CHEMISTRY LETTERS, pp. 499-502, 1981.

A COMPUTER-AIDED SEQUENCE DETERMINATION OF A POLYPEPTIDE FROM THE MASSES AND EDMAN-DEGRADATION OF ITS CONSTITUENT PEPTIDE FRAGMENTS

Yasutsugu SHIMONISHI*, Yeong-Man HONG, Toyoko KITAGISHI, Itsuo KATAKUSE,† Takekiyo MATSUO,†† Hisashi MATSUDA,†† Saburo HARA,††† Tokuji IKENAKA,††† and Yoshiharu IZUMI
Institute for Protein Research, Osaka University, Suita, Osaka 565
† Department of Physics, Faculty of Science, Osaka University,
†† Institute of Physics, College of General Education, Osaka University,
††† Department of Chemistry, Faculty of Science, Osaka University, Toyonaka, Osaka 560

A new method is described for amino acid sequencing of a polypeptide or protein involving Edman-degradation and field-desorption mass spectrometry of its constituent peptides in mixtures. The sequences can be sought from the molecular weights and N-terminal amino acid residues of the constituent peptides and the phenylthiohydantoins released in each cycle of degradation using a computer program.

The Edman method[1] is widely used for amino acid sequencing of polypeptides and proteins. In this paper, we describe a new method for sequencing polypeptides and proteins involving the following procedures: 1) determination by field-desorption mass spectrometry of the molecular weights of constituent peptides in mixtures obtained from a polypeptide or protein by two or more kinds of specific cleavage methods,[2-4] 2) estimation of the phenylthiohydantoins released successively from a mixture of constituent peptide fragments, and 3) computer calculation of amino acid sequences from the data obtained in 1) and 2) using program "PROSEQ1"[5] or "PROSEQ2".[6] We explain the procedure by reporting, as an example, the sequencing of a polypeptide of unknown structure, the N-terminal BrCN fragment of lysozyme of *Streptomyces erythraeus*,[7] which consists of 55 amino acid residues.[8]

The N-terminal BrCN fragment[8] (0.7 mg (110 nmol) and 1.0 mg (160 nmol)) of *Streptomyces erythraeus* lysozyme was digested with trypsin and chymotrypsin, respectively, as described,[4] and the digests were lyophilized. Immediately before mass measurements the lyophilized material was dissolved in a mixture of pyridine and water (1/1, v/v). The solution (1-2 μl containing 1-2 μg of peptides) was applied to an emitter of a mass spectrometer. Field-desorption mass spectra were obtained with a second-order double focusing mass spectrometer,[9] as described.[4] The remaining solution was subjected to Edman-degradation and the resulting phenylthiohydantoins were quantitated not only to estimate roughly the number of peptides in mixtures but also to know the kinds and ratios of the phenylthiohydantoins released in each cycle of degradation in a high-performance liquid chromatograph using a Zorbax ODS column (4.6 mm x 25 cm), under similar conditions to those described.[4] Amino acid sequences were sought using

the computer programs "PROSEQ1"[5] and "PROSEQ2"[6] in the ACOS 700 computer of the Crystallographic Research Center (Institute for Protein Research, Osaka University).

The field-desorption mass spectra of the tryptic and chymotryptic peptides and their peptide fragments obtained after the first-cycle of Edman-degradation are illustrated in Fig. 1. The spectra of tryptic peptides were considerably simple, suggesting that the original polypeptide was specifically digested by this enzyme, whereas those of chymotryptic peptides were rather complicated in the mass region up to 600 atomic mass units. The semi-quantitation of the phenylthiohydantoins released from tryptic and chymotryptic peptides by Edman-degradation is summarized in Table 1. All the possible mass differences before and after degradations of the tryptic and chymotryptic peptides were examined, as described.[4] In this case, the mass spectra were also examined for mass peaks of $[M+2H]^{2+}$, $[M+H-H_2O]^{+}$, $[M+Na]^{+}$, etc. and peptides containing a Lys residue after the first-cycle of degradation.[4] The results inserted in Fig. 1 indicate the presence of six tryptic peptides with R, F, A and N[10] as N-termini and ten chymotryptic peptides with A, N, K, Q, F, W and V as N-termini. It was also found that the mass peaks at m/z=627, 826 and 1623 in tryptic peptides and at m/z=925 and 935 in chymotryptic peptides each contain a Lys residue. These findings are compatible with the presence of 3 Lys residues in the original polypeptide.

The molecular weight of the polypeptide was assumed to be within the range of 6280 and 6293 from the amino acid composition,[8] as shown in Fig. 2. This molecular weight was compared with the sum of the molecular weights of the tryptic and chymotryptic peptides, respectively. The molecular weight of each peptide was obtained by subtracting one proton from the mass value, because the mass peaks of free "underivatized" peptides are generally observed as quasi-molecular ions ($[M+H]^{+}$).[11-12]

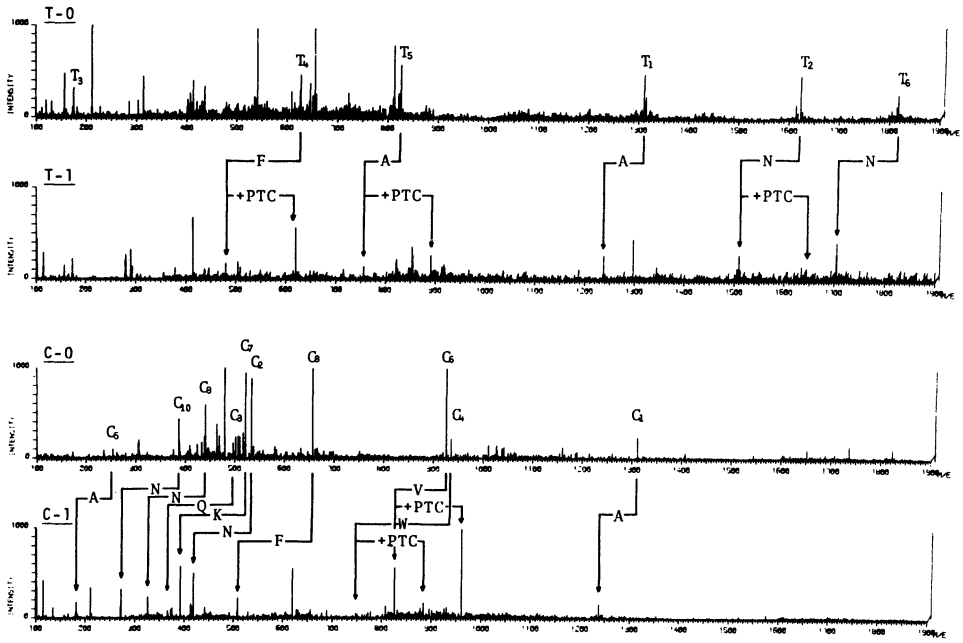

Fig. 1. Field-desorption mass spectra of tryptic (T-0) and chymotryptic (C-0) peptides and their one-step degraded peptide fragments (T-1 and C-1). Individual spectra were measured partly and patched together.

Table 1. Phenylthiohydantoins released from tryptic and chymotryptic peptides of the N-terminal BrCN fragment of *Streptomyces erythraeus* lysozyme by Edman-degradation

	Cycle 1		2		3		4		5		6		7		8		9		10		11	
Asp	62	*24*	30	*3*	46	*26*	8	*4*	4	*2*	3	*3*	33	*20*	8	*5*	3	*6*		*2*	1	*2*
Asn	277*	*96*	113*	*3**	11		2		1	*2*	1	*2*				-*		*5*		*2*		*1*
Glu	59		16*		75	*62*	37	*11*	45	*6*	9	*7*	12	*1*	4	*1*		*2*		*1*		
Gln	-*		11*	-*	65*	*11**	31	*8**	5	*5*	1	*6*	1	*5*	2	*1**				*t**	2	*1*
Ser				*4**	-*	-*		-*					1			-*	5	*5*	2	*2**		*3*
Thr		*1*	64	*80*	13	*8*	25	*1*	7	*12*	7	*3*	10	*1*	3	*1*	3	*2*	1		1	*2*
Gly	28	*6*	51	*12*	33	*3*	39	*25*	43	*28*	34	*16*	10	*7*	19	*3*	8	*5*	19	*14*	6	*6*
Ala	211	*121*	102	*55*	48*	-*	64	*32**	22*	*17*	5	*2*	1		1							
Tyr			23		-*	*22**	25	-*	17*	*6*	4	*4*	3	*9*	2	*7*	5	*4*	2	*2*	2	*4*
Met																						
Hse																						
Val	67	*3*	47	*58*	75	*29*	15	*8*	3	*3*	2	*1*	2	*1*	24	*14*	5	*6*	1	*1*	1	*3*
Pro			2	*40*	50*	*2*	3											*1*		*3*		*1*
Trp	69	*2*			-*		1	*2*														
Phe	97	*44*	3	*2*	1			*9*		*1*			t									*1*
Lys	113	*1*	71	*1*	5		29		21	*1*	1				1	*6*		*2*				
Ile	3	*1*	15	*4*	2	*1*	1		1	*1*	39	*20*	5	*4*	1	*1*	1	*1*				*1*
Leu																						
His																					10	*6*
Arg	8	*26*	8	*4*					3		5											

Normal numerals and those in italics are recoveries (nmol) of the phenylthiohydantoins released from chymotryptic and tryptic peptides, respectively. Ten % of the chymotryptic peptides and 15 % of the tryptic peptides were removed for mass measurements before the 1st, 2nd and 3rd cycles of degradation, respectively. Thr was estimated in addition to Δ-Thr. The separations of Asn and Gln and Ser-derivatives and of Ala and Tyr-derivatives were not satisfactory in some cases. In these cases (*), recoveries were estimated as those of the main component. t means the presence of a trace of material.

Consequently, tryptic peptides were expected to be composed of peptides with mass values at m/z=175(R):T_3, 627(F):T_4, 826(A):T_5, 1310(A):T_1, 1623(N):T_2 and 1818(N):T_6 (amino acid residues in parentheses indicate the N-terminal residue of each peptide). The chymotryptic peptides were suggested to be 253(A):C_5, 386(N):C_{10}, 440(N):C_9, 496(Q):C_3, 521(K):C_7, 533(N):C_2, 656(F):C_8, 925(V):C_6, 935(W):C_4 and 1310(A):C_1. In these values, a peptide with a mass value at m/z=496 was assumed to have Q as the N-terminal residue, because intense mass peaks at 479 and 501 were considered to correspond to $[M+H-NH_3]^+$ and $[M+Na-NH_3]^+$, respectively, derived from 496 ($[M+H]^+$), although the mass peak at m/z=496 was weak. Therefore, the N-terminal residue of the peptide with a mass value at m/z=521 was expected to be K, because one Lys residue was observed in the phenylthiohydantoins in the first-cycle of degradation.

The kind and ratio of the phenylthiohydantoins (Table 1) released from the tryptic and chymotryptic peptides in each cycle of Edman-degradation were selected as integer values within the limits between possible maxima and minima, as described in Fig. 2, although the composition of the phenylthiohydantoins were rather complicated. In these cases, the molecular weight of each peptide, the specificity of the enzymes used and also the expected number of peptides in mixtures were considered. The amino acid sequences were sought from the data obtained above with A and U (U denotes an Hse residue in this paper) as N- and C-terminal residues of the original polypeptide,[8] respectively. The out-put sequence was only one and exactly coincided with the sequence from the N-terminal residue to residue 47, which up to now has been inde-

INPUT DATA

AMINO ACID COMPOSITION

G	7	
A	5	*A
S	2	
P	1	
V	4	
T	3	
C	0	
L	0	
I	2	
N+D	7	
Q+E	6	
K	3	
M	0	
H	1	
F	2	
R	2	
Y	6	
W	3	
U	1	*U
TOTAL	55	

N-TERM C-TERM

MASS VALUES, N-TERMINAL SEQUENCES AND PTH-AMINO ACIDS IN EACH CYCLE

CHYMOTRYPTIC PEPTIDES

		2	3	4	5	6	7	8	9	10	11
252	A	G 1,0	1,0	1,0	1,0	1,0	0	1,0	0	1,0	0
385	N	A 2,0	1,0	2,0	1,0	0	0	0	0	0	0
439	N	S 0	1,0	0	0	0	1,0	0	1,0	1,0	0
495	Q	P 0	1,0	0	0	0	0	0	0	0	0
520	K	V 1,0	2,0	0	0	0	0	1,0	1,0	0	0
532	N	T 2,0	0	1,0	0	1,0	1,0	0	0	0	0
655	F	C 0	0	0	0	0	0	0	0	0	0
924	V	L 0	0	0	0	0	0	0	0	0	0
934	W	I 1,0	0	0	0	1,0	0	0	0	0	0
1309	A	N 2,0	0	0	0	0	0	0	0	0	0
		D 0	1,0	0	0	0	1,0	0	0	0	0
		Q 1,0	2,0	1,0	0	0	0	0	0	0	0
		K 1,0	0	1,0	1,0	0	0	0	0	0	0
		E 1,0	1,0	1,0	1,0	1,0	1,0	0	0	0	0
		M 0	0	0	0	0	0	0	0	0	0
		H 0	0	0	0	0	0	0	0	0	1,0
		F 0	0	0	0	0	1,0	0	0	0	0
		R 0	0	0	1,0	1,0	0	0	0	0	0
		Y 2,0	1,0	2,0	2,0	0	1,0	1,0	1,0	1,0	1,0
		W 0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
		U 1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

TRYPTIC PEPTIDES

		2	3	4	5	6	7	8	9	10	11
174	R	G 0	0	1,0	2,0	2,0	1,0	0	1,0	2,0	2,0
626	F	A 1,0	0	1,0	1,0	0	0	0	0	0	0
825	A	S 1,0	0	0	0	0	0	0	2,0	1,0	2,0
1309	A	P 1,0	0	0	0	0	0	0	0	1,0	0
1622	N	V 1,0	1,0	1,0	0	0	0	1,0	1,0	0	1,0
1817	N	T 2,0	0	0	1,0	1,0	0	0	0	0	1,0
		C 0	0	0	0	0	0	0	0	0	0
		L 0	0	0	0	0	0	0	0	0	0
		I 0	0	0	0	2,0	0	0	0	0	0
		N 0	0	0	0	0	0	0	2,0	1,0	1,0
		D 0	1,0	0	0	0	2,0	0	0	0	1,0
		Q 0	1,0	1,0	1,0	1,0	1,0	0	0	1,0	0
		K 0	0	0	1,0	0	0	1,0	1,0	0	0
		E 0	2,0	1,0	1,0	1,0	0	0	0	0	0
		M 0	0	0	0	0	0	0	0	0	0
		H 0	0	0	0	0	0	0	0	0	1,0
		F 0	0	1,0	0	0	0	0	0	0	0
		R 0	0	0	0	0	0	0	0	0	0
		Y 0	2,0	0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
		W 0	0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
		U 0	0	0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

CANDIDATE SEQUENCES

1 AIVAGIDVSGHQR"/DNQYIINQGKRFAYVKATEGTGYKHPYFAQQYNGSYNIGU

Fig. 2. Computer out-put sequence of the N-terminal BrCN fragment of *Streptomyces erythraeus* lysozyme using "PROSEQ2".[6] U and W were put into as many data as possible, because the former derivative was not identified and the latter residue was possibly destroyed under our experimental conditions.

pendently determined by Edman-degradation alone.[8] The results show that the method can be used for sequencing polypeptides and proteins, and compares favorably with other methods;[1,13] although there may be no problem in sequencing a peptide this long with an automated sequencer.[14] We name the new method the "IPR-sequencing method".

References

1) P.Edman and A.Henschen, "Protein Sequence Determination", ed by J.M.Needleman, Springer-Verlag, Berlin-Heidelberg-New York, pp.232 (1975). 2) Y.Shimonishi, Y.-M. Hong, T.Matsuo, I.Katakuse, and H.Matsuda, *Chem. Lett.*, 1979, 1369. 3) T.Matsuo, I.Katakuse, H.Matsuda, Y.Shimonishi, Y.-M.Hong, and Y.Izumi, *Mass Spectroscopy* (Japan) 28, 169 (1980). 4)Y.Shimonishi, Y.-M.Hong, T.Kitagishi, T. Matsuo, H.Matsuda, and I.Katakuse, *Eur.J.Biochem.*, 112, 251 (1980). 5) T.Kitagishi, Y.-M.Hong, and Y.Shimonishi, *Int.J.Pept.Protein Res.* in press. 6) T.Kitagishi, Y.-M.Hong, and Y.Shimonishi in the 18th Symposium on Peptide Chemistry (Nishinomiya) November 15-16 (1980). 7) T.Morita, S.Hara, and Y.Matsushima, *J.Biochem.*(Tokyo), 83, 893 (1978). 8) S.Hara, M.Akira, E.Taketani, and T.Ikenaka in the 53th Annual Meeting of the Biochemical Society of Japan (Tokyo) October 13-16 (1980). 9) H. Matsuda, *Atomic Masses Fundam.Constants* 5, 185 (1976). 10) R.V.Eck and M.O. Dayhoff, eds, "Atlas of Protein Seqeunce and Structure", National Biomedical Research Foundation Publishers, Silver Spring, Md. (1966); A:Ala, D:Asp, E:Glu, F:Phe, G:Gly, H:His, I:Ile, K:Lys, L:Leu, M:Met, N:Asn, P:Pro, Q:Gln, R:Arg, S:Ser, T:Thr, V:Val, W:Trp, Y:Tyr. 11) H.U.Winkler and H.D.Beckey, *Biochem.Biophys.Res.Commun.*, 46, 391 (1972). 12) S.Asante-Poku, G.W.Wood, and D.E.Schmidt, Jr., *Biomed.Mass Spectrom.*, 2, 121 (1975). 13) J.A.Kellcy, H.Nau, H.-J.Forster, and K.Bieman, *Biomed.Mass Spectrom.*, 2, 315 (1975). 14) P.Edman and G.Begg, *Eur.J.Biochem.*, 1, 80 (1967).

(Received February 23, 1981)