Protein Science (1994), *3*:1347-1350. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Protein identification in DNA databases by peptide mass fingerprinting

PETER JAMES,[1] MANFREDO QUADRONI,[1] 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 (ETH), 8092 Zürich, Switzerland

(RECEIVED May 2, 1994; ACCEPTED May 24, 1994)

Abstract: Proteins can be identified using a set of peptide fragment weights produced by a specific digestion to search a protein database in which sequences have been replaced by fragment weights calculated for various cleavage methods. We present a method using multidimensional searches that greatly increases the confidence level for identification, allowing DNA sequence databases to be examined. This method provides a link between 2-dimensional gel electrophoresis protein databases and genome sequencing projects. Moreover, the increased confidence level allows unknown proteins to be matched to expressed sequence tags, potentially eliminating the need to obtain sequence information for cloning. 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.

Keywords: DNA database; expressed sequence tags (EST); mass spectrometry; protein identification

Peptide mass fingerprinting, the identification of a protein in a database using a set of molecular masses of peptides generated by a specific digestion, is emerging as a reliable and rapid alternative to peptide sequencing by Edman degradation or mass spectrometry. The idea was first put forward by W.J. Henzel in a poster presentation at the Third Symposium of The Protein Society in Seattle, 1989. The idea appeared to lay dormant for a while until a flurry of papers appeared in the middle of 1993 (Henzel et al., 1993; James et al., 1993; Mann et al., 1993; Pappin et al., 1993; Yates et al., 1993). The importance of this method as a means of linking 2-dimensional (2D) gel databases to protein databases was stressed by all of the groups working on the problem. One of the great advantages of 2D electrophoresis is the ability to follow the coordinated change in the expression and posttranslational modification of a large number of proteins simultaneously. Two-dimensional gels of cells in different states can be analyzed by computer and the changes quantitated (Taylor et al., 1982); for example, comparative protein maps of cells and tissues in normal and pathological states are being developed for use as a diagnostic tool (Appel et al., 1991). Two-dimensional gel technology is being applied in many other fields of medicine: in the molecular epidemiology of viruses and bacteria (Cash, 1991), in studying the immune response (Kovarova et al., 1992), and in postimplantation changes in organs (Praxmayer et al., 1992). The acquisition of a mass fingerprint takes ca. 10 min for a matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometer and 30 min for a capillary HPLC run on a quadrupole instrument, and requires only tens of femtomoles for detection. Since upwards of 200 or more proteins can be isolated in sufficient quantities for digestion (<10 pmol) from a single experiment (running multiple gels and then digestions in parallel), mass fingerprinting allows rapid identification of known proteins and provides a unique tag for unknowns, complementing the 2D analysis.

The major drawback of all the programs described so far is that they only use protein sequences, or protein sequences obtained by translation from the cDNA sequence. Computerized extraction of the correct reading frame of cDNA sequences is possible, but the complete extraction of sequences from data produced by the various genome projects is impractical at the moment due to difficulties such as predicting boundaries for small exons/introns, reading frame shifts, and the occurrence of sequences within introns of one protein that code for another protein, among others. Potentially the most useful source of sequence information, which is inaccessible to autotranslation, is the rapidly increasing number of expressed sequence tags (EST), small cDNA sequences obtained from random-primed cDNA libraries (Adams et al., 1991). In release 37 of the EMBL database there are over 4,000 such sequences present, coding on average for approximately 100-150 amino acids. However, in order to extract the protein data, the tags must be translated in all 6 reading frames. The use of such a database immediately poses problems for the algorithms described so far.

In order to obtain reasonable scores, strategies have been used such as establishing a molecular weight estimate or mass win-

Reprint requests to: Peter James, Protein Chemistry Laboratory, Biochemistry III, ETH-Zentrum, 8092 Zürich, Switzerland; e-mail: bcmass@czheth5a.bitnet.

dow, setting an accuracy tolerance, limiting the number of mismatches allowed, or using a scoring system weighted according to the frequency of occurrence of a mass in the protein within a given mass range. These restrictions are not readily applicable to sequences derived from genomic or EST data because only partial or fragmented sequences are represented. We have therefore developed an algorithm that allows a flexible tolerance for matching accuracy, with the score being calculated as the sum of the reciprocal logarithms of the probabilities of each match happening at random within the mass range searched. No constraints are placed on the search parameters. All the protein fingerprint data here were generated using protein digests performed according to Henzel et al. (1993) and Lee and Shively (1990) and analyzed by capillary HPLC coupled to a Finnigan MAT (San Jose, California) TSQ710 triple quadrupole mass spectrometer as described previously (James et al., 1993). A protein database search takes 2 min on a DEC 5000 workstation; the same search of the DNA database requires ca. 30 min. Every time a DNA database search is performed, the algorithm translates the entire database (in all 6 reading frames for DNA) and calculates the mass fingerprints (taking into account any user-specified modified amino acids like pyridylethylcysteine or chemical modification such as deuteration or acetylation).

One of the main problems of mass mapping is determining the certainty of the search result. Digests that produce only a few peptides, or in which the amount of material is so low that mass accuracy suffers, can produce inconclusive results, as can proteins that are not in the database. One established mass spectrometric (MS) technique that can greatly increase the confidence levels in database searching is hydrogen–deuterium exchange. This has already been shown to be an effective tactic in helping interpret MS/MS spectra for peptide sequencing (Sepetov et al., 1993). The number of exchangeable hydrogens in a peptide is sequence dependent, so peptides with similar masses may be distinguished after exchange (A, F, G, I, L, M, and V all have 1 exchangeable hydrogen; C, D, E, H, S, T, W, and Y have 2; K, N, [carboxymethylcys] and Q have 3; and R has 5). The result of a protein database search using a tryptic digestion of creatine kinase B from chicken gizzard is shown in Table 1. Only 6 masses were observed in the spectrum, two of which came from a contaminating protein. The search of the protein database produced a set of low scores with creatine kinase at position 3. After deuterium exchange of the same digest, the new masses, when run against the deuterated protein database, produced a similar set of low scoring results with creatine kinase at position 7. Table 1 shows the expected shifts for each of the observed masses and the difference between the observed and theoretical shifts. Creatine kinase shows the closest correlation between calculated and observed (the single mass difference observed for one of the peptides was due to combined rounding up and down errors in data collection since only integral values were used). The deuterium exchange did not improve the quality of the data; however, because the results of the digests are orthogonal, combining them shows conclusively that the target protein was creatine kinase B from the chicken, because none of the other top 50 scoring proteins appeared in the deuterated top 50.

Another approach to the generation of orthogonal data is the use of mass fingerprints from 2 or more different digestions. The use of a single digestion mass fingerprint is much less effective when searching in DNA databases constructed by translation of all 6 reading frames (Table 2). If one compares the score difference (delta) between the protein used for digestion and the next nonrelated protein, it is obvious that the confidence level for the DNA search is much lower than that for the protein database. The results for the dual digestions, however, are much clearer: the delta improves dramatically, as does the confidence level, because for all of the searches shown only the target and closely related proteins appear in the top 50 of both scoring lists. The use of an approximate molecular weight (estimated by gel elec-

Table 1. *Using deuterium exchange to increase identification confidence*[a,b]

T search position	Protein	Acc. no.	Matching masses	Theoretical mass shifts	Delta (observed − calculated) shifts	D-T search position
1	Ubiquinone-binding protein, human	P14927	691, 759, 1,134, 1,231	14, 14, 21, 21	2, 1, 4, 3	[c]
2	Proline-specific permease, *Saccharomyces cerevisiae*	P15380	688, 759, 1,134, 1,231	17, 10, 27, 19	4, 3, 10, 0	[c]
3	Creatine kinase B, *Gallus gallus*	P05122	685, 692, 759, 1,232	13, 11, 13, 19	0, 1, 0, 0	7
4	Patatin B1 precursor, *Solanum tuberosum*	P15476	575, 759, 1,133	13, 13, 19	3, 0, 2	[c]
5	Patatin B2 precursor, *S. tuberosum*	P15477	575, 759, 1,133	13, 13, 19	3, 0, 2	[c]
6	Patatin class 1 precursor, *S. tuberosum*	P11768	575, 759, 1,133	13, 13, 19	3, 0, 2	[c]
7	40S Ribosomal protein S4, human	P15880	685, 759	11, 15	2, 2	[c]
8	40S Ribosomal protein S4, *Mus musculus*	P25444	685, 759	11, 15	2, 2	[c]
9	40S Ribosomal protein S2, *Drosophila melanogaster*	P31009	685, 759	11, 15	2, 2	[c]
10	59.1-kDa Protein in RPOA3, *Giardia lamblia*	P25203	575, 685, 1133	10, 13, 23	0, 0, 6	[c]

[a] A tryptic digestion of creatine kinase B from chicken brain was carried out, and the peptide masses obtained using a Brukker (Fallanden, Switzerland) MALDI-TOF mass spectrometer were used to search SwissProt release 26:

Weights observed	575, 685, 691, 759, 1,133, 1,232
Deuterated weights observed	585, 698, 703, 772, 1,150, 1,251
Mass shift observed	10, 13, 12, 13, 17, 19

[b] T search, database search using tryptic digestion; Acc. no., SwissProt database accession number; D-T search, database search after deuterium exchange of the tryptic digests.

[c] Not in top 50 scores.

Table 2. *Comparison of single and dual digestion searches in protein and DNA databases*[a]

				Protein database search				DNA database search			
				Single digest		Dual digest		Single digest		Dual digest	
Target protein	Acc. no.[b]	Digest 1[c]	Digest 2	Position	Delta	Position	Delta	Position	Delta	Position	Delta
ATP/ADP carrier protein T1, human	P12235	Trypsin	CNBr	2	−9.2	1	87.1	1	7.3	1	37.8
M6 antigen, human	X64364	Trypsin	AspN	[d]		[d]		2	−16.3	1	57.4
Lambda receptor, *Escherichia coli*	P02943	LysC	Trypsin	1	57.8	1	155.8	1	3.9	1	58.3
Citrate carrier, *Klebsiella pneumoniae*	P31602	Trypsin	AspN	1	54.5	1	120.5	2	−2.1	1	29.8
10-kDa Chaperonin, *Mycobacterium bovis*	P15020	V8	Trypsin	1	31.4	1	176.9	1	5.6	1	119.8
Na/K ATPase alpha 1, rat	P06685	LysC	CNBr	1	37.8	1	143.1	1	15.8	1	81.6
Lipid binding protein P2, bovine	P07926	Trypsin	V8	1	19.5	1	107.7	2	−7.7	1	40.9
Phospholipase C-alpha, rat[e]	X12355	AspN	Trypsin	1	38.6	1	122	2	[f]	1	72.7
Apolipoprotein AI, human	P02647	LysC	Trypsin	1	54.5	1	139.9	1	10.1	1	102.5
Pectate lyase E, *Erwinia chrysanthemi*	P18210	Trypsin	AspN	1	45.5	1	150.4	1	[g]	1	67.2
Enoylpyruvate transferase, *E. coli*	P28909	AspN	LysC	1	49.1	1	177.1	1	31.1	1	109.2
Na/K ATPase beta 1, rat	P07340	LysC	CNBr	1	44.6	1	131.4	1	30.4	1	92.4
Phenylalanine ammonium lyase, parsley	P24481	Trypsin	Trypsin-D	1	45.9	1	80.8	1	9.3	1	27.3
Average of scores				1.08	39.1	1	132.8	1.31	6.72	1	69.0

[a] The examples used were samples submitted to the laboratory for Edman sequence analysis, so the database searching was performed blind and then compared to the results obtained from the Edman analysis. The databases used were SwissProt releases 25, 26, and 27 and EMBL releases 35, 36, and 37. The protein data shown above were obtained by capillary HPLC-MS using a Finnigan MAT TSQ700 triple quadrupole mass spectrometer as previously described (James et al., 1993). Digestions were carried out on polyvinylidene fluoride-blotted samples as described by Henzel et al. (1993).

[b] SwissProt database accession number.

[c] Digest 1 is the digestion used for the single digest search.

[d] Not present in SwissProt release 27.

[e] The same sequence appears in the SwissProt database described as ER-60 protease under the accession number P11598.

[f] The highest score was also a putative phospholipase C from another tissue type.

[g] Equal scoring with an unrelated sequence, alpha-1-113, 5′-terminal region, SwissProt database accession number M22993.

trophoresis) gives clearer results for single digest searches, but is by far no means as effective as a dual data set. This orthogonal approach can be extended into multidimensional searches, because the algorithm allows for multiple digestions and chemical modifications, e.g., the acetylation of amino groups and methylation of carboxyl groups, which is commonly used in MS/MS peptide sequencing (Hunt et al., 1986) as an to aid spectral interpretation.

Currently we are working on automating the data collection process for HPLC-MS. Half of the peptide digest is injected onto the HPLC, and a computer program records the most intense ions and performs MS/MS on those that are above a certain threshold (a similar program has been developed by T.D. Lee and was presented at the 1993 American Society for Mass Spectrometry meeting). The run is then repeated for either the deuterated or methylated peptides, so within an hour or so one can generate data for a 2D search and have MS/MS data that may directly yield some sequence information. MALDI-TOF MS usually only requires a small fraction of a digest to obtain a spectrum, so aliquots can be used for deuterium exchange, acetylation, or methylation. Provided that a suitable matrix is used, the modifications could even be carried out sequentially on the probe tip.

Multidimensional searches (using deuteration, chemical modification, or several digestions) allow matches to be obtained with a high degree of confidence by a comparison of the scoring lists for each of the searches and the combined search, and false positives from unsequenced proteins can be excluded. We have been able to identify proteins from matches to ESTs in the database, though this could only be achieved by using dual digestion data. A major drawback is that >90% of all ESTs contain noncoding characters in the sequence. If the character occurs at a position that does not affect the outcome of translation, the algorithm accepts it; if more than 1 amino acid is possible, an X is placed in the translation product and after digestion peptides containing an X are discarded. A second limitation is that the length of the translation product limits the effectiveness of the search to proteins from the organism from which the ESTs were obtained. The ability to match proteins to such tags allows the synthesis of oligonucleotides for cloning without having to obtain sequence information.

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. An experimental World Wide Web server has been set up at the address http://cbrg.inf.ethz.ch which requires a client with forms capability.

Supplementary material on Diskette Appendix

Appendix 1 (SUPLEMNT directory, James.SUP subdirectory, file James.doc): (1) Help file for e-mail server; (2) Example of search for protein not yet in the database; (3) EST search example.

Appendix 2 (SUPLEMNT directory, James.SUP subdirectory, file James.alg): Description of algorithm.

References

Adams MD, Kelley JM, Gocayne JD, Dubnik M, Polymeropoulus MH, Xiao H, Merril M, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC. 1991. Complementary DNA sequencing: Expressed sequence tags and Human Genome Project. *Science 252*:1651–1656.

Appel RD, Hochstrasser DF, Funk M, Vargas JR, Pellegrini C, Muller AF, Scherrer JR. 1991. The MELANIE project: From a biopsy to automatic protein map interpretation by computer. *Electrophoresis 12*:722–735.

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

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

Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. 1986. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci USA 83*:6233–6237.

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

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

Lee TD, Shively JE. 1990. Enzymatic and chemical digestion of proteins for mass spectrometry. *Methods Enzymol 193*:351–360.

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

Pappin D, Hojrup P, Bleasby A. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol 3*:327–332.

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

Sepetov NF, Issakova OL, Lebl M, Swiderek K, Stahl DC, Lee TD. 1993. The use of hydrogen–deuterium exchange to facilitate peptide sequencing by electrospray tandem mass spectrometry. *Rapid Commun Mass Spec 7*:58–62.

Taylor J, Anderson NL, Scandora AE, Villard KE, Anderson NG. 1982. The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin Chem 28*:861–866.

Yates JR, Speicher S, Griffin PR, Hunkapillar T. 1993. Peptide mass maps—A highly informative approach to protein identification. *Anal Biochem 214*:397–408.