BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **221,** 609–613 (1996)
ARTICLE NO. 0643

Rapid Protein Identification Using N-Terminal "Sequence Tag" and Amino Acid Analysis

Marc R. Wilkins,*,† Keli Ou,* Ron D. Appel,† Jean-Charles Sanchez,† Jun X. Yan,* Olivier Golaz,* Vince Farnsworth,‡ Paul Cartier,‡ Denis F. Hochstrasser,† Keith L. Williams,*,[1] and Andrew A. Gooley*

**Macquarie University Centre for Analytical Biotechnology, Sydney, New South Wales 2109, Australia; †Central Clinical Chemistry Laboratory and Medical Computing Centre, University of Geneva, CH 1211 Geneva 14, Switzerland; and ‡Beckman Instruments Inc., Fullerton, California*

Received March 5, 1996

Proteins can be identified by amino acid analysis and database matching, but it is often desirable to increase the confidence in identity through the use of other techniques. Here we describe a rapid protein identification method that uses Edman degradation to create a 3 or 4 amino acid N-terminal "sequence tag," following which proteins are subjected to amino acid analysis protein identification procedures. Edman degradation methods have been modified to take only 23 min per cycle, and rapid amino acid analysis techniques are used. The Edman degradation and amino acid analysis is done on a single PVDF membrane-bound protein sample. A computer database matching program is also presented which uses both amino acid composition and "sequence tag" data for protein identification. This method represents the most inexpensive, accurate, and rapid means of protein identification, which is ideal for the screening of proteomes separated by 2-D gel electrophoresis. The creation of N-terminal Edman degradation "sequence tags" prior to peptide mass fingerprinting of samples should also be useful.

There is an increasing demand for rapid and accurate protein identification techniques to enable the cost-effective identification of thousands of proteins from 2-D reference gels of whole organisms or tissues. Amino acid (AA) analysis, in conjunction with molecular weight (MW) and pI estimates from gels, has been shown to be useful in this role (1–3). In one study, 75% percent of proteins analysed from an *E. coli* reference map were identified correctly from a single protein spot for each sample (3). Importantly, scores produced by computer identification programs allowed high confidence to be placed in identities that showed "correct" score patterns, and also suggested when protein identities needed to be confirmed by other means (3). Often when score patterns do not provide high confidence in protein identification, the correct identification is ranked within the list of best-matching proteins (2,3).

Two methods have been reported recently as means of increasing confidence in protein identities assigned by rapid techniques. Cordwell *et al.* (4) have shown that proteins can be identified with very high confidence by a combination of peptide mass fingerprinting and AA analysis techniques. Mann and Wilm (5) suggested that the combination of peptide masses and partial sequence information from tandem mass spectrometry provides a highly specific means of protein identification. Whilst both techniques are powerful approaches to protein identification, peptide mass fingerprinting requires lengthy sample preparation prior to sample analysis, and mass spectrometry needs expert operators for data analysis, particularly that from tandem mass spectrometry.

In this manuscript, we combine rapid AA analysis techniques with an N-terminal "sequence tag" created by Edman degradation to achieve protein identification with high confidence. Methods of Edman degradation are described that have a cycle time of 23 min, and modifications to a recently described computer database matching program (3) are presented which allow lists of best-matching proteins by AA composition to be checked with N-terminal sequence tags.

[1] Corresponding author: Fax: 61-2-850-8174.

0006-291X/96 $18.00

MATERIALS AND METHODS

Sample preparation, amino acid analysis, and N-terminal Edman degradation. Proteins from human sera and *E. coli* were separated by micropreparative 2-D gel electrophoresis, blotted to polyvinylidene difluoride (PVDF) membranes, and stained with amino black as described (6). Protein pI and MW were estimated from gels (7). To generate N-terminal sequence tags, single PVDF spots carrying protein were subjected to 3 or 4 cycles of rapid Edman degradation on a Porton Instruments 2020 Sequenator equipped with a Beckman System Gold on-line PTH analyser. Steps used for the rapid Edman degradation cycles are shown in Table 1. After spots were subjected to Edman degradation, they were then used for AA analysis according to Wilkins *et al.* (3). This involved the hydrolysis and AA extraction of 12 proteins in 2 h, following which proteins were analysed in an automated Fmoc-based AA analysis system (8,9).

Protein identification by computer database searching. Recently, we described a protein identification program, AA-CompIdent, which matches protein AA composition, pI and MW against the SWISS-PROT database to produce a list of best-matching proteins from the database ranked by a score (3). This program has now been modified to allow N-terminal sequence tag data to be used in protein identification. This is achieved by checking the entire sequence of all best-matching proteins by AA composition for the presence of sequence identical to the sequence tag. Proteins in the list that contain the sequence tag are marked with an asterisk. The computer output shows the predicted N-terminal sequence of all best-matching proteins, allowing the user to check if the sequence tag is N-terminal. Figure 1 shows a typical search result. Note that this computer program is accessible via the world wide web through the ExPASy server: http://expasy.hcuge.ch/ch2d/aacompi.html

RESULTS AND DISCUSSION

Rapid Edman Degradation Cycles Allow High Throughput

Conventional Edman degradation is optimised to produce high repetitive yield at each cycle. Whilst high repetitive yield is necessary for the sequencing of long stretches of proteins (30 to 40 residues), is not necessary if only 3 or 4 N-terminal residues are to be removed to create a sequence tag. Accordingly, we modified Edman degradation methods to shorten cycle time, at the expense of repetitive yield. These methods produced high quality, easily interpreted sequence data, with a cycle time of 23 min. By comparison, the normal program takes 47 min (see Table 1). This allowed the creation of sequence tags in under 100 min. This is a significant saving of machine time and reagents over conventional degradation methods, allowing the tagging of 5 proteins a day with minimal operator intervention. In multi-cartridge sequencers, throughput could be increased to 10 proteins per day.

```
SpotNb ALBU
===============
Tagging: DAH
pI:    5.71    Range: ( 5.21,    6.21)
Mw:   66822    Range: (46775,   86869)

The HUMAN entries having pI and Mw values in the specified range:

Rank  Score  Protein       pI     Mw      N-terminal Sequence
=============================================================================
*  1      8  ALBU_HUMAN    5.67   66472   DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQC
   2     46  CG2A_HUMAN    6.10   48536   MLGNSAPGPATREAGSALLALQQTALQEDQENIN
   3     48  NCF2_HUMAN    5.88   59733   MSLVEAISLWNEGVLAADKKDWKGALDAFSAVQD
   4     54  FETA_HUMAN    5.53   66477   RTLHRNEYGIASILDSYQCTAEISLADLATIFFA
   5     57  GRK4_HUMAN    6.19   57693   MELENIVANSLLLKARQEKDYSSLCDKQPIGRRL
   6     57  GBP2_HUMAN    5.54   66638   QLAGPNEPIDNTKGQLVVNPEALKILSAITQPVV
   7     58  SYG_HUMAN     5.88   77530   MDGAGAEEVLAPLRLAVRQQGDLVRKLKEDKAPQ
   8     60  ICA6_HUMAN    5.55   54672   MSGHKCSYPWDLQDRYAQDKSVVNKMQQRYWETK
   9     61  KNLC_HUMAN    5.94   64786   MSTMVYIKEDKLEKLTQDEIISKTKQVIQGLEAL
  10     62  HBI_HUMAN     5.35   51804   MTAEEMKATESGAQSAPLPMEGVDISPKQDEGVL
=============================================================================
```

FIG. 1. Search result from modified AACompIdent program for human serum albumin, showing SWISS-PROT entry for 10 best matches, and 40 residues of their predicted N-terminus. The program marks proteins which carry the sequence tag with an asterisk. Note how the sequence tag (DAH) is present only in ALBU_HUMAN and is found at the N-terminus. In the best 20 matches the sequence tag was found in 2 other proteins, but not at their N-termini.

TABLE 1
Major Differences between Normal and Rapid Edman Degradation Cycles for the Sequencing of PVDF-Bound Proteins (Full Instructions for Rapid Cycles Are Available from the Corresponding Author)

Program	Time (sec)	Cartridge temp. (°C)	Flask temp. (°C)
Normal Program			
Converting ATZ	716	45	56
R2 Coupling	580	45	56
Drying R2	300	50	56
Primary R3 Cleaving	476	35	56
Secondary R3 Cleaving	475	35	56
Other steps	486		
Total time =	2814		
Rapid Program			
Converting ATZ	118	45	65
R2 Coupling	300	45	65
R2 Coupling	100	55	65
Drying R2	180	55	65
R3 Cleaving	340	35	65
Other steps	342		
Total time =	1380		

Note. R2, di-isopropylethylamine; R3, anhydrous trifluoroacetic acid; ATZ, anilinothiazolinone. Other steps include reagent pressurising, line flushing, and drying.

Protein Identification by Amino Acid Composition and N-Terminal Sequence Tag

In theory, there are either 400 (20^2) or 8000 (20^3) possible 3 residue N-terminal sequence tags of proteins, depending on whether methionine is the N-terminal amino acid. There are 8000 (20^3) or 160000 (20^4) possible 4-residue sequence tags for the same reason. Due to the immense specificity of sequence data, N-terminal sequence tags should offer a powerful means of screening lists of best-matching proteins generated by matching protein AA compositions against databases, producing very high confidence in protein identities. To test this approach, 8 known proteins from a human sera 2-D gel and 4 known proteins from an *E. coli* 2-D gel were subjected to 3 or 4 cycles of rapid Edman degradation, and the same samples then used for AA analysis. A single PVDF-bound protein spot from a single gel was used in each case. Amino acid composition data and N-terminal sequence tag, in conjunction with estimated protein pI and MW, were matched against the SWISS-PROT database using pI and MW windows of ±0.5 units and ±30% respectively. Matching was done only against database entries for the species of interest, and no calibration proteins were used.

Database matching with the combination of protein AA composition, pI, and MW ranked the correct protein identity as #1 in 10 out of 12 cases (Table 2). This is an identification rate similar to that reported previously (3). However, only 2 of the 12 matches showed the distinctive "correct" score pattern that gives high confidence in the #1 ranked protein as the correct identity. Such score patterns are seen where the score of the #1 ranked protein is 30 or less and the score of the #2 ranked protein divided by the score of the #1 ranked protein is greater than 2. These give an indication of the goodness of fit of compositional data with a particular database entry as well as the exclusivity of the match (3).

When database matching was undertaken with the combination of sequence tag, AA composition, pI and MW, the sequence tag allowed the unequivocable identification of 10 out of 12 proteins (Table 2). For these 10 proteins, the sequence tag was found only at the N-terminus of the correct protein identity, as in Figure 1. The high specificity of the sequence tag approach was further highlighted by the fact that the sequence tag was not found in any other best-matching protein in

TABLE 2
Results from AA Composition Identification and Sequence Tag Identification of 12 Proteins from 2-D Gels (See Figure 1 for a Sample Set of Results for One Protein)

Protein identification by AA composition	Score of rank #1	Score of rank #2/ rank #1	N-terminal sequence tag	Number of proteins carrying sequence tag[a]	
				N-terminal	Elsewhere
Correct ID ranked #1:					
A1AT_HUMAN	21	1.2	EDP	1	2
ALBU_HUMAN	8	5.8	DAH	1	2
APA1_HUMAN	30	3.5	DEP	1	0
APA2_HUMAN	80	1.1	Blocked	—	—
APC3_HUMAN	107	1.2	SEA	1	1
FIBG_HUMAN	36	1.1	YVA	2[b]	0
HPT1_HUMAN	59	1.9	VDS	1	0
ALF_ECOLI	27	1.5	XKIF	1	4
ATPB_ECOLI	27	1.1	ATG	1	0
GLYA_ECOLI	21	1.5	MLKR	1	1
Correct ID not ranked #1:					
TTHY_HUMAN	124	1.0	GPT	1	1
PTHP_ECOLI	170	1.3	MFQ	1	0

[a] In list of top 20 proteins generated by matching AA composition, pI and MW against SWISS-PROT database entries for the species of interest.

[b] Matches were against fibrinogen gamma chains a and b, two almost identical proteins ranked #1 and #2 by AA composition matching.

4 cases, and in the remaining 6 cases the sequence tag was found only at non N-terminal positions of other proteins. The N-terminal sequence tag did not completely resolve the identity of one protein spot, as the same sequence tag was found at the N-terminus of fibrinogen gamma chain a and fibrinogen gamma chain b. Nevertheless, as these proteins were ranked #1 and #2 by AA composition identification, it was clear that the correct identity was one of these closely related molecules. The final protein had a blocked N-terminus (APA2_HUMAN), which made it impossible to create a sequence tag. However in the list of 20 best-matching proteins by AA composition for this sample, only APA2_HUMAN showed glutamine as the N-terminal amino acid, modified to pyroglutamate to create a blocked N-terminus (10). Thus in this case, the fact that the rank #1 protein is known to be blocked has increased the likelihood that it represents the correct protein identity. With blocked protein samples, the use of quantitative AA analysis after Edman degradation checks the possibility that sequence was not obtained due to a lack of protein. This is always an important consideration if protein sequence cannot be determined.

CONCLUSIONS

We have described a rapid method for increasing confidence in protein identities assigned by AA composition, involving an N-terminal sequence tag generated by rapid Edman degradation. This is done sequentially on a single protein sample from a single 2-D gel. Whilst the approach applies the sequence tag concept of Mann and Wilm (5), it does not suffer from the shortcomings of mass spectrometry. The data generated is quickly and easily interpreted. Nevertheless, the creation of Edman degradation sequence tags prior to peptide mass fingerprinting of samples should also be useful. When used for the identification of proteins from 2-D gels, the AA composition and N-terminal sequence tag approach will be instrumental in the confirmation of open reading frames from genome sequencing projects, and defining the N-termini of proteins. If proteins are blocked, the computer program can also be used to establish which best-matching proteins carry, for

example, a certain sequence generated by internal peptide mapping. This method has the potential to rapidly assign identities to 50 proteins per week with very high confidence. It is therefore an ideal means for screening proteomes separated by 2-D gel electrophoresis, linking protein maps to genomes.

ACKNOWLEDGMENTS

We acknowledge the financial assistance of GBC Scientific Instruments (Dandenong, Victoria, Australia). KLW and AAG acknowledge the financial support of an Australian Research Council Grant. DH acknowledges the assistance of a Montus Foundation Grant and the Swiss National Fund for Scientific Research. MRW was the recipient of an Australian Postgraduate Research Award, and JXY is a MUCAB PhD Scholar.

REFERENCES

1. Jungblut, P., Dzionara, M., Klose, J., and Wittmann-Leibold, B. (1992) *J. Prot. Chem.* **11,** 603–612.
2. Hobohm, U., Houthaeve, T., and Sander, C. (1994) *Anal. Biochem.* **222,** 202–209.
3. Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Humphery-Smith, I., Williams, K. L., and Hochstrasser, D. F. (1996) *Bio/Technology* **14,** 61–65.
4. Cordwell, S., Wilkins, M. R., Cerpa-Poljak, A., Gooley, A. A., Duncan, M., Williams, K. L., and Humphery-Smith, I. (1995) *Electrophoresis* **16,** 438–443.
5. Mann, M., and Wilm, M. (1994) *Anal. Chem.* **66,** 4390–4399.
6. Golaz, O., Hughes, G. J., Frutiger, S., Paquet, N., Bairoch, A., Pasquali, C., Sanchez, J.-C., Tissot, J. D., Appel, R. D., Appel, R. D., Walzer, C., Balant, L., and Hochstrasser, D. F. (1993) *Electrophoresis* **14,** 1223–1231.
7. Bjellqvist, B., Hughes, G., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., and Hochstrasser, D. (1993) *Electrophoresis* **14,** 1023–1031.
8. Haynes, P. A., Sheumack, D., Kibby, J., and Redmond, J. W. (1991) *J. Chromatogr.* **540,** 177–185.
9. Ou, K., Wilkins, M. R., Yan, J. X., Gooley, A. A., Fung, Y., Scheumack, D., and Williams, K. L. (1996) *J. Chromatogr. A,* in press.
10. Bairoch, A., and Boeckmann, B. (1994) *Nucl. Acids Res.* **22,** 3578–3580.