director of the NIH-supported Mass Spectrometry Facility. Dr. Papayannopoulos is currently a senior scientist with Biogen, a biotechnology company in Cambridge, MA. His research interests focus on applications of mass spectrometric methodologies to peptide and protein structure problems.

Scheme 1.[a] General Strategy for Protein Sequencing



[a] Reproduced from ref 8 with permission of John Wiley and Sons Ltd. Copyright 1994.

For mass spectrometric amino acid sequencing, the protein is first treated with a reducing agent to convert any –S–S– bonds to –SH groups, which are then alkylated to stabilize the reduced form and to, at least partially, unfold the protein to facilitate proteolysis with enzymes of relatively high specificity (Scheme 1). The mass difference between the native protein and the reduced and alkylated protein (determined by MALDI or ESI) establishes the number of cysteines present. The individual enzyme digests, which are generally rather complex mixtures, are then fractionated by reversed-phase high-performance liquid chro-

* Address correspondence to Prof. Klaus Biemann, Dept. of Chemistry, Rm. 56-010, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307.

† Present address: Biogen Inc., 14 Cambridge Ctr., Cambridge, MA 02142.

(1) Barber, M.; Bordoli, R. A.; Sedgwick, R. D.; Tyler, A. N. *J. Chem. Soc., Chem. Commun.* **1981**, 325–327.

(2) Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. *Int. J. Mass Spectrom. Ion Processes* **1987**, *78*, 53–68.

(3) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64–71.

(4) *Tandem Mass Spectrometry*; McLafferty, F. W., Ed.; John Wiley & Sons: New York, 1983.

(5) Biemann, K.; Martin, S. A. *Mass Spectrom. Rev.* **1987**, *6*, 1–76.

0001-4842/94/0127-0370$04.50/0

matography (rpHPLC) to generate simpler peptide mixtures (one to five components). The amino acid sequence of the individual peptides is then determined by CID-MS. To establish a previously unknown protein sequence, one must use more than one enzyme of different specificity (for example, trypsin that cleaves only at arginine and lysine, and endoGlu-C that cleaves only at glutamic acid at pH 4 and also at aspartic acid at pH 8) to obtain overlapping sets of peptides and sufficient redundancy in the data. Glycoproteins are first treated with enzymes that remove the carbohydrate portions. A more detailed description of the experimental procedure will be given later during the discussion on the sequencing of the thioredoxin isolated from the photosynthetic bacterium *Chloroflexus aurianticus*.

Thioredoxins

Thioredoxins are a class of ubiquitous redox proteins that occur in all living systems, from single cell organisms to humans. They are generally 100–110 amino acids long and have a highly conserved active site encompassing -Trp-Cys-Gly-Pro-Cys- or -Trp-Cys-Ala-Pro-Cys-. The reversible interchange of the two cysteines in the reduced (free thiols) and oxidized (disulfide bridges) forms is the basis of the diverse biological functions of these proteins.[11] The amino acid sequences retain a certain degree of homology from bacteria to mammals, and this can be used to establish their evolutionary history.[12]

As the most recent, and in some ways the most interesting, example illustrating the power of the mass spectrometric approach to peptide and protein structure, unpublished work from our laboratory on the sequencing of the thioredoxin from *Chloroflexus* will be discussed. This photosynthetic bacterium is also a very interesting organism from the standpoint of evolution, and its thioredoxin shows some different properties from those of others.

The molecular weight of the native protein was determined first by MALDI using a Vestec 2000 time-of-flight (TOF) mass spectrometer, and then by ESI[13] on a double-focusing mass spectrometer, in MS-1 of the two spectrometers comprising the JEOL HX110/HX110 tandem instrument. The MALDI value for the $(M + H)^+$ ion was *m/z* 12 012, i.e., $M_r = 12\,011$. ESI on the high-resolution instrument (Figure 1a) revealed a multiplicity of molecule ions, as shown (Figure 1b) for the $(M + 8H)^{8+}$ ion, indicating the presence of at least three species of M_r 11 996, 12 010, and 12 025. The protein was reduced with triethylphosphine and alkylated with 4-vinylpyridine. It should be noted that, because of the relatively low resolution obtained

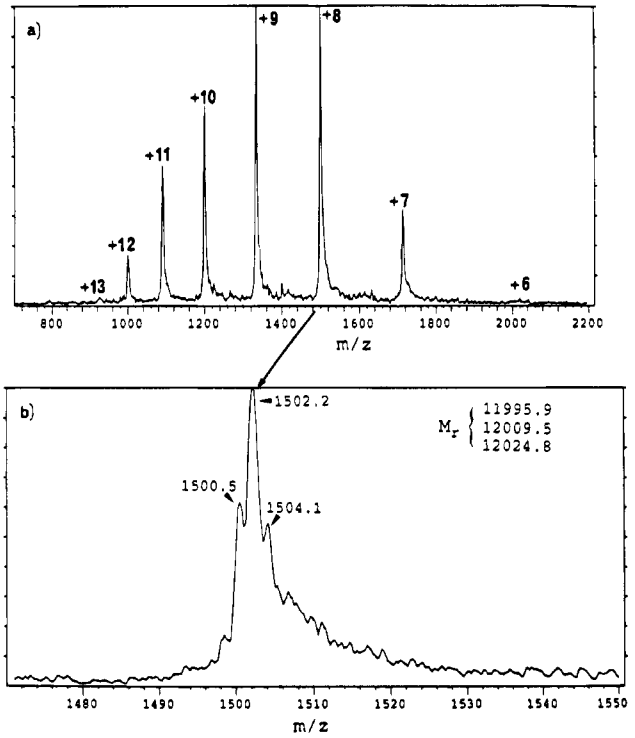


Figure 1. (a) ESI mass spectrum of *Chloroflexus* thioredoxin. (b) Region of $(M + 8H)^{8+}$ expanded.

in both the MALDI process and ESI, the molecular weight measurements are isotopically averaged values (i.e., C = 12.011, etc.), while the data obtained on peptides by FAB and CID on a magnetic sector mass spectrometer are based on monoisotopic masses (i.e., $^{12}C = 12.000\,000$).

Separate aliquots corresponding to 4–5 nmol each of the reduced and alkylated protein were digested with trypsin, chymotrypsin, and endoGlu-C, followed by endoLys-C, respectively. The conditions were essentially those used in the structure determination of the human erythrocyte glutaredoxin[14] discussed later. Each digest was separated by rpHPLC into a number of fractions. Fraction 7 of the chymotryptic digest (Figure 2) gave a FAB mass spectrum that revealed the presence of at least seven components, three of which clustered between *m/z* 1000 and 1050 with the monoisotropic (^{12}C-only) $(M + H)^+$ ions differing by consecutive 14 units (Figure 3). One major advantage of tandem mass spectrometry over the conventional Edman degradation is the fact that it is not necessary to purify each peptide prior to analysis, which would require considerable effort and more material. One can achieve the analogous separation in MS-1, by selecting one $(M + H)^+$ ion after the other (as long as they differ by at least 1 mass unit) for CID and recording the unique product ion spectrum (with MS-2) of each component independent of the presence of others.[15]

The CID spectra obtained from the precursor ions of *m/z* 1019.5, 1033.5, and 1047.5 are shown in Figure 4a–c. The labeling of the peaks corresponds to the

(6) Biemann, K. *Annu. Rev. Biochem.* **1992**, *61*, 977–1010.
(7) *Methods in Enzymology. Mass Spectrometry, Vol. 193*; McCloskey, J. A., Ed.; Academic Press: San Diego, 1990.
(8) Biemann, K. In *Biological Mass Spectrometry: Present and Future*; Matsuo, T., Caprioli, R. M., Gross, M. L., Seyama, Y., Eds.; Wiley: Sussex, 1994; pp 275–297.
(9) Mathews, W. R.; Johnson, R. S.; Cornwell, K. L.; Johnson, T. C.; Buchanan, B. B.; Biemann, K. *J. Biol. Chem.* **1987**, *262*, 7537–7545.
(10) Johnson, R. S.; Biemann, K. *Biochemistry* **1987**, *26*, 1209–1214.
(11) Holmgren, A. *Annu. Rev. Biochem.* **1985**, *54*, 237–271.
(12) Hartman, H.; Syvanen, M.; Buchanan, B. B. *Mol. Biol. Evol.* **1990**, *7*, 247–254.
(13) Cody, R. B.; Tamura, J.; Musselman, B. D. *Anal. Chem.* **1992**, *64*, 1561–1570.
(14) Papov, V. V.; Gravina, S. A.; Mieyal, J. J.; Biemann, K. *Protein Sci.* **1994**, *3*, 428–434.
(15) Biemann, K. *Anal. Chem.* **1986**, *58*, 1289A–1300A.

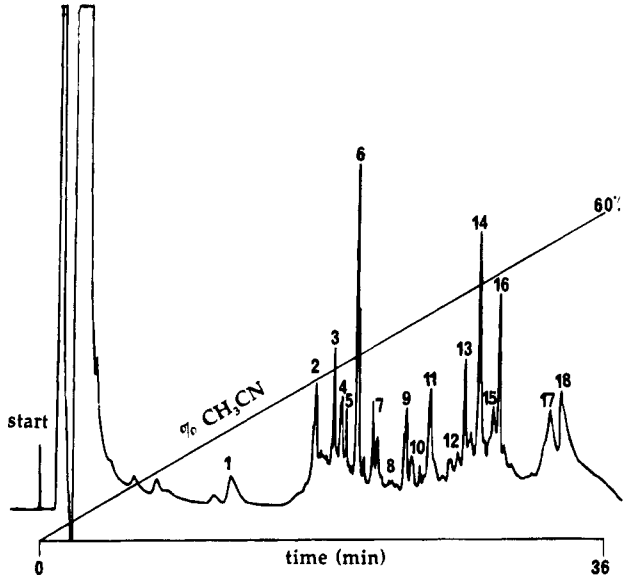


Figure 2. HPLC chromatogram of the chymotryptic digest of *Chloroflexus* thioredoxin.

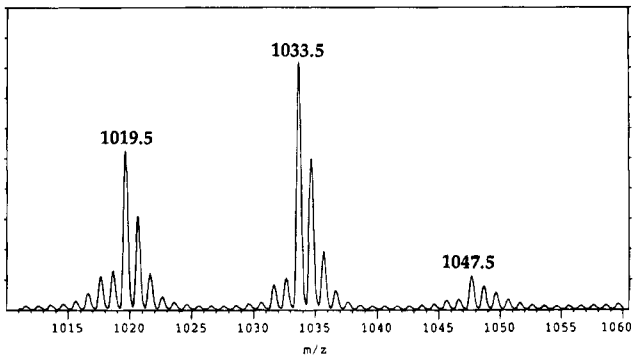


Figure 3. Expanded region of the FAB mass spectrum of fraction 7 in Figure 2.

notations summarized in Scheme 2.[16] It is obvious that these three spectra exhibit a very similar pattern, except that all peaks above a_5 and d_6 are displaced by 14 and 28 units in Figure 4b and Figure 4c, respectively, with respect to Figure 4a.

These CID spectra are relatively simple. Because of the presence of the strongly basic arginine at the N-terminus, the protonating hydrogen of the $(M + H)^+$ precursor ion preferentially localizes the positive charge there[17] and charge-remote fragmentation[18] leads to the preponderance of the N-terminal a_n and d_n ions. The latter permit the differentiation of leucine from isoleucine[17] and are only formed under the high energy (>1 keV) collision conditions achievable in magnetic sector instruments, in contrast to triple quadrupole tandem mass spectrometers. The spectra shown in Figure 4a,b were first interpreted to indicate the isobaric sequences Arg-Glu-Ile-Phe-Asp-Ala-Gly-Val-Leu and that in which Gly(7) is replaced by Ala. This is because the peaks labeled d_6 and a_6 in Figure 4a would be of the same mass as the a_6 and a_7 ions for the -Ala(6)-Gly(7)- sequence and the mass shift in the spectrum shown in Figure 4b would similarly

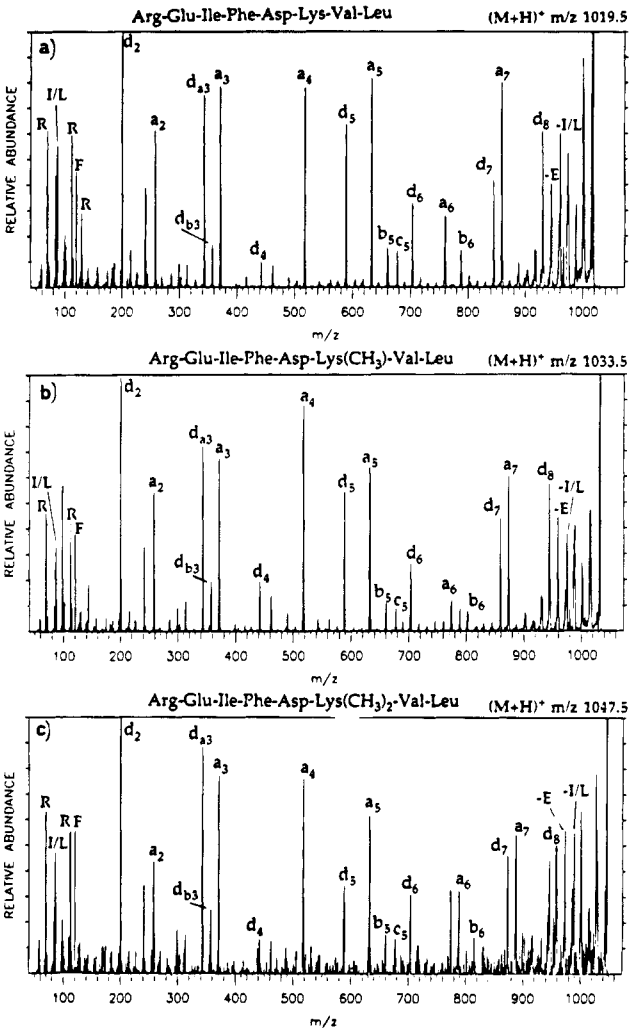
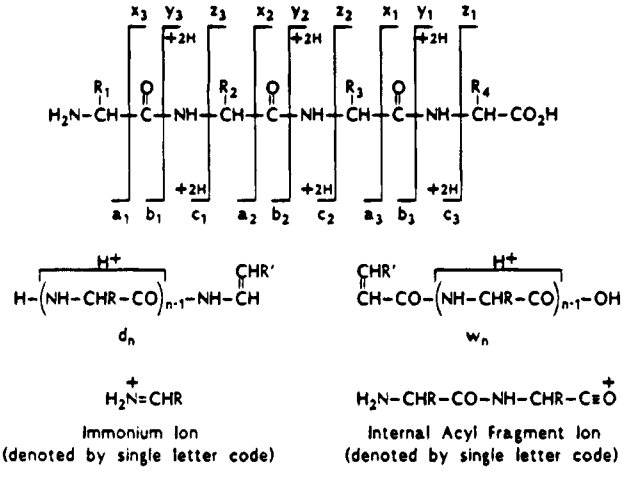


Figure 4. CID spectra of the ^{12}C species of the three components shown in Figure 3.

Scheme 2.[a] Fragment Ions Produced by CID from Protonated Peptides



[a] Reproduced from ref 6 with permission. Copyright 1992 Annual Reviews Inc.

support the -Ala(6)-Ala(7)- sequence. It was only when the CID spectrum of the minor component of $(M + H)^+ = m/z$ 1047.5 (Figure 4c) was measured that this interpretation became doubtful because it would require α-aminobutyric acid (Abu), a non-protein amino acid, in position 7. At that point, one had two choices: either to synthesize the peptides with Ala-

(16) Biemann, K. *Biomed. Environ. Mass Spectrom.* **1988**, *16*, 99–111.
(17) Johnson, R. S.; Martin, S. A.; Biemann, K. *Int. J. Mass Spectrom. Ion Processes* **1988**, *86*, 137–154.
(18) Tomer, K. B.; Crow, F. W.; Gross, M. L. *J. Am. Chem. Soc.* **1983**, *105*, 5487–5488.

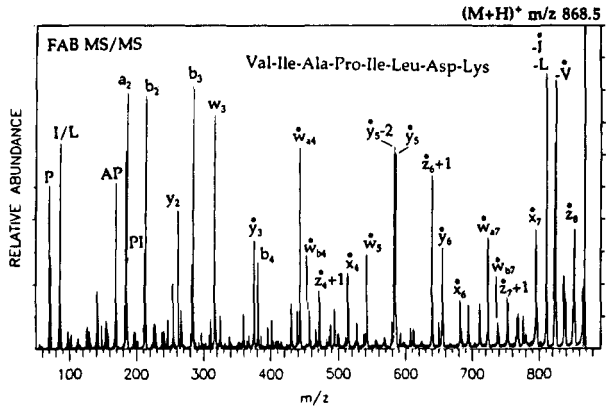


Figure 5. CID spectrum of a *Chloroflexus* thioredoxin tryptic peptide of $(M + H)^+ = m/z$ 868. Fragment ions with a dot above their label shift by 1 unit in the CID spectrum of the m/z 869 precursor, which turns out to be a mixture of the ^{13}C isotope of the above peptide and the ^{12}C isotope of a second peptide 1 unit heavier. This is a result of Leu-6 substitution by Asn (see also Figure 6).

Gly, Ala-Ala, and Ala-Abu in positions 6 and 7 and determine their CID spectra for comparison or to carry out an Edman degradation on this HPLC fraction, which, as mentioned earlier, contained at least seven peptides. The latter approach was chosen and revealed the absence of Ala in position 6 but indicated Lys-PTH instead. Thus, the sequence shown in Figure 4a-c must be correct and the mass differences of 14 and 28 units reside in the side chain of lysine (i.e., its N_ϵ-monomethyl and N_ϵ,N_ϵ-dimethyl homologues). In principle, it would also be possible to differentiate between the Ala-Gly and Lys sequences (and their homologues) by carrying out an exact mass measurement on the $(M + H)^+$ ion, which differs by 0.036 Da for the two possibilities.

While the CID spectra shown in Figure 4 are relatively simple, that of another peptide $(M + H)^+ = m/z$ 868.5 in a fraction of the tryptic digest is more complex (Figure 5). It consists of a number of different ion types, mainly those retaining the charge at the C-terminus due to the C-terminal lysine, while the N-terminal a_n and b_n ions include only the first four amino acids. The frequent w_n ions permit the identification of the two isoleucines and the single leucine. The w_3 ion corresponds to the loss of a C_3H_7 group, indicating that leucine is in position 6 (third from the C-terminus), while the pairs w_{a4}, w_{b4} and w_{a7}, w_{b7} are due to loss of C_2H_5 and CH_3, respectively, in accordance with the mechanism of this side chain cleavage.[19]

When the precursor ion at m/z 868 was examined, it was observed that the isotope pattern was different from that expected for a peptide of this composition because the $^{13}C_1$ component was too abundant (Figure 6). This suggested the presence of another peptide of $(M + H)^+ = m/z$ 869. When this ion was subjected to CID, a spectrum very similar to Figure 5 was obtained, except that all peaks became doublets because the precursor ion now included the $^{13}C_1$ species of the m/z 868 ion. However, for those peaks marked with a dot, the 1 mass unit higher ion of the doublet was now more abundant because it was due to the ^{12}C species of the m/z 869 component. It follows that in this

(19) Johnson, R. S.; Martin, S. A.; Biemann, K.; Stults, J. T.; Watson, J. T. *Anal. Chem.* **1987**, *59*, 2621–2625.

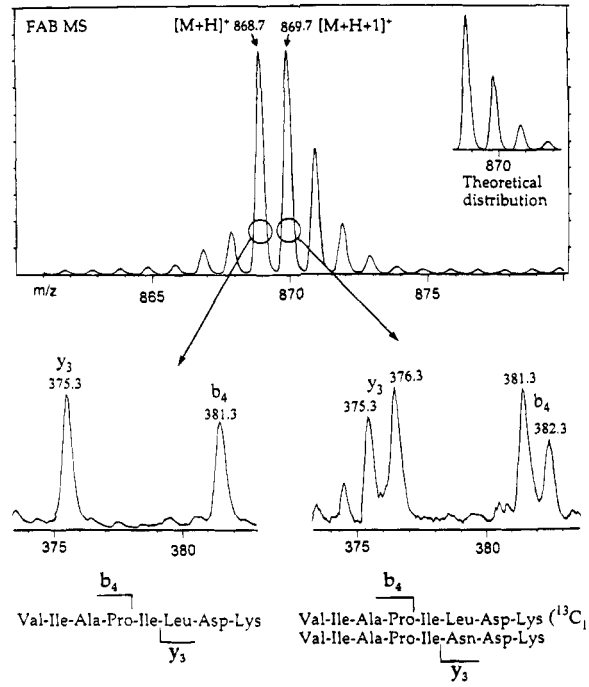


Figure 6. FAB mass spectrum (top) of the $(M + H)^+$ region of the tryptic peptide, the CID spectrum of which is shown in Figure 5. As is apparent from the isotope pattern when compared to the theoretical one (insert), two peptides differing by 1 unit are actually present. Comparison of the CID spectrum of m/z 868 (^{12}C isotope of lighter peptide) with that of m/z 869 (^{12}C isotope of heavier peptide *and* ^{13}C isotope of lighter peptide) reveals the position in the peptide sequence where amino acid substitution has occurred: The b_4 fragment ion (bottom) remains at m/z 381.3 in both spectra, indicating that the first four amino acids of the peptide remain unchanged. However, the y_3 fragment ion shifts by 1 unit, from m/z 375.3 to m/z 376.3, indicating that the substitution has occurred in the last three amino acids. Other fragment ions (Figure 5) help pinpoint the substitution at position 6.

peptide the leucine in position 6 (41 in the final sequence) is replaced by asparagine. The regions of the y_3 and b_4 ions are expanded in the lower portion of Figure 6 as an example. The complete sequence of the *Chloroflexus* thioredoxin is shown in Figure 7 including the supporting data. It will be noted that positions 58 and 59 are also occupied by two amino acids, Asn vs Asp and Thr vs Val. While the former could be due to deamidation during the isolation and purification of the protein, the latter, as well as the Leu vs Asn duplicity at position 41 discussed above, must be due to the fact that this organism has multiple genes coded for its thioredoxin. On the other hand, the methylation of Lys-104 must be a posttranslational event.

Now, as the sequence determination of the *Chloroflexus* thioredoxin has been completed, one can calculate the expected molecular weights for the three components that have been revealed by the ESI spectrum shown in Figure 1. For the lightest component, which would contain Leu at position 41, Asn at 58, Val at 59, and Lys at 104, $M_r = 11\,996$; for MeLys and Me_2Lys at 104, $M_r = 12\,010$ and 12 024, respectively, in excellent agreement with the experimental values (Figure 1b). The mass resolution of ESI is, however, not sufficient to determine that the lightest combination is one component, and that the other is represented by 41 = Asn, 58 = Asp, and 59 = Thr. The combined mass difference of these two combina-

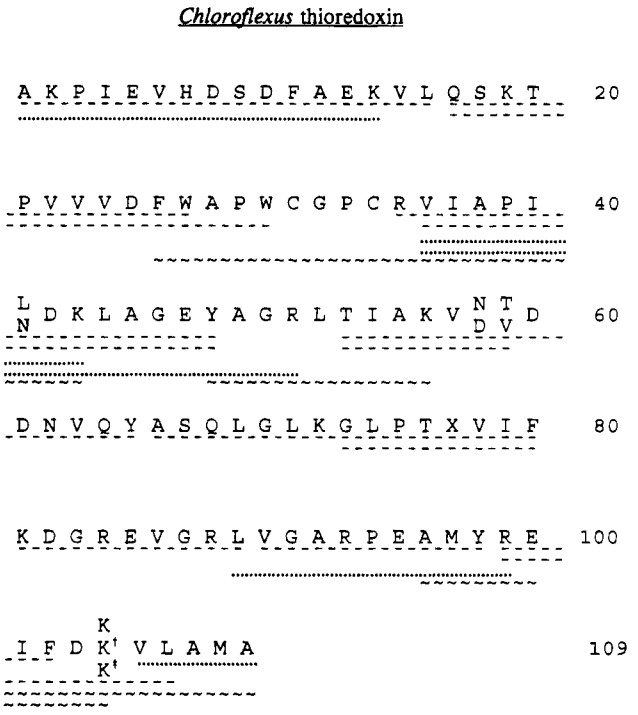


Figure 7. Sequence of *Chloroflexus* thioredoxin. All peptides sequenced by CID tandem mass spectrometry are underlined as follows: chymotryptic peptides with dashed lines (– – –), tryptic peptides with dotted lines (···), and peptides obtained from a combined endoGlu-C, endoLys-C digest with wavy lines (~ ~ ~). The notations K† and K‡ stand for N_ϵ-methyllysine and N_ϵ,N_ϵ-dimethyllysine, respectively.

tions is only +4 Da, which could not be resolved at *m/z* 12 000 under the conditions used.

Figure 8 represents a summary of the primary structures of the five thioredoxins[9,10,20,21] that we have determined exclusively by tandem mass spectrometry, including the one from *Chloroflexus* discussed above. It should be noted that the sequence shown for the *Chromatium* thioredoxin in Figure 8 shows Leu in position 42 and Ile in positions 38, 60, and 72, while they were not differentiated (and thus denoted Xle) in the original work.[10] This identification was possible by reinterpretation of the original CID spectra after we had discovered the w_n ions.[19]

Glutaredoxins

Another category of thiol–disulfide exchange enzymes also occurring in organisms from single cell bacteria to mammalian systems is the glutaredoxins, originally called thiotransferases. Their biological activities are, to a certain extent, related to those of the thioredoxins, and they also have a similar active site, namely, two cysteines separated by two amino acids, the first one generally being proline. The structures and properties of glutaredoxins have been the subject of a recent, extensive review.[22] They also are redox proteins of molecular masses in the 12 kDa region and show a higher degree of homology than the thioredoxins.

We have applied the same strategy discussed in the previous section to the determination of the primary structure of two mammalian glutaredoxins from rabbit bone marrow[23] and from human red blood cells (hRBC Grx)[14] and corrected[24] the one published[25] for the protein isolated from calf thymus. Of the few known mammalian glutaredoxins, all are acetylated at the N-terminus and thus not amenable to direct Edman sequencing, a fact which caused some problems with the earlier structures. The sequences of four glutaredoxins are compared in Figure 9. Once we had determined the complete sequence of the protein from rabbit bone marrow and found that the N-terminus was Ac-Ala-Gln-Glu, the two other sequences known at that time and which began with Ac-Gln-Ala-Ala for the pig liver thioltransferase[26] and pyro-Glu-Ala-Ala for the calf thymus protein,[25] respectively, became suspect.

The error in the former was actually caused by a misinterpretation of the CID spectrum obtained with a single double-focusing mass spectrometer operated in the so-called "linked scan" mode,[27] rather than a true tandem mass spectrometer as used in our studies. That error was soon corrected by the authors of the earlier report[26] based on the DNA sequence of the pig liver gene.[28] The problem with the calf thymus glutaredoxin was more severe, as it not only had an incorrect N-terminal sequence but also lacked the 68–71 stretch (-Thr-Val-Pro-Arg-) present in both the pig liver[28] and the rabbit bone marrow[23] enzymes. Obviously, a short tryptic peptide had been missed (amino acid 67 is arginine).

It was a simple matter to correct this structure by digesting reduced and alkylated calf thymus glutaredoxin with chymotrypsin and measuring the *m/z* values of the $(M + H)^+$ ions of the peptides in the digest. Chymotrypsin was chosen to avoid missing the tetrapeptide again. The chymotryptic digest indeed produced all peptides expected from a sequence that included the "missing" tetrapeptide and the modified N-terminus, i.e., $(M + H)^+$ = *m/z* 718.4 and 478.2, respectively. The CID spectrum of the former revealed the sequence Thr-Val-Pro-Arg-Val-Phe, and that of the latter indicated Ac-Ala-Gln-Phe. Thus in a single experiment, the correct structure shown in Figure 9 was established.[24] Had MALDI or ESI been already available at that time (1988) in our laboratory, a single molecular weight measurement would have indicated directly that the published structure[25] could not be correct, but the FAB and CID spectra on the digest were still more conclusive and specific.

The amino acid sequence of the glutaredoxin isolated from human red blood cells[29] presented some surprises. First, when the isolate was reduced prior to alkylation, the molecular mass measured by MALDI-TOF-MS *decreased* from 11 841 Da to 11 688 Da,

(20) Johnson, T. C.; Yee, B. C.; Carlson, D. E.; Buchanan, B. B.; Johnson, R. S.; Mathews, W. R.; Biemann, K. *J. Bacteriol.* **1988**, *170*, 2406–2408.
(21) Johnson, R. S.; Mathews, W. R.; Biemann, K.; Hopper, S. *J. Biol. Chem.* **1988**, *263*, 9589–9597.
(22) Wells, W. W.; Yang, Y.; Gan, Z.-R. *Adv. Enzymol.* **1992**, *66*, 149–201.
(23) Hopper, S.; Johnson, R. S.; Vath, J. E.; Biemann, K. *J. Biol. Chem.* **1989**, *264*, 20438–20447.
(24) Papayannopoulos, I. A.; Gan, Z.-R.; Wells, W. W.; Biemann, K. *Biochem. Biophys. Res. Commun.* **1989**, *159*, 1448–1454.
(25) Klintrot, L.-M.; Höög, J.-O.; Jörnvall, H.; Holmgren, A.; Luthman, M. *Eur. J. Biochem.* **1984**, *144*, 417–423.
(26) Gan, Z.-R.; Wells, W. W. *J. Biol. Chem.* **1987**, *262*, 6699–6703.
(27) Jennings, K. R.; Dolnikowski, G. G. In *Methods in Enzymology. Mass Spectrometry, Vol. 193*; McCloskey, J. A., Ed.; Academic Press: San Diego, 1990; pp 37–61.
(28) Yang, Y.; Wells, W. W. *Gene* **1989**, *83*, 339–346.
(29) Mieyal, J. J.; Stark, D. W.; Gravina, S. A.; Dothey, C.; Chung, J. S. *Biochemistry* **1991**, *30*, 6088–6097.

```
Chlorobium  Ala Gly Lys Tyr Phe Glu Ala Thr Asp Lys Asn Phe Gln Thr Glu 15
Chromatium  Ser Asp Ser Ile Val His Val Thr Asp Asp Ser Phe Glu Glu Glu
 R. rubrum                  Met Lys Gln Val Ser Asp Ala Ser Phe Glu Glu Asp
    Rabbit       Val Lys Gln Ile Glu Ser Lys Ser Ala Phe Gln Glu Val Leu
 Chlorofl.       Ala Lys Pro Ile Glu Val His Asp Ser Asp Phe Ala Glu Lys

Chlorobium  Ile Leu Asp Ser Asp Lys Ala Val Xle Val Asp Phe Trp Ala Ser 30
Chromatium  Val Xle Lys Ser Pro Asp Pro Val Leu Val Asp Tyr Trp Ala Asp
 R. rubrum  Val Leu Lys Ala Asp Gly Pro Val Xle Val Asp Phe Trp Ala Glu
    Rabbit  Asp Ser Ala Gly Asp Lys Leu Val Val Val Asp Phe Ser Ala Thr
 Chlorofl.  Val Leu Gln Ser Lys Thr Pro Val Val Val Asp Phe Trp Ala Pro

Chlorobium  Trp Cys Gly Pro Cys Met Met Xle Gly Pro Val Ile Glu Gln Leu 45
Chromatium  Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Val Leu Asp Glu Ile
 R. rubrum  Trp Cys Gly Pro Cys Arg Gln Xle Ala Pro Ala Leu Glu Glu Leu
    Rabbit  Trp Cys Gly Pro Cys Lys Met Ile Lys Pro Phe Phe His Ala Leu
 Chlorofl.  Trp Cys Gly Pro Cys Arg Val Ile Ala Pro Ile Leu Asp Lys Leu
                                                        Asn

Chlorobium  Ala Asp Asp Tyr Glu Gly Lys Ala Ile Ile Ala Lys Xle Asn Val 60
Chromatium  Ala Asp Glu Tyr Ala Gly Arg Val Lys Xle Ala Lys Xle Asn Ile
 R. rubrum  Ala Thr Ala Leu Gly Asp Lys Val Thr Val Ala Lys Ile Asn Ile
    Rabbit  Ser Glu Lys Phe Asn Asn Val Val Phe Ile Glu Val Asp Val Asp
 Chlorofl.  Ala Gly Glu Tyr Ala Gly Arg Leu Thr Ile Ala Lys Val Asn Thr
                                                            Asp Val

Chlorobium  Asp Glu Asn Pro Asn Ile Ala Gly Gln Tyr Gly Xle Arg Ser Ile 75
Chromatium  Asp Glu Asn Pro Asn Thr Pro Pro Arg Tyr Gly Ile Arg Gly Ile
 R. rubrum  Asp Glu Asn Pro Gln Thr Pro Ser Lys Tyr Gly Val Arg Gly Ile
    Rabbit  Asp Cys Lys Asp Ile Ala Ala Glu Cys Glu Val Lys Cys Met Pro
 Chlorofl.  Asp Asp Asn Val Gln Tyr Ala Ser Gln Leu Gly Leu Lys Gly Leu

Chlorobium  Pro Thr Met Leu Ile Xle Lys Gly Gly Lys Val Val Asp Gln Met 90
Chromatium  Pro Thr Leu Met Leu Phe Arg Gly Gly Glu Val Glu Ala Thr Lys
 R. rubrum  Pro Thr Leu Met Ile Phe Lys Asp Gly Gln Val Ala Ala Thr Lys
    Rabbit  Thr Phe Gln Phe Phe Lys Lys Gly Gln Lys Val Gly Glu Phe Ser
 Chlorofl.  Pro Thr Xle Val Ile Phe Lys Asp Gly Arg Glu Val Gly Arg Leu

Chlorobium  Val Gly Ala Leu Pro Lys Asn Met Ile Ala Lys Lys Ile Asp Glu 105
Chromatium  Val Gly Ala Val Ser Lys Ser Gln Leu Thr Ala Phe Leu Asp Ser
 R. rubrum  Ile Gly Ala Leu Pro Lys Thr Lys Leu Phe Glu Trp Val Glu Ala
    Rabbit  Gly Ala Asn Lys Glu Lys Leu Glu Ala Thr Ile Asn Glu Leu Leu
 Chlorofl.  Val Gly Ala Arg Pro Glu Ala Met Tyr Arg Glu Ile Phe Asp Lys
                                                                Lys†
                                                                Lys‡

Chlorobium  His Ile Gly
Chromatium  Asn Xle
 R. rubrum  Ser Val
 Chlorofl.  Val Leu Ala Met Ala
```

Figure 8. Summary of amino acid sequences of thioredoxins determined by tandem mass spectrometry. The notations Lys† and Lys‡ stand for N_ϵ-methyllysine and N_ϵ,N_ϵ-dimethyllysine, respectively.

indicating that a small molecule or a few even smaller molecules were covalently attached to one or more of the cysteines of the protein via a disulfide bond. Secondly, upon treatment of the reduced material with 4-vinylpyridine, which converts $-SH$ to $-SCH_2CH_2$-(C_5H_4N) and thus adds 105 Da per cysteine present, the molecular mass increased by 526 Da to 12 214 Da. This corresponds to five cysteines, which is in contrast to the few other known (at that time) mammalian glutaredoxins (Figure 8) that have only four, two at the active site at positions 22 and 25, and two others toward the C-terminus (positions 78 and 82). Because of this unexpected result, the experiment was repeated with iodoacetamide as the alkylating agent, but the result was the same: five cysteines. This additional Cys was found to be located close to the N-terminus (position 7) on the basis of the CID spectrum of one of the tryptic peptides obtained from ethylpyridylated hRBC Grx (Figure 10). The presence of the alkylated cysteine is clearly indicated by the peak labeled et-Py at *m/z* 106, corresponding to protonated vinylpyridine produced via a CID-generated reversal of the alkylation reaction, and the peak at *m/z* 979.2 labeled $-$et-Py that arises from the loss of the ethylpyridyl group from the precursor ion, $(M + H)^+$. The presence of an Ac-Ala group rather than the isobaric N-terminal Leu or Ile is indicated by the b_1 ion, which we have only observed for N-acylated peptides.

```
                                                                    15
PiG:  Ac-Ala Gln Ala Phe Val Asn Ser Lys Ile Gln Pro Gly Lys Val Val
CTG:  Ac-Ala Gln Ala Phe Val Asn Ser Lys Ile Gln Pro Gly Lys Val Val
RMG:  Ac-Ala Gln Glu Phe Val Asn Ser Lys Ile Gln Pro Gly Lys Val Val
hEr:  Ac-Ala Gln Glu Phe Val Asn Cys Lys Ile Gln Pro Gly Lys Val Val

                                                                    30
PiG:      Val Phe Ile Lys Pro Thr Cys Pro Phe Cys Arg Lys Thr Gln Glu
CTG:      Val Phe Ile Lys Pro Thr Cys Pro Tyr Cys Arg Lys Thr Gln Glu
RMG:      Val Phe Ile Lys Pro Thr Cys Pro Tyr Cys Arg Lys Thr Gln Glu
hEr:      Val Phe Ile Lys Pro Thr Cys Pro Tyr Cys Arg Arg Ala Gln Glu

                                                                    45
PiG:      Leu Leu Ser Gln Leu Pro Phe Lys Glu Gly Leu Leu Glu Phe Val
CTG:      Leu Leu Ser Gln Leu Pro Phe Lys Gln Gly Leu Leu Glu Phe Val
RMG:      Ile Leu Ser Glu Leu Pro Phe Lys Gln Gly Leu Leu Glu Phe Val
hEr:      Ile Leu Ser Gln Leu Pro Ile Lys Gln Gly Leu Leu Glu Phe Val

                                                                    60
PiG:      Asp Ile Thr Ala Thr Ser Asp Thr Asn Glu Ile Gln Asp Tyr Leu
CTG:      Asp Ile Thr Ala Ala Gly Asn Ile Ser Glu Ile Gln Asp Tyr Leu
RMG:      Asp Ile Thr Ala Thr Ser Asp Met Ser Glu Ile Gln Asp Tyr Leu
hEr:      Asp Ile Thr Ala Thr Asn His Thr Asn Glu Ile Gln Asp Tyr Leu

                                                                    75
PiG:      Gln Gln Leu Thr Gly Ala Arg Thr Val Pro Arg Val Phe Ile Gly
CTG:      Gln Gln Leu Thr Gly Ala Arg Thr Val Pro Arg Val Phe Ile Gly
RMG:      Gln Gln Leu Thr Gly Ala Arg Thr Val Pro Arg Val Phe Leu Gly
hEr:      Gln Gln Leu Thr Gly Ala Arg Thr Val Pro Arg Val Phe Ile Gly

                                                                    90
PiG:      Lys Glu Cys Ile Gly Gly Cys Thr Asp Leu Glu Ser Met His Lys
CTG:      Gln Glu Cys Ile Gly Gly Cys Thr Asp Leu Val Asn Met His Glu
RMG:      Lys Asp Cys Ile Gly Gly Cys Ser Asp Leu Ile Ala Met Gln Glu
hEr:      Lys Asp Cys Ile Gly Gly Cys Ser Asp Leu Val Ser Leu Gln Gln

                                                                   105
PiG:      Arg Gly Glu Leu Leu Thr Arg Leu Gln Gln Ile Gly Ala Leu Lys
CTG:      Arg Gly Glu Leu Leu Thr Arg Leu Lys Gln Met Gly Ala Leu Gln
RMG:      Lys Gly Glu Leu Leu Ala Arg Leu Lys Glu Met Gly Ala Leu Arg Gln
hEr:      Ser Gly Glu Leu Leu Thr Arg Leu Lys Gln Ile Gly Ala Leu Gln
```

Figure 9. Summary of amino acid sequences of glutaredoxins: PiG, pig; CTG, calf thymus; RMG, rabbit bone marrow; hEr, human erythrocyte from red blood cells.

Another unexpected result was the fact that both Asn and Asp were found in position 51, indicating the apparent presence of two isoenzymes. This difference was deduced from the data of an endoAsp-N digest that contained, among many others, three peptides of $(M + H)^+ = m/z$ 520.3, 856.3, and 1356.7. They span the regions 46–50 and 51–57, if position 51 is Asp, and 46–57, if position 51 is Asn. Their CID spectra confirmed these assignments. Recently, the cDNA sequence encoding human glutaredoxin was determined, and it was found that the code for amino acid position 51 is AAC, i.e., asparagine and not aspartic acid (W. W. Wells and E. B. Meyer, unpublished). The duplicity found in the protein sequence is thus most likely an artifact of the isolation or purification of the protein, which is not uncommon under certain favorable circumstances.

The unexpected decrease in molecular weight of the hRBC Grx upon treatment with a reducing agent (triethylphosphine), as mentioned earlier, is readily explained by the fact that, in the last step of the purification of glutaredoxins, bis(hydroxyethyl) disulfide ($HOCH_2CH_2SSCH_2CH_2OH$) is added to ensure that the disulfide bridge of the active site is closed. This reagent probably reacted with one of the three remaining free cysteines to form a covalent adduct of the type $-SSCH_2CH_2OH$, adding 76 Da each. Two of these, or three in a 1:2:1 ratio, would account for the decrease in molecular mass (by 153 Da) upon reduction.

Conclusion

The work summarized here demonstrates that the various mass spectrometric techniques available today

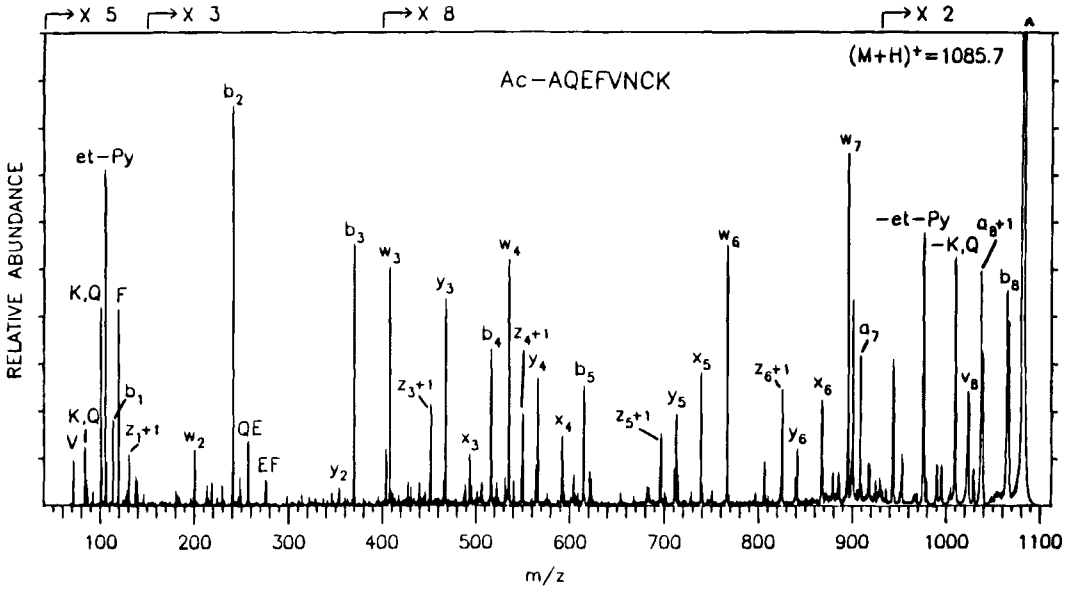


Figure 10. CID mass spectrum of the N-terminal tryptic peptide $(M + H)^+ = 1085.7$ from human glutaredoxin. Reproduced from ref 14 with the permission of Cambridge University Press. Copyright 1994.

permit the determination of the primary structure of proteins in the 10–15 kDa range available in nanomole amounts. There is no reason to believe that this is any limit; larger proteins will just take more effort and time and probably more material. Modifications, such as N-terminal acylation, methylation of lysine, isoforms, etc., that may not be amenable to the conventional Edman degradation or may not be easily detected are readily dealt with or recognized. The throughput of mass spectrometric sequencing is also much faster than the Edman method, which additionally requires pure single peptides, while the tandem mass spectrometer can sequence mixtures. For the same overall efficiency, the cost of one tandem mass spectrometer is less than the number of automated Edman sequencers needed to achieve the same throughput. The latter method can be used to determine a considerable part of the N-terminal sequence of an intact protein, as long as it is not blocked by acylation or a pyroglutamic acid. However, the ability to accurately measure by mass spectrometry the molecular weight of the native product (or what is assumed to be the native product) provides a stringent check on the completeness of the final sequence.

Another approach for the indirect determination of the amino acid sequence of a protein is to sequence the gene encoding it. There one has to keep in mind that any posttranslational modifications, such as the methylation of Lys-104 in the *Chloroflexus* thioredoxin or the N-acetylation of the glutaredoxins, cannot be deduced from the DNA sequence. Furthermore, if the protein is encoded by multiple genes of slightly different DNA sequences corresponding to a multiplicity of amino acids on certain sites (as in the *Chloroflexus* case), there is the risk that only one of the genes would be cloned, thus missing the other variants.

While we have discussed here the sequencing of proteins using a four-sector tandem mass spectrometer, other methodologies are available. Foremost among these are LC–MS systems, particularly HPLC coupled to an ESI triple-quadrupole mass spectrometer, which eliminates some of the sample handling (e.g., collections of fractions). It also substantially increases sensitivity, thus decreasing the amount of material required. The collision spectra of doubly charged (or, less easily, more multiply charged) peptide ions can also be interpreted in terms of their amino acid sequence.

This methodology has been improved to the extreme in Hunt's laboratory[30] for the identification, at the subpicomole level, of peptides presented to the immune system by the major histocompatibility complex molecules. For that purpose, it is not necessary (although it would be desirable) to determine the sequence of all peptides present in that very complex mixture, but it suffices to do so for the more abundant components. It is also only necessary to obtain a CID spectrum that reduces the number of possible sequences to only a few, which can then be synthesized for biological testing. On the other hand, the sequencing of a protein is complete only when the positions of all amino acids are identified correctly, unambiguously, and preferably redundantly, as shown in Figure 7 (which, for the sake of brevity, does not include all the data obtained). If only one peptide is missed in an LC–MS/MS experiment, the entire digest must be reinjected to recover the missing data set.

From Figures 8 and 9, it may appear that the homology among these two sets of proteins would cumulatively aid the sequencing of other members of these families. This is by no means the case. Each new one has to be sequenced independently, and the homology facilitates only the final alignment of the peptides. The four thioredoxins listed in Figure 8 have in common amino acids 23–35, the stretch that includes the active site, and positions 38, 40, and 61 (i.e., only 16 out of 105–110 amino acids). Even among the nonmammalian proteins, only 14 other sites are occupied by the same amino acid. Thus, it is necessary to interpret all CID spectra independent of homology. An example of the potential pitfalls is the case of the *Chloroflexus* peptide discussed earlier in connection with Figure 4a, where the Ala-Gly instead of Lys interpretation would be supported by the presence of Ala in that position in the *R. rubrum*

(30) Hunt, D. F.; et al. *Science* **1992**, *256*, 1817–1820.

thioredoxin. The glutaredoxins (Figure 9) are more homologous, but there, too, it would be difficult to recognize more than a few proteolytic peptides on the basis of their molecular weight.

Future developments in instrumentation will make mass spectrometric protein sequencing less expensive and more widely accessible. The LC–ESI triple-quadrupole system mentioned above is one direction. Another mass analyzer, the "ion trap",[31] may eventually play a significant role as a rather simple and inexpensive device for peptide sequencing. Its present shortcoming is the inability of providing a complete CID spectrum ranging from the low-mass immonium ions, indicative of the presence of various amino acids, all the way to the $(M + H)^+$ precursor ion. Whether the MS^n experiments proposed[31] for overcoming this problem for unknown sequences will be practically useful remains to be seen.

Recently, a modification in the operation of a reflectron-type TOF-MS has been shown to make it possible to obtain fragment ion spectra of peptides similar to those produced by low-energy CID.[32,33] The precursor ion $(M + H)^+$ of the peptide, generated by MALDI, undergoes fragmentation either by collision with the matrix molecules in the plume formed by the impinging laser beam or with residual gas molecules in the flight tube. Because all ions once accelerated retain the same velocity even if they later fragment into an ion and a neutral species, their flight time is unchanged and they are recorded as the *m/z* value of the precursor $(M + H)^+$ ion in a linear TOF-MS like the one used in our laboratory. When the voltage on the reflecting lens is appropriately changed in a reflectron TOF-MS, these ions formed by postsource decay (PSD) are observed at their proper *m/z* value. Because of the relative simplicity of TOF-MS and the high efficiency (i.e., sensitivity) of MALDI, PSD may become a very useful approach to peptide sequencing. Improvements in instrumentation, such as the "curved field reflectron",[34] may further improve the performance and simplify operation.

A major part of the work summarized in this Account was to a great extent carried out by W. R. Mathews, R. S. Johnson, and V. V. Papov. S. Apfalter contributed to the structure determination of the Chloroflexus thioredoxin during a six-month leave from the Technical University of Vienna. The proteins were provided by B. Buchanan (University of California, Berkeley), S. Hopper (University of Pittsburgh), J. J. Mieyal (Case Western Reserve University), and W. W. Wells (Michigan State University). The latter also kindly informed us about the DNA sequence of the human glutaredoxin gene prior to publication. The work was supported by grants from the National Institutes of Health (GM05472 and RR00317).

(31) Cooks, R. G.; Cox, K. A. In *Biological Mass Spectrometry: Present and Future*; Matsuo, T., Caprioli, R. M., Gross, M. L., Seyama, Y., Eds.; Wiley: Sussex, 1994; pp 179–197.
(32) Kaufmann, R.; Spengler, B.; Lützenkirchen, F. *Rapid Commun. Mass Spectrom.* **1993**, *7*, 902–910.
(33) Yu, W.; Vath, J. E.; Huberty, M. C.; Martin, S. A. *Anal. Chem.* **1993**, *65*, 3015–3023.
(34) Cornish, T. J.; Cotter, R. J. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 781–785.