Vol. 195, No. 1, 1993 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
August 31, 1993 Pages 58-64

PROTEIN IDENTIFICATION BY MASS PROFILE FINGERPRINTING

PETER JAMES*[1], MANFREDO QUADRONI[1,2], ERNESTO CARAFOLI[1,2], AND GASTON GONNET[3]

[1]Protein Chemistry Laboratory of the Department of Biology, [2]Institute for Biochemistry, and [3]Institute for Scientific Computation, Swiss Federal Institute of Technology (E.T.H.), 8092 Zürich, Switzerland

Received June 21, 1993

SUMMARY: We have developed an algorithm for identifying proteins at the sub-microgram level without sequence determination by chemical degradation. The protein, usually isolated by one- or two-dimensional gel electrophoresis, is digested by enzymatic or chemical means and the masses of the resulting peptides are determined by mass spectrometry. The resulting mass profile, i.e.,the list of the molecular masses of peptides produced by the digestion, serves as a fingerprint which uniquely defines a particular protein. This fingerprint may be used to search the database of known sequences to find proteins with a similar profile. If the protein is not yet sequenced the profile can serve as a unique marker. This provides a rapid and sensitive link between genomic sequences and 2D gel electrophoresis mapping of cellular proteins.

Protein identification traditionally involves Edman degradation, which requires the isolation of the target in amounts sufficient to obtain direct (N terminal or internal) sequence information. The development of sensitive HPLC separation methods for phenylthiohydantoin amino acid detection and the redesign of the traditional automated Edman sequencers have been combined to produce gas phase machines with a sensitivity in the low picomole range (1) which are commercially available. The subsequent development of 'blot sequencing' allowed N-terminal sequencing of proteins isolated from complex mixtures using SDS-PAGE on supports compatible with the new sequencers without the need for the proteins to be first purified to homogeneity (2). The technique was extended to proteins isolated by 2D gel electrophoresis with its phenomenal resolving power (3), and to the acquisition of internal sequence information by 'on blot' digestion and microbore HPLC separation of the resulting peptides (4). Recently peptide sequencing by tandem mass spectrometry (5) has become a viable alternative with comparable

* To whom reprint requests should be addressed.

Abbreviation: 2D, two-dimensional.

0006-291X/93 $4.00

sensitivity. The limiting factor of both methods is mainly the difficulty in handling (sub) picomole amounts of peptides obtained from digestions (for an overview see (6)).

Within the next few years the sequences of entire genomes of organisms will become available; the complete sequence for yeast chromosome III having already been published (7). The complete sequence of the *E. coli* genome will probably be completed by 1995 and that of yeast by 1998. Larger projects such as the *Drosophila* (8) and human genomes (9) are targeted for completion within the next twenty years. The problem of protein identification can then be reduced to how to extract information from a protein which will match it uniquely to a sequence in the genomic databases? We propose that protein recognition be performed by fragmenting the protein with a chemical or a protease with a high bond specificity (such as cyanogen bromide or trypsin) and then determining the masses of the peptides produced to give a set of molecular masses termed the mass profile (10). The method described here allows one to search a database of mass profiles generated by theoretical fragmentation of all known protein sequences with the experimentally determined profile, a process we term mass profile fingerprinting. A mass profile of a protein digest can be rapidly obtained using HPLC-electrospray ionisation mass spectrometry or Laser desorption time of flight mass spectrometry.

MATERIALS AND METHODS

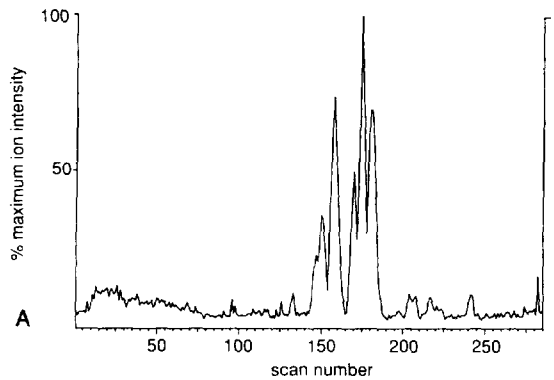
Gel electrophoresis. Intact cells were disrupted and treated with DNase and then placed directly into the sample buffer for electrophoresis. Standard two dimensional gel electrophoresis was carried out (3) and the gels stained using the zinc imidazole method (11). Standard staining methods with dyes such as Coomassie blue or by the silver impregnation can not be used since they (or the accompanying fixation procedures) cause chemical modifications of the proteins which makes the fingerprinting unviable. All the results shown here were generated with 10 picomoles of protein.

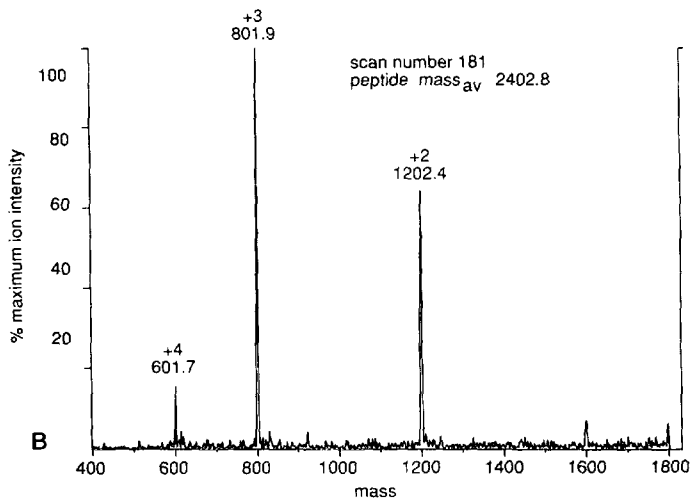
Protein digestion and generation of the mass profile. The protein of interest was cut out of the gel and the protein digested 'in situ', with a protease such as trypsin or Asp-N (12, 13). The peptides were extracted from the gel pieces and redigested with the same protease to ensure complete digestion before injection onto a modified capillary HPLC system (14). The peptides were separated on an Applied Biosystems (Foster City, CA, USA) model 140A HPLC with 785 detector equipped with a Z cell from LC Packings Ltd (Zürich, Switzerland) using a C18 capillary column from the same firm. The flow of 2 µl per minute was directed into a TSQ710 triple quadrupole mass spectrometer (Finnigan MAT, San Jose, CA, USA) using an electrospray ionisation interface (Analytica of Branford, Branford, CT, USA). This device produces a fine spray of liquid particles containing the peptides which rapidly diminish in size due to coulombic repulsion and to drying with a heated gas.(15) The peptides are entirely solvent free by the time they reach the mass spectrometer. The amount of peptide required for the determination of a molecular mass on such a machine is of the order of tens of femtomoles (18). Scans were accumulated from 400 to 2000 mass units in 4 seconds. The PEPMAP program used to match the predicted peptides to the chromatogram obtained is supplied by Finnigan MAT.

The mass fingerprinting algorithm. A complete description of the algorithm is given in chapter 20 of "A Tutorial on Computational Biochemistry using Darwin" which can be requested by E-mail from wertli@inf.ethz.ch. The database searching from a mass profile is offered as a free service by an automatic server at the ETH, Zürich. For information send an electronic message to the address cbrg@inf.ethz.ch with the line; help mass search, or help all.

RESULTS AND DISCUSSION

Mass profile generation and data base searching. The process of data acquisition and fingerprint searching is outlined in Fig. 1 using a digest of 10 pmols of bovine brain calmodulin





C

Score	n	k	AC	DE
85.4	10	4	P02594	CALMODULIN. ELECTRIC EEL
85.4	10	4	P02593	CALMODULIN. BOVINE, and others
85.4	10	4	P21251	CALMODULIN. SEA CUCUMBER
83.8	11	4	P02599	CALMODULIN. SLIME MOLD
83.8	11	4	P11120	CALMODULIN. MUSHROOM
65.4	9	3	P05932	CALMODULIN. BETA. SEA URCHIN
64.1	10	3	P07181	CALMODULIN. FRUIT FLY, and others
64.1	10	3	P02595	CALMODULIN. SCALLOP
64.1	10	3	P11121	CALMODULIN. SEA SQUIRT
64.1	10	3	P05419	NEO-CALMODULIN. CHICKEN

D

No.		M+nH(AV)	Scan	Sequence
T1	1521.7	0		ADQLTEEQIAEFK
T2	957.1	186		EAFSLFDK
T3	907.0	159		DGNGTITTK
T4	806.0	177		ELGTVMR
T5	4072.5	178		SLGQNPTEAELQDMINEVD ADGNGTIDFPEFLTMMAR
T6	147.2	0		K
T7	278.4	0		MK
T8	1094.1	206		DTDSEEEIR
T9	522.6	0		EAFR
T10	508.6	0		VFDK
T11	1266.3	150		GDNGYISAAELR
T12	2360.7	0		HVMTNLGEKLTDEEVDEMIR
T13	2491.7	204		EANIDGDGEVNYEEFVQMMTAK

as an example. A trace of the total ion intensity over the mass range scanned is plotted against the scan number (Fig. 1A), each scan representing a mass chromatogram. The data are obtained manually after the run; the scans covering a peptide peak are summed and the peptide mass is calculated from the masses of the various charge states present (Fig. 1B). The mass spectrometer is tuned to give average masses and each mass is checked against a blank run of the protease itself to exclude any autoproteolysis products or impurities coming from the gel. An arbitrary low mass cut off of 400 mass units is used to exclude chemical noise from the solvents, and any very small peptides. The list of masses (the mass profile) is then run against the database. This search provides a list of probable candidates for the unknown protein (Fig. 1C), but in order to confirm the identity, each of the candidate sequences is matched against the raw data (Fig. 1D). If the masses of all peptides (partial digestion of the protein and disulphide linked peptides can be allowed for as well) would match, the unknown protein is identified with a high degree of confidence.

If some peptides do not match one can start searching for possible post-translational modifications which could account for the differences. Alternatively the non-matching peptides could be sequenced in a second run by on-line HPLC tandem mass spectrometry. This would allow isoform differences to be quickly identified. In the example shown using calmodulin (Fig. 1D), two large peaks are unmatched, scans 170 and 181 (tryptic peptides T6, T7, T9, T10 came in the flow through). Scan 170 contains a peak of 1563.6 which is 42 mass units higher than that of the expected N-terminal peptide, T1, which is in agreement with the finding that calmodulin is N-terminally acetylated. The second peptide, T12, in scan 181, mass 2402.8, is a product of incomplete digestion and is 42 mass units higher than predicted. This reflects the known trimethylation of lysine 115, which prevents the expected tryptic cleavage and causes the mass increment of 42.

Algorithm robustness: Mass accuracy, unsequenced proteins and multiple species.

Two tryptic digestions of beta-lactoglobulin were used to test the effect of mass accuracy on the mass profile searching. For the first digest, the mass spectrometer was tuned to unit resolution to obtain accurate mass measurements (+/- 0.01). A second digestion was used with the instrument tuned to low resolution (1-2 units wide at 50% peak height). The results of the mass profile searches are shown in Tables 1A and 1B. The correct match should have been bovine beta-

---

Figure 1. Procedure for data accumulation and analysis by mass profile mapping. The ion current trace for a tryptic digest of 10 pmol calmodulin is shown (A) together with a mass spectrum from scan 181 (B). The spectrum shows the multiply charged species used to calculated the mass of the singly charged peptide. The average molecular masses of the non-charged fragment $(M)_{AV\ extracted}$ from the ion current trace (1264.8, 955.9, 804.3, 507.3, 2401.8) form the fingerprint or profile which is used to search the database. Table (C) shows the output from the computer program. The score is the quality of the match between the given masses and a protein in the database, the higher the score the better the match; n, shows the number of masses that are produced by a theoretical digest of the found protein occurring between the minimum and maximum mass values used in the profile search; k, indicates the number of masses which were successfully matched against those of the products of the theoretical digestion; AC is the accession number in the Swissprot database and DE gives the name and species from which the protein sequence was determined. The sequence of electric eel calmodulin was downloaded from the database and the theoretical tryptic fragments and partial digests were matched with the original raw data using the program PEPMAP. The output of the program (D) shows a list of the peptides and the scan, if any, in which they can be found.

Table 1A Mass profile search of tryptic fragments from bovine beta-lactoglobulin with a mass accuracy of +/- 0.2.The database was searched using the following peptide masses determined from a 10 pmol tryptic digestion of bovine beta-lactoglobulin : 932.5, 1064.2, 673.5, 836.5, 915.5.

Score	n	k	AC	DE
93.1	7	5	P02755	BETA-LACTOGLOBULIN. WATER BUFFALO
78.7	6	4	P02754	BETA-LACTOGLOBULIN. PRECURSOR BOVINE
78.7	6	4	P02757	BETA-LACTOGLOBULIN SHEEP
76.0	7	4	P02756	BETA-LACTOGLOBULIN. PRECURSOR. GOAT
51.2	9	3	P10834	PET 54 PROTEIN. S. CEREVISIAE
49.2	7	3	P27076	60S RIBOSOMAL PROTEIN K. MARXIANUS
49.0	5	2	P16448	LUXH PROTEIN. VIBRIO HARVEYI
48.7	16	3	P11972	SST2 PROTEIN. S. CEREVISIAE
47.2	4	2	P09660	ACETYLCHOLINE RECEPTOR e CHAIN. RAT
46.9	8	2	P24222	HYPOTHETICAL PROTEIN. S. GRISEUS

lactoglobulin, however this sequence was inferred from that of the cDNA and still contains the signal peptide. The sequence of buffalo protein is almost identical but it does not contain the signal peptide and so the N-terminal peptide from the digestion can be matched correctly. The data show that the algorithm clearly has the ability to find matches when present and that the mass accuracy is not the limiting factor (systematic measurement errors are corrected for by the algorithm). The wider window for matching the peptides due to the greater inaccuracy of the masses in Table 1B leads to the appearance of large proteins which show a few random matches to the masses used, this can be corrected for by placing a maximum value for the mass of the protein to be matched (e.g. two times that estimated from the gel). This wide tolerance has an important consequence in that the sensitivity of detection can be increased by decreasing the resolution of the instrument, i.e. by trading accuracy against improved sensitivity. It is possible to further reduce sample handling by simply analysing the unseparated digest using a laser time of flight mass spectrometer (17) which usually has a lower mass accuracy of around 0.2% (for non-reflectron units).

Until the genome projects are finished, the possibility remains of picking up false matches when the protein being fingerprinted is not in the database. This has been tested for

Table 1B Mass profile search of beta lactoglobulin with a fragment mass accuracy of +/- 1.5. The database was searched using the equivalent fragments as for Table 2B but the mass spectrometer was detuned to give a peak width at 50% peak height of 1-2 mass units. The values obtained were: 931.3, 1064.5, 674.1, 838, 913.5.

Score	n	k	AC	DE
58.2	7	5	P02755	BETA-LACTOGLOBULIN. BUFFALO
57.8	19	3	P18163	LONG-CHAIN-FATTY-ACID-COA LIGASE. RAT
49.4	6	3	P02754	BETA-LACTOGLOBULIN PRECURSOR. BOVINE
49.4	6	3	P02757	BETA-LACTOGLOBULIN. SHEEP
49.3	9	3	P05413	FATTY ACID-BINDING PROTEIN. HUMAN
49.1	10	2	P24460	CYTOCHROME P450 IIB11. DOG
48.9	10	2	P09397	STREP RESISTANCE PROTEIN. S. GRISEUS
48.7	12	4	P07742	RDP REDUCTASE. MOUSE
47.7	8	2	P08593	MS18 PROTEIN. S. CEREVISIAE
47.5	7	3	P02756	BETA-LACTOGLOBULIN PRECURSOR. GOAT

Table 1C Mass profile search of a protein not in the database. The database was searched using the following masses of fragments from a tryptic digestion of a virus replication inhibiting plant protein: 1078.2, 1380.7, 1609.5, 1219.6, 917.6, 1111.9.

Score	n	k	AC	DE
69.6	18	2	P13510	CATION EFFLUX PROTEIN A. EUTROPHUS
55.4	15	2	P23739	SUCRASE-ISOMALTASE. RAT
53.4	4	2	P02670	KAPPA CASEIN PRECURSOR. GOAT
51.3	11	3	P02858	GLYCININ G4 PRECURSOR. SOYBEAN
50.5	4	3	P10051	AMINOGLYCOSIDE TRANSFERASE. C. DIVERSUS
50.1	4	3	P17311	DNA PACKAGING PROTEIN. PHAGE T4
49.6	31	3	P11454	ENTEROBACTIN SYNTHETASE. E.COLI
49.6	9	2	P04070	PROTEIN C PRECURSOR. HUMAN
48.9	8	3	P12054	RLX PROTEIN. STAPHYLOCOCCUS AUREUS.
48.7	15	3	P17889	INITIATION FACTOR IF-2. BACILLUS SUBTILIS

several proteins and has not been found to be a problem. Table 1C shows the result of a search using a tryptic fingerprint of a plant protein which inhibits viral replication which is being currently sequenced and is not in the database as yet. In the absence of a sequence which would match the mass profile input the search produced a set of values around 60-70 which is the 'noise level' for two fairly close matches happening at random. The search was carried out using 6 values but only 2 matched. One must take into account the number of masses used in the search, the resultant score, and especially the number of matches found. Clearly the results of the search show that the protein is not closely related to anything in the database. The top scoring proteins can be downloaded from the data base and matched directly against the HPLC trace obtained from the mass spectrometer to provide a conclusive answer (Fig. 1D).

A further problem which may arise is the occurrence of multiple proteins in the sample. In order to illustrate this point, the results of a profile search using a digest of casein kinase II holoenzyme (α and β subunits)are shown (Table 1D). The subunits are found with scores proportional to the number of their respective peptides used in the search . Thus multiple proteins are not problematical, the scoring being approximately proportional to their molar ratios.

Table 1D Mass profile search using a digestion of a mixture of proteins. A tryptic digestion of casein kinase α and β subunits was performed and the following masses were obtained and used to search the database: 489.3, 698.5, 1083.5, 622.4, 756.5, 1105.5, 1686.3, 1309.6, 1872.4, 955.5, 778.7.

Score	n	k	AC	DE
88.0	19	5	P19139	CASEIN KINASE II, ALPHA CHAIN. RAT
86.9	20	5	P19138	CASEIN KINASE II, ALPHA CHAIN. HUMAN
85.9	21	5	P28020	CASEIN KINASE II, ALPHA' CHAIN. X. LAEVIS
83.4	8	4	P13862	CASEIN KINASE II, BETA CHAIN. HUMAN and others
83.4	8	4	P28021	CASEIN KINASE II, BETA CHAIN. X. LAEVIS
82.1	48	5	P07168	WIDE HOST PROTEIN A. TUMEFACIENS
81.4	9	4	P07312	CASEIN KINASE II, BETA CHAIN. BOVINE
73.9	23	4	P13650	GLUCOSE DEHYDROGENASE-B. ACINETOBACTER
73.4	37	5	P11922	INVASIN. YERSINIA PSEUDOTUBERCULOSIS
71.3	65	9	P12798	PHOSPHORYLASE B KINASE BETA. RABBIT

The method described does not require the isolation of pure peptides and subsequent handling (with consequent losses) as does chemical sequencing. It can be combined with on-line sequence determination by tandem mass spectrometry using collisionally activated dissociation to generate partial sequence information. If the protein or a homologue is known, the mass profile is sufficient for identification, if unknown the sequence information may provide enough information to allow oligonucleotide probe construction for cloning. This method provides a sensitive link between the 2D gel electrophoresis protein maps of cells and the sequences in the genome (18). The probability of obtaining an internal sequence from 10 pmoles of protein in a gel by chemical degradation is poor, however only 4 or 5 peptides extracted with a >1% yield would be sufficient for identification by this method. Mass fingerprinting offers far greater sensitivity and rapidity which is of utmost importance in identifying proteins for 2D gel database construction.

REFERENCES

1. Hewick, R. M., Hunkapiller, M. W., Hood, L. E. and Dreyer, W. J. (1981) J. Biol. Chem. 256, 7990-7997.
2. Aebersold, R. H., Teplow, D. B., Hood, L. E. and Kent, S. B. (1986) J. Biol. Chem. 261, 4229-4238.
3. O'Farrell, P. H. (1975) J. Biol. Chem. 250, 4007-4021.
4. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. (1987) Proc. Natl. Acad. Sci. U.S.A. 84, 6970-6974 .
5. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C.R. (1986) Proc. Natl. Acad. Sci. U.S.A. 83, 6233-7623.
6. Tempst, P., Link, A. J., Riviere, L. R., Fleming, M. and Elicone. C. (1990) Electrophoresis 11, 537-553.
7. Oliver S. G., et al. (1992) Nature 357, 38-46.
8. Hartl, D. L. and Lozovskaya, E. R. (1992) Comp. Biochem. and Physiol. B: Comp. Biochem. 103, 1-8.
9. Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) Science 245, 1434-1435.
10. Henzel, W.J., Stults, J.T., and Watanabe C. (1989) Abstracts of the 3rd Protein Scoiety Symposium, Seattle (July 29-August 2).
11. Ortiz, M. L., Calero, M., Fernandez Patron, C. F., Castellanos, L. and Mendez, E. (1992) FEBS letters 296, 300-304.
12. Rosenfeld, J., Capdeville, J., Guillemot, J. C., and Ferrara P. (1992) Anal. Biochem. 203, 173-179.
13. Kawasaki H., and Suzuki, K. (1990) Anal. Biochem. 186, 264-268.
14. Davis, M. T. and Lee, T. D. (1992) Protein Science 1, 935-944.
15. Whitehouse, C. D., Dreyer, R. N., Yamashita, M. and Fenn, J.B. (1985) Anal. Chem. 57, 675-681.
16. Hunt, D. F. et al. (1992) Science 255, 1261-3126.
17. Hillenkamp, F., Karas, M., Beavis, R. C., Chait, B.T. (1991) Anal. Chem. 63, 1193A-1203A.
18. Celis, J. E. et al. (1991) Faseb Journal 5, 2200-2208.