to distinguish between the two alternative sequences given above it would be sufficient to know whether Leu-Glu or CMCys-CMCys appeared first in the digest.

(*b*) A second possibility is to use an enzyme to break down the peptide into smaller fragments that can be investigated. Again, because the problem is one of differentiating between a limited number of known sequences, the choice of enzyme could be straightforward. In the present case chymotrypsin should give either CMCys-Asn or Gln-CMCys-Asn as one of the peptides and the identification of one of these two products should decide which sequence is present. In fact the occurrence of the tetrapeptide Asn-Tyr-CMCys-Asn in the 'even' peptides and the tripeptide Tyr-CMCys-Asn in the 'odd' peptides is sufficient to select the first alternative as the correct one.

The elimination of possible sequences to leave only two alternatives is crucially dependent on being able to distinguish aspartic acid residues from asparagine and glutamic acid from glutamine. This may sometimes pose a problem because the identification needs to be absolute and not merely inference from the net charge on the molecule. Thus the unequivocal identification of Glu-Gln and Glu-Asn is essential to establish the A-chain sequence. This was not done in the present work but could have been achieved by use of the Edman procedure on the significant peptides. However, it is obvious that the techniques used in the present work to separate and identify the dipeptides are too slow and cumbersome to make the sequence procedure attractive in its present form. Nevertheless the practical success achieved, especially when viewed in the light of the theoretical aspects discussed in the Appendix, suggests that if a rapid and reliable method of separating and identifying dipeptides can be developed, the technique will become extremely valuable. An obvious possibility, which it is desirable to pursue further, is the use of mass spectroscopy of suitable derivatives of dipeptides, probably in conjunction with gas chromatography, to effect some preliminary separation. This approach would have the added advantage of coping with the problem of identification of side-chain amides and the whole procedure could ultimately become a standard routine with a computer print-out of the final sequence.

Note added in proof. Since submission of this paper my attention has been drawn to a recent article by McDonald *et al.* (1971), who conclude that 'a prolyl residue constitutes an impasse for dipeptidyl aminopeptidase I if it occurs in a peptide substrate at any position other than at the initial NH_2 (sic) terminus'.

I acknowledge the kindness of Dr. F. H. C. Stewart in synthesizing and making available the tetrapeptide used in this work, and also acknowledge the expert technical assistance of Mr. R. Cranston.

References

Blombäck, B., Blombäck, M., Edman, P. & Hessel, B. (1966) *Biochim. Biophys. Acta* **115**, 371

Callahan, P. X., McDonald, J. K. & Ellis, S. (1969) *Fed. Proc. Fed. Amer. Soc. Exp. Biol.* **28**, 661

Gray, W. R. (1967) *Methods Enzymol.* **11**, 139

McDonald, J. K., Zeitman, B. B., Reilly, T. J. & Ellis, S. (1969) *J. Biol. Chem.* **244**, 2693

McDonald, J. K., Callahan, P. X., Ellis, S. & Smith, R. E. (1971) in *Tissue Proteinases* (Barrett, A. J. & Dingle, J. T., eds.), p. 77, North-Holland, Amsterdam

Metrione, R. M., Neves, A. G. & Fruton, J. S. (1966) *Biochemistry* **5**, 1597

Rowlands, R. J. & Lindley, H. (1972) *Biochem. J.* **126**, 685

Shemyakin, M. M., Ovchinnikov, Yu. A., Vinogradova, E. I., Kiryushkin, A. A., Feigina, M. Yu., Aldanova, N. A., Alakhov, Yu. B., Lipkin, V. M., Miroshnikov, A. I., Rosinov, B. V. & Kazaryan, S. A. (1970) *FEBS Lett.* **7**, 8

Thompson, E. O. P. & O'Donnell, I. J. (1966) *Aust. J. Biol. Sci.* **19**, 1139

Wood, K. R. & Wang, K. T. (1969) *Biochim. Biophys. Acta* **133**, 369

APPENDIX

A Theoretical Investigation into the Potential Usefulness of the Cathepsin C 'Domino' Technique

By R. J. ROWLANDS and H. LINDLEY

Division of Protein Chemistry, Commonwealth Scientific and Industrial Organization, Parkville (Melbourne), Vic. 3052, *Australia*

(*Received* 8 *October* 1971)

In the experimental case of the A-chain of insulin investigated in the main paper (Lindley, 1972), deduction of the sequence from the 'odd' and 'even' sets of dipeptides follows a similar strategy to that used in the game of dominoes and gives only two alternative solutions for this 21-residue peptide. In

Table 1. *Ambiguities in sequence of tryptic peptides by 'domino' strategy*

Length of peptide	No. studied	No. unique	Mean no. of possibilities	Worst case of ambiguity
4	136	136	1	—
5	88	88	1	—
6	117	117	1	—
7	92	92	1	—
8	84	76	1.1	2
9	100	99	1.01	2
10	55	55	1	—
11	51	28	1.6	4
12	97	37	2.0	8
13	66	40	1.6	12
14	31	12	3.7	14
15	19	5	3.6	18
16	69	20	4.3	24
17	32	23	1.6	6
18	21	7	2.8	12
19	52	16	2.6	6
20	9	2	10	20
21	14	1	5.4	16
22	12	1	11	16
23	11	1	3.7	6
24–27	15	3	19	60
28–30	7	2	48	352
31–40	11	1	165	1008
41–50	6	0	8000	25920

assessing the general value of this approach to sequence determination it is important to know how frequently sequence ambiguities arise by using this simple technique. Accordingly a computer program was written to take a known protein sequence, generate the tryptic peptides and from these obtain the 'odd' and 'even' dipeptide series. All possible sequences consistent with these dipeptides, the information from the Edman steps and the enzyme specificity were then generated. Most of the protein sequences cited by Dayhoff & Eck (1967–68) have been studied in this way. In all 1195 tryptic peptides of four or more residues have been used, and Table 1 shows a summary of the results obtained. The extent of ambiguity is far less than we had expected and it seems probable that very little additional information is necessary to arrive at the correct sequence for peptides of up to approx. 25 residues, but beyond this the extent of ambiguity increases markedly. However, Table 1 also shows that 97% of all the tryptic peptides examined are less than 25 residues long, so it would appear that the method has considerable practical potential. Moreover, although the number of possibilities may sometimes seem to be large, when the alternatives are examined it becomes obvious that relatively few decisions have to be made to deduce the correct sequence.

As an example of such a situation the B-chain of bovine insulin may be considered. When the 'domino' strategy is applied to the *N*-terminal 22 residue peptide 16 alternative solutions are obtained. These are given in Table 2, and the correct sequence is shown underlined. Examination of this table shows that only three decisions are needed to arrive at the correct sequence. Because the first dipeptide to be released in the 'odd' series is Val-Asn the number of possibilities is immediately reduced from 16 to six. Also, because Gly-Ser is liberated much earlier than Gly-Glu in the 'odd' series of dipeptides, two more possible alternatives are eliminated. Finally, the appearance of Val-Glu before either Val-Cys or Tyr-Leu (also in the 'odd' series) yields a unique solution. In retrospect, it can be seen that decision number 2 is redundant and only decisions 1 and 3 are needed to choose the correct sequence. However, this conclusion could only have been made with prior knowledge of the sequence.

The above example illustrates the possible usefulness of the technique, but as noted in the main paper, the practical application will depend on the development of a technique for rapidly and completely characterizing a mixture of dipeptides (or dipeptide derivatives), preferably with some degree of quantitation, so that possible duplication or even more

Table 2. *Sixteen possible variations of sequence of bovine insulin B-chain from 'odd' and 'even' dipeptide results*

Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Ala-Leu-Tyr-Leu-Val-Cys-Gly-Ser-His-Leu-Val-Glu-Arg
Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Ala-Leu-Val-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Glu-Arg
Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Cys-Gly-Glu-Ala-Leu-Val-Glu-Arg
Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Glu-Ala-Leu-Val-Cys-Gly-Glu-Arg
Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Cys-Gly-Glu-Ala-Leu-Tyr-Leu-Val-Glu-Arg
Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-Leu-Val-Cys-Gly-Glu-Arg
Phe-Val-Cys-Gly-Glu-Ala-Leu-Tyr-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Glu-Arg
Phe-Val-Cys-Gly-Glu-Ala-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Glu-Arg
Phe-Val-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Ala-Leu-Val-Glu-Arg
Phe-Val-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Glu-Ala-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Arg
Phe-Val-Cys-Gly-Ser-His-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Ala-Leu-Tyr-Leu-Val-Glu-Arg
Phe-Val-Cys-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Arg
Phe-Val-Glu-Ala-Leu-Tyr-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Cys-Gly-Glu-Arg
Phe-Val-Glu-Ala-Leu-Tyr-Leu-Val-Cys-Gly-Ser-His-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Arg
Phe-Val-Glu-Ala-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Cys-Gly-Glu-Arg
Phe-Val-Glu-Ala-Leu-Val-Cys-Gly-Ser-His-Leu-Tyr-Leu-Val-Asn-Gln-His-Leu-Cys-Gly-Glu-Arg

frequent replication of a dipeptide sequence would be detected. Mass spectrometry of a suitable dipeptide derivative appears to be the most suitable technique for this. Because there are only 400 dipeptides, one can envisage the eventual compilation of a library of mass spectra of all 400 dipeptide derivatives. After this, complete interpretation of the mass spectrum of quite complex mixtures of dipeptides should be a relatively straightforward procedure with the aid of a computer.

References

Dayhoff, M. O. & Eck, R. V. (1967–68) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Spring

Lindley, H. (1972) *Biochem. J.* **126**, 683