The *N*-Terminal Amino Acid Sequence of Plastocyanin from *Stellaria media* L.

AN EXERCISE TO ESTABLISH CRITERIA FOR THE IDENTIFICATION OF RESIDUES FROM A SEQUENATOR

By BARRY G. HASLETT and DONALD BOULTER

Department of Botany, University of Durham, Durham City DH1 3LE, U.K.

(Received 28 April 1975)

The *N*-terminal amino acid sequence of plastocyanin from *Stellaria media* L. (chickweed) has been determined by means of an automatic sequencer. Amino acid derivatives obtained during the degradation were identified by a combination of t.l.c., g.l.c. and regeneration of the parent amino acid. The analytical results are presented for each sample, and the criteria used to establish the amino acid sequence are discussed.

The development of automated techniques for sequential degradation of peptides and proteins (Edman & Begg, 1967) has been a major advance in the field of protein chemistry. These methods are now in routine use in a number of laboratories, one of the main advantages being the speed at which sequences can be determined. Although primarily used to determine *N*-terminal amino acid sequences of proteins, these techniques are increasingly being used in the determination of complete sequences (see, e.g., Kolb *et al.*, 1974). Instead of laboriously purifying large numbers of small peptides obtained from several enzymic digests, manually degrading each and obtaining overlapping sequences, the strategy is changing. Large fragments obtained by specific cleavages at infrequent amino acid residues are degraded automatically and the phenylthiohydantoin derivatives of the amino acids identified; the order of these large fragments in the complete sequence can often be deduced without the need for overlapping-sequence data.

By using automated techniques, however, less redundant information is available for the establishment of the amino acid sequence of a polypeptide or protein than is the case with manual techniques, and therefore the identification of phenylthiohydantoin derivatives obtained during the automated degradation must be unequivocal. The present paper indicates our criteria for the acceptance of the identification of derivatives as exemplified by the determination of the *N*-terminal amino acid sequence of plastocyanin from *Stellaria media* L. The criteria apply only to the residues obtained during degradation in a spinning-cup sequenator (Edman & Begg, 1967).

Materials and Methods

Preparation of plastocyanin

Plastocyanin was extracted and purified from *Stellaria media* L. (chickweed) by the methods

described by Ramshaw *et al.* (1973), the only modification being that leaves that had been frozen at -20°C for 1 month were used as starting material instead of an acetone-dried powder. The yield of purified plastocyanin was 5 mg/20 kg fresh wt. of whole plant.

Automated sequencing

The protein was desalted before sequence determination by passage through a column (1 cm $\times$ 10 cm) of Amberlite MB-1 ion-exchange resin (BDH Chemicals Ltd., Poole, Dorset, U.K.) and subsequently freeze-dried.

Edman degradation was carried out with a Beckman 890C automatic sequencer; 5 mg of protein (450 nmol) was applied to the sequencer in 0.4 ml of 70% (v/v) formic acid. Before the degradation, a cycle was carried out in the absence of phenyl isothiocyanate in order to stabilize the protein film. A single degradation was performed with the 'fast protein' programme as supplied with the instrument (Beckman programme no. 072172C). All sequencer chemicals were obtained from Pierce Chemical Co., Rockford, Ill., U.S.A.

Resultant thiazolinone derivatives were converted into the more stable phenylthiohydantoin derivatives by heating in 1.0 M-HCl at 80°C for 10 min.

Identification of phenylthiohydantoin derivatives of amino acids

Phenylthiohydantoin derivatives of amino acids were identified by a combination of methods (see Scheme 1). Preliminary identification was achieved by t.l.c. on pre-coated silica-gel sheets impregnated with a fluorescent indicator (Eastman-Kodak Ltd., Rochester, N.Y., U.S.A.). Chromatograms were developed with the solvent systems of Jeppsson & Sjöquist (1967). Derivatives were seen as fluorescence-quenched spots and were identified by their position in relation to those of a standard mixture.

Confirmation of the preliminary identification was

Scheme 1. Scheme for the identification of phenylthiohydantoin derivatives of amino acids obtained during the automated degradation of chickweed plastocyanin

For details, see the text. Regeneration from the aqueous phase was only carried out if no identification was made from the organic phase.

achieved by using g.l.c. or after regenerating the parent amino acid in constant-boiling HI (Inglis *et al.*, 1971). The parent amino acid was identified either by using a Locarte amino acid analyser or by dansylation with subsequent polyamide-layer chromatography of the Dns-amino acid (Ramshaw *et al.*, 1970).

G.l.c. was carried out with a Varian 1400 gas chromatograph, methods described by Pisano & Bronziert (1969) and Beckman (1973) being used. The column contained Chromasorb W AW-DCMS (100–120 mesh) coated with SP 400 (Phase Separations Ltd., Queensferry, Flint., Wales, U.K.). Silylation of samples was performed by the 'on-column' method (Beckman, 1972) with *N,O*-bis(trimethylsilyl)acetamide as the silylating agent.

Results

The *N*-terminal amino acid sequence of chickweed plastocyanin is shown in Table 1, together with the methods used to identify each residue. The residues appear to have been identified unambiguously with the exception of numbers 36, 37 and 39. Proline, phenylalanine and alanine were found by t.l.c. and g.l.c. of sample 36, but phenylalanine was judged to have been due to carry-over from phenylalanine in sample 35 and alanine to background. Histidine in sample 37 was identified after regeneration of the parent amino acid from the aqueous phase, since t.l.c. and g.l.c. are ineffective in the identification of this residue. The valine in sample 39 was identified

Table 1. Amino acid sequence of chickweed plastocyanin with the results of analyses of each sample from the sequencer

Procedures for the analyses of samples are given in the text. Phenylthiohydantoin derivative of norleucine (50 nmol) was added to each tube of the sequencer fraction collector. This internal standard allowed the quantitative determination of derivatives by g.l.c. and amino acid analysis. The values in parentheses represent yields of residue in nmol.

Sequence position	T.l.c. identification	G.l.c. identification	Identification of parent amino acid after regeneration		Residue
			Amino acid analysis	Dansylation followed by polyamide-layer chromatography	
1	Ala	Ala (177.5), Leu/Ile (21.2), Val (15.2)			Alanine
2	Glu	Glu (88.4), Ala (23.0), Leu/Ile (9.2)			Glutamic acid
3	Val	Val (92.0), Ala (20.6), Leu/Ile (12.2)			Valine
4	Leu/Ile*, Val	Leu (144.6), Val (32.0), Ala (12.4)			Leucine
5	Leu/Ile	Leu (240.4), Val (20.2), Ala (14.6)			Leucine
6	Leu/Ile, Gly*	Gly (35.8), Leu/Ile (75.6), Val (18.2), Ala (12.4)			Glycine
7	Ala*, Gly	Ala (88.2), Gly (21.4), Leu/Ile (27.4), Val (21.2)			Alanine
8	Ser/Thr		Ala (94.7), Asp (26.4), Gly (26.0), Glu (18.0)	Ala*, Gly, Glu, Asp	Serine
9	Asp	Asp (49.0)	Asp (87.0), Ala (45.6), Gly (26.2)	Asp*, Ala, Lys, Gly	Aspartic acid
10	Gly*, Glu, Lys	Gly (25.2), Ala (21.0), Val (9.2), Leu/Ile (13.8)			Glycine
11	Gly	Gly (17.6), Ala (20.6), Val (9.0), Leu/Ile (10.8)			Glycine
12	Leu/Ile*, Gly	Leu (80.8), Gly (11.8), Val (6.5), Ala (10.1)			Leucine
13	Ala*, Leu, Ile	Ala (66.8), Val (7.4), Leu/Ile (39.2)			Alanine
14	Phe*, Ala	Phe (61.6), Ala (24.4), Val (6.6), Leu/Ile (11.0)			Phenylalanine
15	Val*, Phe	Val (68.2), Phe (47.2), Ala (15.0), Leu/Ile (13.6)			Valine
16	Pro*, Val	Pro (42.2), Val (28.8), Phe (12.4), Ala (11.0)			Proline
17	Asn		Asp (69.4), Gly (25.2), Ala (24.0)	Asp*, Ala, Gly	Asparagine Aspartic acid
18	Thr, Asp*	Asp (61.8)			
19	Leu/Ile*, Asp	Ile (47.8), Phe (13.0), Val (6.8), Ala (5.3)			Isoleucine
20	*Ser/Thr		Ala (44.0), Asp (30.0), Gly (21.4)	Ala*, Asp, Gly	Serine
21	Val	Val (36.0), Leu/Ile (16.0), Phe (6.6), Ala (6.0)			Valine
22	Ala*, Val	Ala (53.4), Val (25.0), Leu/Ile (6.4)			Alanine
23	Ala*, Val	Ala (58.0), Val (11.6), Phe (11.2)			Alanine
24	Gly, Ala	Gly (12.4), Ala (13.6), Val (8.8)			Glycine Glutamic acid
25	Glu	Glu (30.7), Asp (5.6)			
26	Lys*, Glu		Lys (45.2), Asp (60.7), Ser (56.0), Glu (15.0)	Lys*, Gly, Glu	Lysine
27	Glu, Leu/Ile*	Ile (26.2), Phe (8.8), Ala (6.4)			Isoleucine

Table 1.—*continued*

Sequence position	T.l.c. identification	G.l.c. identification	Identification of parent amino acid after regeneration		
			Amino acid analysis	Dansylation followed by polyamide-layer chromatography	Residue
28	Ser/Thr*		α -Aminobutyrate (24.8), Lys (15.2)	α -Aminobutyrate, Gly, Lys	Threonine
29	Phe	Phe (28.8), Leu/Ile (9.2)			Phenylalanine
30	Lys*, Phe		Lys (41.2), Phe (21.8), Ile (9.4)	Lys*, Phe	Lysine
31	Asn*, Asp		Asp (29.6), Lys (11.0), Gly (16.0)	Asp*, Lys	Asparagine
32	Asn		Asp (36.8), Gly (21.0), Ala (15.8)	Asp*, Lys	Asparagine
33	Ala*, Asn	Ala (11.2), Phe (6.2)			Alanine
34	Ala, Gly	Gly (5.8), Ala (6.6)			Glycine
35	Phe*, Gly	Phe (17.8), Ala (5.4)			Phenylalanine
36	Pro, Phe, Ala	Pro (3.1), Phe (10.6), Ala (3.1)			Proline
†37	Nothing		His (10.2), Lys (6.7)	His	Histidine
38	Asn		Asp (19.8), Gly (8.1), Ser (7.1)	Asp*, Gly, Ala	Asparagine
39	Val	Val (3.4), Phe (4.2)			Valine
40	Val	Val (5.2), Phe (2.3)			Valine

* The residue in greater concentration as estimated qualitatively.

† Histidine identified after regeneration of the aqueous phase.

by t.l.c. and g.l.c., although the latter showed the presence of some phenylalanine, which was adjudged to be due to background. Homology lends confidence to these identifications, since proline-36 and histidine-37 are invariant, and residue 39 has always been isoleucine or valine in 30 *N*-terminal sequences of higher-plant plastocyanin, 12 of which have been supported by data obtained by manual methods. However, consideration of homology is no substitute for firm identification and must be used with caution.

The results obtained with residue 26 are noteworthy; they suggest that the sample used for analysis of regenerated amino acids was contaminated with free amino acids and are misleading in the identification of residue 26. This emphasizes the necessity of employing more than one method of analysis of derivatives.

It is also noteworthy that, owing to the inefficiency of conversion of the glycine thiazolinone, the phenylthiohydantoin derivative of glycine (samples 6, 10, 11, 24 and 34) occurred in relatively low yields. This has to be taken into account when identifying this amino acid derivative.

The phenylthiohydantoin derivatives of serine and threonine (samples 8, 20 and 28) showed only faint spots on t.l.c., as a result of some β -elimination in these compounds. After regeneration, these

derivatives were identified as alanine and α -aminobutyrate respectively.

Discussion

Criteria for the identification of residues from automated sequence analysis have been suggested by Smithies *et al.* (1971) and by Niall (1973).

Smithies *et al.* (1971) described a quantitative scheme for the identification of thiazolinones obtained during automated degradation in which the parent amino acids were regenerated from the derivatives and identified by quantitative amino acid analysis. By means of a computer, the data were corrected for background, carry-over and losses in yield. To accept an identification, the yield of a residue had to fall within confidence limits based on the expected yield of that residue when corrected data were plotted in semi-logarithmic form against cycle number.

The main criticism of this system is that the criteria for establishing an identification are too inflexible. The yields of some amino acid derivatives are extremely variable and unpredictable, so that reliable 'correction factors' for them are unobtainable. The complex calculations involved also require the routine use of computing facilities.

Another criticism is that the only method employed by Smithies *et al.* (1971) for the identifi-

cation of amino acid derivatives is that of regenerating the parent amino acid and subjecting this to quantitative amino acid analysis. Sole reliance on this method of identification can lead to erroneous results as a consequence of contamination of analyser samples with free amino acids, artifacts or peptide fragments washed from the sequencer cup (see sample 26 in the Results section). Further, it is advisable to analyse the first few samples from a degradation quickly to observe the progress of the experiment; this cannot be done if the only method of identification used is regeneration and analysis of parent amino acids.

The methods of identification used by Niall (1973) consist of a combination of preliminary t.l.c. followed by quantitative g.l.c. After measurement of total yield relative to an internal standard of the phenylthiohydantoin derivative of norleucine, background (average yield of that residue from the previous two or three cycles) is subtracted and a correction made for carry-over occurring in the subsequent cycle. The criteria for establishing an identification is that its corrected yield, i.e. specific yield, should be at least twice the background for the amino acid derivative, and that no other amino acid derivative should be more than 30% above its background in the sample.

The system used by Niall (1973) has several advantages over that of Smithies *et al.* (1971). First, it uses more than one method of identification. Secondly, the criteria for establishing an identification are more flexible than those of Smithies *et al.* (1971), since they are related to the background yields at specific residue positions. However, it is not practical to adhere rigidly to these criteria. To calculate background yields of each derivative in every sample, it is necessary to regenerate the parent amino acids from each sample and subject these to quantitative amino acid analysis. This is time-consuming, and in fact it would appear that Niall (1973) does not do this either, but estimates the background in each sample qualitatively by the strength of fluorescence-quenched spots on t.l.c. The results obtained with sample 7 (see Table 1 and below) also illustrate the difficulties involved in adhering rigidly to the criteria of Niall (1973).

It is suggested that no objective quantitative rules can be applied to establishing the identification of phenylthiohydantoin derivatives of amino acids without the risk of ambiguous results. The main reason for this is the diverse range of chemical properties exhibited by the derivatives, leading to variation in the yields of the different derivatives obtained during automated degradation, together with the fact that the yield of a derivative can depend on the nature of its immediate neighbours in the sequence (see Brandt & von Holt, 1974). It is important for the main derivative of each sample to be identified by at least two independent methods

of analysis, although this is impractical for the phenylthiohydantoin derivatives of histidine, arginine and cysteine acid, which are analysed solely by regeneration and identification of the parent amino acid. Generally, the yield of a single derivative should increase so as to be significantly above the background yield of it and other derivatives, but no strict quantitative rule exists to define this difference. Unless the residue is repeated in the next sequence position, the yield of derivative should then decrease. As the degradation proceeds, available yield data for the identified residues are plotted against cycle number. This reveals any anomalously low yields after account is taken of those derivatives that are expected to have a lower recovery. Quantitative determination of the derivatives is an essential requirement for establishing the 'cleanness' of the protein and of the degradation; however, in those instances where the parent amino acid is regenerated, it is unnecessary, in our view, to perform a quantitative amino acid analysis, since the results in Table 1 show that the identification of derivatives after regeneration of corresponding parent amino acids is equally accurate whether quantitative amino acid analysis or the qualitative dansyl procedure is used.

Finally, it is necessary to point out any weaknesses in the evidence for an identification when constructing the sequence. To exemplify the application of our criteria, the arguments for the assignment at residue positions 6, 7, 10 and 11 are given below.

Although the yield of leucine/isoleucine at position 6 was 75.6 nmol (Table 1), the data can be interpreted by our criteria to exclude leucine/isoleucine at this position. Thus the yield of leucine/isoleucine is not greater than or approximately equal to the yield of leucine/isoleucine at position 5 i.e. 240.4 nmol; furthermore glycine increases in amount at position 6 and subsequently decreases in position 7 and residue 6 is therefore assigned as glycine. It is of interest to note that whereas glycine was recorded as the major component in sample 6, g.l.c. revealed a higher yield of leucine/isoleucine at this position. This indicates that qualitative t.l.c. alone is not sufficient for the establishment of the major derivative. However, since glycine is the actual residue at position 6, these data also emphasize the necessity of using more than one method of identification.

In considering residue position 7 the use of our more flexible criteria avoids the ambiguous result obtained by the application of Niall's criteria. By using his method, only 'background' (calculated from the preceding samples) and 'carry-over' (calculated from the subsequent sample) are taken into account for comparison with the yield of residue in the sample considered. The background yield of the phenylthiohydantoin derivatives of leucine/isoleucine in position 7 would be calculated by Niall's (1973) method from the average yield of this residue in

samples 1–3, since leucine occurs at positions 4 and 5. The value calculated is approx. 14.0 nmol. However, in position 7 the yield of the phenylthiohydantoin derivatives of leucine/isoleucine is 27.4 nmol, i.e. nearly twice the background value owing to occurrence of leucine at positions 4 and 5. By strict adherence to the criteria of Niall (1973) one would have to identify residue 7 as either alanine or leucine/isoleucine. Our criteria do not suggest this assignment since the yield of leucine/isoleucine does not increase at position 7. Since alanine is the only derivative to increase significantly in yield relative to the preceding sample, the identification is alanine.

In positions 10 and 11, alanine was absent in t.l.c., indicating that 'alanine' seen in g.l.c. (Table 1) was due to contaminating organic matter. This again emphasizes the need for more than one method of identification. The yield of glycine in sample 11 is 30% lower than that in sample 10; even so, it is also identified as glycine. Owing to variable conversion of glycine and the lower accuracy of quantitation of this amino acid due to its poor response on g.l.c., less reliance can be placed on the relative yield data for this residue. On average, we have found the yield of glycine to be 25–30% lower than that expected from a semi-logarithmic plot of the data, but up to 50% lower yields have been recorded. This situation emphasizes the need to evaluate the total information relating to a particular sample rather than only one part. The identification of glycine as residue 11 was also based on the fact that it was the major derivative in the sample as judged from two independent methods of identification (see Table 1) and that no other derivatives fulfilled any of our criteria.

The use of a sequenator is especially valuable in evolutionary studies in which amino acid sequences

of the same protein from many organisms are to be determined. Not only can an identification of a residue be compared homologously, but actual yield data can also be directly compared in sequences in which the neighbouring residues are either the same or different. From consideration of this kind, experience is built up that is far more useful than a set of inflexible quantitative rules. In these circumstances, methods that allow the faster accumulation of data are to be preferred.

We thank Mr. J. Gilroy and Mrs. E. Jones for invaluable technical assistance, and the Nuffield Foundation for financial support.

References

- Beckman (1972) *In Sequence*, issue no. 3, p. 5, Beckman Instruments, Palo Alto
- Beckman (1973) *In Sequence*, issue no. 4, p. 1, Beckman Instruments, Palo Alto
- Brandt, W. F. & von Holt, C. (1974) *Eur. J. Biochem.* **46**, 419–429
- Edman, P. & Begg, G. (1967) *Eur. J. Biochem.* **1**, 80–91
- Inglis, A., Nicholls, P. & Roxburgh, C. (1971) *Aust. J. Biol. Sci.* **24**, 1247–1250
- Jeppsson, J.-O. & Sjöquist, J. (1967) *Anal. Biochem.* **18**, 264–269
- Kolb, E., Harris, J. I. & Bridgen, J. (1974) *Biochem. J.* **137**, 185–197
- Niall, H. D. (1973) *Methods Enzymol.* **27**, 942–1010
- Pisano, J. J. & Bronziert, T. J. (1969) *J. Biol. Chem.* **244**, 5597–5607
- Ramshaw, J. A. M., Thompson, E. W. & Boulter, D. (1970) *Biochem. J.* **119**, 535–539
- Ramshaw, J. A. M., Brown, R. H., Scawen, M. D. & Boulter, D. (1973) *Biochim. Biophys. Acta* **303**, 269–273
- Smithies, O., Gibson, D., Fanning, E. M., Goodflesh, R. M., Gilman, J. G. & Ballantyne, D. L. (1971) *Biochemistry* **10**, 4912–4921