Acta Chemica Scandinavica B 32 (1978) 380—386

The Primary Structure of Soybean (*Glycine max*) Leghemoglobin *c* *

GUNNEL SIEVERS, MARJA-LIISA HUHTALA and NILS ELLFOLK

Department of Biochemistry, University of Helsinki, Unioninkatu 35, SF-00170 Helsinki 17, Finland

The primary structure of soybean (*Glycine max* cv. *Fiskeby*) leghemoglobin *c* has been determined. The polypeptide chain consists of 143 amino acid residues and has a molecular weight of 15 950. The sequence of leghemoglobin *c* completely differs at six positions from that of component *a*. Microheterogeneity was observed at six positions, one alternative always being identical with the corresponding amino acid residue in component *a*. Leghemoglobins c_1 and c_2 differ only at position 143, which is lysine in c_1 and phenylalanine in c_2. The discrepancies between the sequences of leghemoglobin *c* reported by Hurrell and Leach (*FEBS Lett. 80* (1977) 23) and by us are discussed in detail.

Soybean (*Glycine max*) root nodules contain leghemoglobin, which has many features in common with animal globins (*cf.* Ref. 1). The protein consists of one polypeptide chain with a molecular weight of about 16 000 and has protoheme as the prosthetic group. Leghemoglobin can be resolved into at least four components, *a*–*d*,[2] of which leghemoglobin *a* and leghemoglobin *c* are the main ones. Recently leghemoglobin *c* has been further resolved into subcomponents c_1 and c_2.[3]

The amino acid sequence of soybean (cv. *Fiskeby*) leghemoglobin *a* has been determined,[4,5] as have the sequences of leghemoglobins from kidney bean (*Phaseolus vulgaris*),[6] broad bean (*Vicia faba*),[7] and yellow lupin (*Lupinus luteus*).[8] All these primary structures are closely similar.

In this study a detailed report on the primary structure of soybean (cv. *Fiskeby*) leghemoglobin *c* is given. Preliminary reports have been recently published by us,[9] and independently the sequence of leghemoglobin c_2 has been reported by Hurrell and Leach.[10] The two primary structures reported differ at several positions, which stresses the need of a detailed account.

MATERIALS AND METHODS

Soybean leghemoglobin *c* was prepared as previously described,[2] except that the eluting buffer used in the DEAE-cellulose chromatography was sodium acetate, pH 5.6, because leghemoglobin was found to be less stable at pH 5.2.[11] Leghemoglobin *c* was resolved into its subcomponents c_1 and c_2 using a gentle gradient from 0.02 to 0.2 M sodium acetate buffer, pH 5.6. The apoprotein of leghemoglobin *c* was prepared as previously described.[12] In all, about 600 mg protein was used in this study.

Unresolved apoleghemoglobin *c* was digested with trypsin[13] and thermolysin[14] and apoleghemoglobin c_2 with chymotrypsin.[15] Secondary hydrolysis of the peptides was also performed with subtilopeptidase A, pepsin or dilute acid cleavage in 0.03 M hydrochloric acid.[16] The peptides formed were separated by ion exchange chromatography, high voltage paper electrophoresis and paper chromatography.[17] The amino acid composition of the peptides was determined on a modified Beckman/Spinco 120 B amino acid analyzer.

Subtractive Edman and dansyl-Edman degradation, leucine aminopeptidase and carboxypeptidase A were used to elucidate the sequences of the peptides.[16] The *N*-terminus of the protein was determined by the Edman method and the *C*-terminus by hydrazinolysis.[17]

The amide and acidic groups were established from the electrophoretic mobility of the

* Detailed evidence for the amino acid sequence of leghemoglobin *c* has been deposited with the British Library at Boston Spa, Wetherby, West Yorkshire, UK, as Supplementary Publication No. SUP 90 034 (19 pages).

peptides.[18] In some cases the position of asparagine and glutamine was confirmed after Edman degradation and identification of the phenylthiohydantoin (PTH–) derivatives.[16]

Peptide nomenclature. The peptides are numbered consecutively from the *N*-terminus using arabic numerals. In cases of microheterogeneity the peptide pairs are named *A* and *B*. A capital letter before the peptide number indicates the methods of hydrolysis of the protein, as follows: *T*, trypsin; *Th*, thermolysin; *C*, chymotrypsin; *S*, subtilopeptidase A; *P*, pepsin; *A*, dilute hydrochloric acid. The subfragments are numbered consecutively from the *N*-terminus of the peptide fragment.

RESULTS

The *N*-terminal sequence of soybean leghemoglobin c is *NH_2-Gly-Ala-Phe-Thr* as determined by the Edman procedure. Hydrazinolysis of the apoprotein gave 39 % phenylalanine, 34 % lysine and 27 % alanine.

The fractionation of the tryptic peptides on Dowex 1 × 2 is shown in Fig. 1. Fourteen fractions were obtained, each of which contained several peptides. The fractions were further fractionated on Dowex 1 × 2, Dowex 50 × 2, paper electrophoresis and paper chromatography. The hydrophobic peptides *T3* and *T16B* were not found in the fractions. *T3* was purified from a tryptic digest of the apoprotein by high voltage paper electrophoresis at pH 6.5 and then at pH 1.9. In both cases it remained on the starting line and was eluted from the paper with 50 % acetic acid and repeatedly precipitated by dilution. Peptide *T16B*, *Ala-Phe*, was found as a part of the chymotryptic peptides *C10* and *C11*. In all, twenty-one tryptic peptides, one ditryptic peptide and free lysine were obtained, but the two microheterogeneous forms of peptides *T2* and *T3* could not be separated and were studied as mixtures. After purification, *T1*, *T4A*, *B*, *T6 – T8*, *T10A*, *B – T13*, *T15* and *T16A*, *B* were directly sequenced by subtractive Edman or dansyl-Edman procedures. The larger peptides

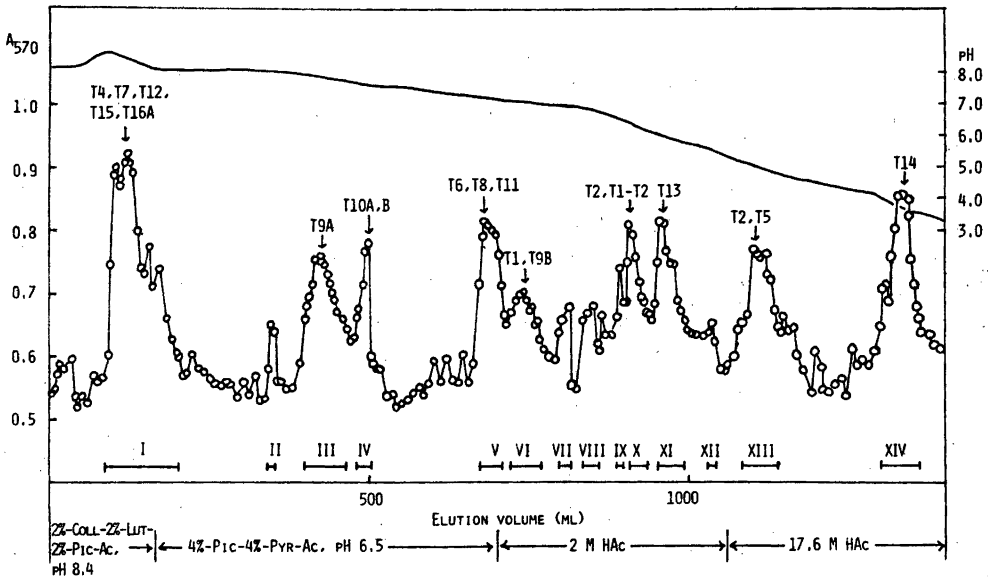


Fig. 1. Fractionation of tryptic peptides from apoleghemoglobin c on Dowex 1 × 2. Tryptic peptides from 32.6 μmol of apoleghemoglobin were fractionated on a Dowex 1 × 2 column (2 × 80 cm), equilibrated to pH 8.2 at 35 °C in a buffer composed of 2 % 2,4,6-collidine – 2 % α-picoline – 2 % 2,6-lutidine, pH adjusted to 8.2 with acetic acid. Elution was started with the pH 8.2 buffer at a flow rate of 22.4 ml/h; 4.3 ml fractions were collected. After 185 ml a gradient was set up with 500 ml of buffer, pH 8.2, in a closed mixing chamber into which 516 ml 4 % α-picoline – 4 % pyridine – acetic acid buffer, pH 6.5, was introduced. The elution was continued with 357 ml 2 M acetic acid and finally with glacial acetic acid. 0.1 ml from each fraction was taken for alkaline ninhydrin test. The pooled fractions are shown with bars.

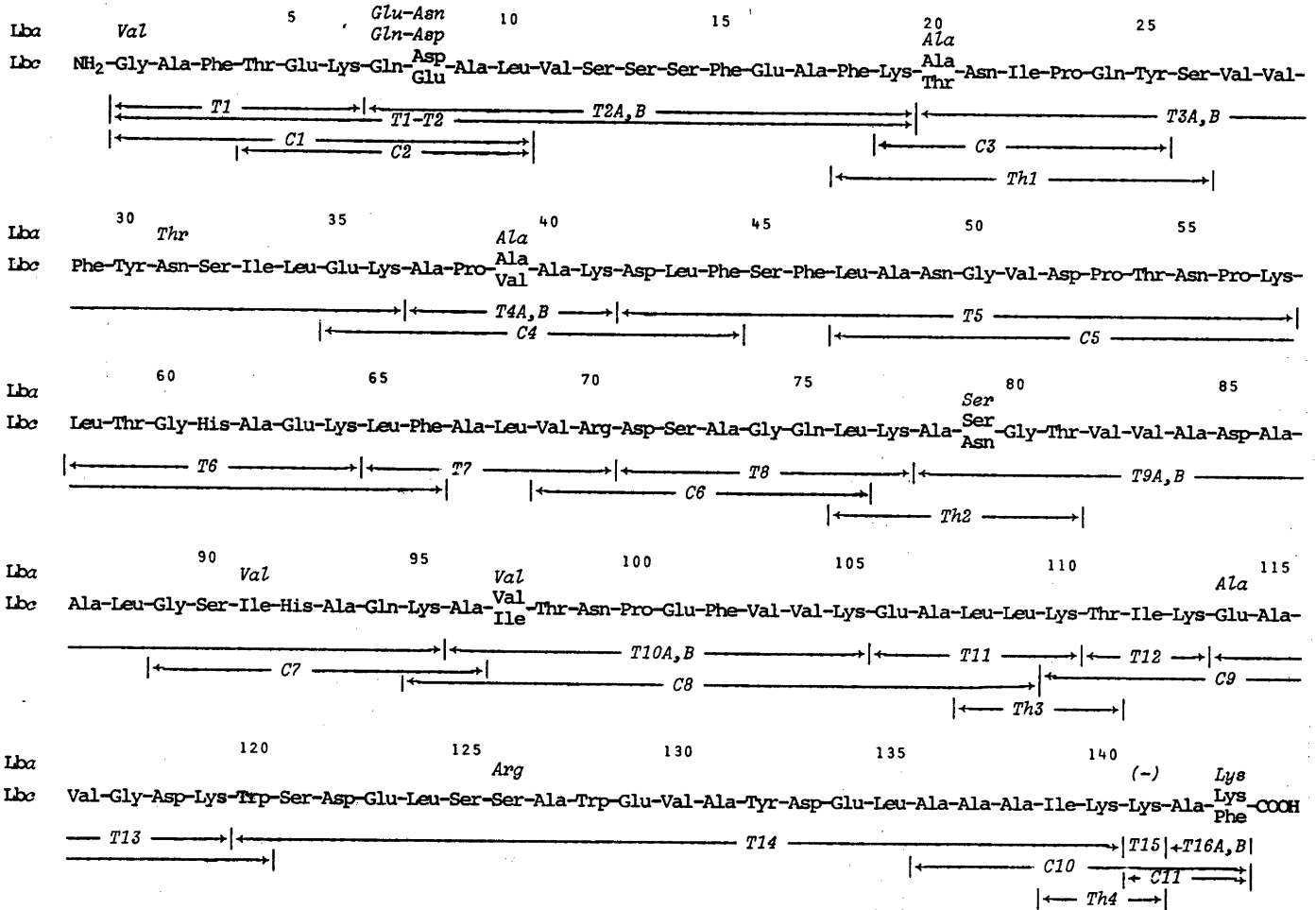


Fig. 2. The amino acid sequence of soybean leghemoglobin *c*. The tryptic (*T*), chymotryptic (*C*) and thermolytic (*Th*) peptides are shown with bars. In cases of microheterogeneity peptide *A* contains the upper alternative and peptide *B* the lower. The corresponding amino acids in leghemoglobin *a* (Lb*a*) are also given when these differ from those in leghemoglobin *c* (Lbc).

T2, *T3*, *T5*, *T9A*, *B* and *T14* were first hydrolyzed with enzymes or hydrochloric acid before the sequences of the fragments were determined.

From the chymotryptic hydrolysis the peptides containing lysine and arginine were isolated and sequenced. In cases of microheterogeneity the peptide pairs were not resolved but used as such. At four positions the order of the tryptic peptides could not be established by chymotryptic peptides. Apoleghemoglobin *c* was therefore hydrolyzed with thermolysin. A great number of peptides were isolated but only those four peptides are given that uniquely overlap peptide borders *T2/T3*, *T8/T9*, *T11/T12* and *T14/T15*.

The complete amino acid sequence of leghemoglobin *c* is shown in Fig. 2, which also gives the corresponding amino acids in leghemoglobin *a* when these differ from those in leghemoglobin *c*. At six positions there are alternative amino acids, giving six pairs of tryptic peptides in addition to nine single ones. One *Lys-Lys* bond is found at position 140 – 141. The polypeptide chain consists of 143 amino acid residues, giving an approximate molecular weight of 15950 for leghemoglobin *c*. Leghemoglobin *c*, like leghemoglobin *a*, contains only two histidines at positions 61 and 93, which are assumed to be the heme-binding histidines.

Supplementary publication

Detailed evidence for the amino acid sequence of the protein has been deposited as Supplementary Publication SUP 90 034 with the British Library (Lending Division) for storage on microfiche. This evidence comprises:

(1) Figures showing elution diagrams of the rechromatography of fractions I, III, XIII and XIV from Fig. 1 on Dowex 1 × 2 and fractions V, VI, IX, X, XI on Dowex 50 × 2.

(2) Tables containing the amino acid compositions, the methods of purification, the electrophoretic mobilities at pH 6.5, and R_{Leu} in butanol – acetic acid – water (4:1:5, v/v) of the tryptic, thermolytic and chymotryptic peptides.

(3) Tables showing the individual sequence evidence for each peptide. This consists of the results from dansyl-Edman, subtractive Edman, leucine aminopeptidase and carboxypeptidase degradations, fragmentations of the primary peptides with different hydrolysis methods and the net charge and the sequence of the subfragments.

DISCUSSION

The *N*-terminus of leghemoglobin *c* is glycine, indicating a single polypeptide chain. Hydrazinolysis, however, gave three amino acids: lysine, phenylalanine and alanine. No peptide was found that could give any evidence for alanine as a *C*-terminus. It has been shown that, despite repeated precipitation and dialysis, free amino acids may remain secondarily bound to proteins.[19] The amino acid composition of the hydrophobic peptide *T3*, which was purified by precipitation, always contained two alanines, but only one could be found in the sequence. This shows that free alanine is tightly bound to the peptide, which may be the reason for the alanine found by hydrazinolysis. Unresolved leghemoglobin *c* is therefore considered to have two *C*-terminal amino acids, lysine and phenylalanine.

Two alternative amino acids occur at six positions (8, 20, 39, 79, 97 and 143), and one of them is always identical with the corresponding amino acid residue in leghemoglobin *a* (Fig. 2). The tryptic peptide pairs containing the first five alternative amino acids are as a rule eluted in the same fractions from the ion exchange column, showing that no charge difference exists (Fig. 1). The only exception is peptide *T9A* with the *N*-terminus *Ala-Ser-Gly*, which is eluted in fraction III and peptide *T9B* with the *N*-terminus *Ala-Asp-Gly*, which is eluted in fraction VI. However, the thermolytic peptide *Th2* (Fig. 2), which contains both alternatives, has a net charge of +1, showing that in *T9B* the asparagine must be in the deamidated form. Deamidation of asparagine has been shown to occur easily in an *Asn-Gly* bond,[20] and it is clear that the glutamine of peptide *T2* can also be easily deamidated because about half of *T2* is found in fraction X and the rest in fraction XIII. In all five cases mentioned above the alternative amino acids could be due to a point mutation in the corresponding codon.[21] At position 143, however, which has lysine or phenylalanine, these codons differ in all three bases.[21] Lysine is the *C*-terminus of leghemoglobin *a*. In contrast, in the closely related kidney bean leghemoglobin *a* the corresponding amino acid is tyrosine,[6] which in hemoglobins has been shown to replace frequently phenylalanine.[22]

Comparison of the tryptic peptides from the unresolved leghemoglobin *c* with the chymotryptic peptides from leghemoglobin c_2 shows that the same microheterogeneity that occurs in the unresolved protein is also present in component c_2. The only difference found is in the *C*-terminus, which in leghemoglobin c_2 is phenylalanine (found also by Hurrell and Leach [10]), indicating that lysine must be the *C*-terminus of component c_1. Resolution of the two *c*-components on DEAE-cellulose is consequently due to the charge difference between lysine and phenylalanine.

The discrepancy between the sequence of leghemoglobin c_2 as reported by Hurrell and Leach [10] and that given by us necessitates a closer scrutiny of the results at the positions where variation occurs.

One possible reason for the deviations may be in the soybean varieties used in the two laboratories. In our laboratory soybean cv. *Fiskeby* was used whereas *Lincoln* was used in the laboratory of Hurrell and Leach. In a study on the evolutionary stability of cytochrome *c*-551 in different strains of *Pseudomonas aeruginosa* and *P. fluorescens* Ambler found that at a few positions in the cytochrome chains a variation of some amino acid residues occurred commonly.[23] Glutamic acid in one strain could be replaced by aspartic acid in another, glycine by serine or aspartic acid and valine by isoleucine. Some of the variations in the leghemoglobin *c* sequences, but not necessarily all, may thus depend on differences between the soybean varieties. Another reason for the deviations may depend on the reliability of the results obtained. Most of the differences in the leghemoglobin *c* sequences lie in regions, for which Hurrell and Leach determined the sequences using automatic sequence analyses. No tryptic or other type of peptides from these regions have been separately sequenced. This probably is a consequence of the small amounts of protein the Australian group used for the sequence study. Presumably they did not have enough material to check the results so that mistakes are possible.

The differences between the two sequences are collected in Table 1 and compared with the corresponding amino acid residues in other known leghemoglobin sequences. In the tryptic peptide *T1* we did not find any aspartic acid at position No. 5, whereas peptide *T2* as well as the chymotryptic peptides *C1* and *C2* all contained *Glu 2.5* and *Asp 0.5*. Dansyl-Edman degradation gave as residue 8 equal amounts of glutamic acid and aspartic acid. The unresolved peptide *T3* included *Thr 0.6*, *Asp 2.0* and *Ala 2.0*, and the dansyl-Edman degradation of the peptide resulted in both alanine and

Table 1. Differences between the amino acid sequences of soybean leghemoglobin *c* (*G*Lbc) as reported in this study (A) and by Hurrell and Leach (B), compared to the amino acid residues at corresponding positions in the sequences of soybean leghemoglobin *a* [4,5] (*G*Lb*a*), kidney bean leghemoglobin *a* [6] (*Ph*Lb*a*), broad bean leghemoglobin I [7] (*V*LbI), and lupin leghemoglobin 1 [8] (*L*LbI).

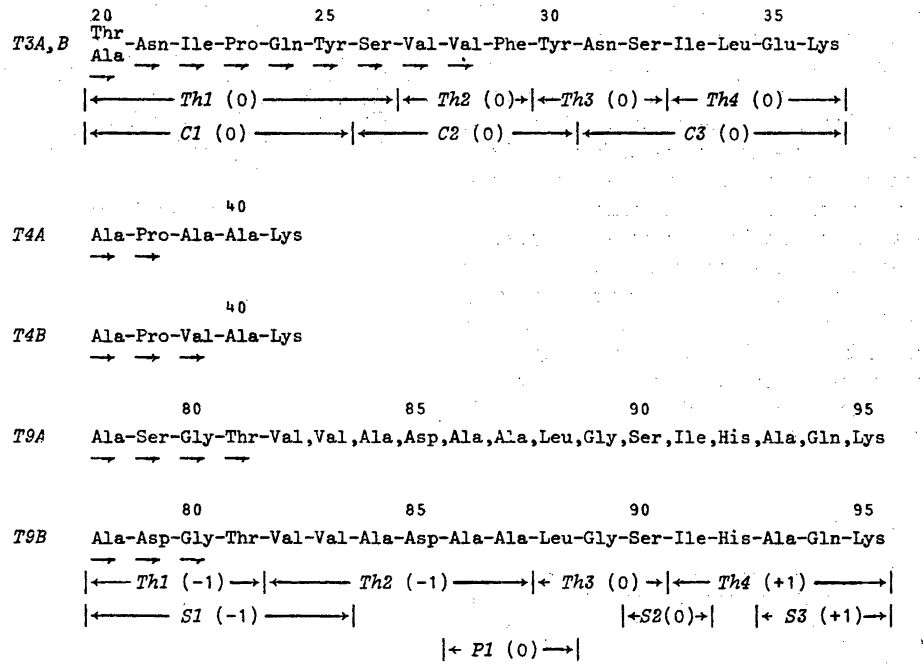
Tryptic peptide	Residue number	*G*Lbc A	B	*G*Lb*a*	*Ph*Lb*a*	*V*LbI	*L*LbI
T1	5	Glu	Glu/Asp	Glu	Glu	Glu/Asp	Asp
T2	8	Glu/Asp	Glu	Asp/Asn	Glu	Glu	Val
T3	20	Ala/Thr	Thr	Ala	Gly	Gln	Ala
	31	Asn	Thr	Thr	Thr	Thr	Thr
T4	38	Pro	Phe	Pro	Pro	Pro	Pro
	39	Ala/Val	Ala	Ala	Ala	Thr	Gly
	40	Ala	Val	Ala	Ala	Ala	Ala
T5	52	Asp	Asn	Asp	Asp	Gly	Glu
T7	67	Ala	Gly	Ala	Gly	Gly	Lys
T9	79	Ser/Asn	–	Ser	Asn	Thr	Asn
	80	Gly	–	Gly	Gly	Gly	Gly
	88	Leu	Ser	Leu	Leu	Asp	Leu
T10	97	Val/Ile	Ile	Val	Val	Val	Val
T16	143	Lys/Phe	Phe	Lys	Tyr	Ala	Ala

threonine as the *N*-terminal amino acid. Hydrolysis of *T3* with thermolysin and chymotrypsin gave the peptide fragments shown in Table 2. The amino acid composition of the thermolytic peptide *T3-Th3* is Tyr_1 Asp_1 Ser_1 with no trace of threonine. Without doubt the amino acid residue 31 in the sequence determined by us is asparagine, not the threonine found by Hurrell and Leach in the sequence of leghemoglobin c_3 and which occurs in all other leghemoglobins (Table 1). All the differences mentioned above may depend on differences in the varieties used. This may also be true for the deviations found at positions 52 (Asp⟷Asn) and 67 (Ala⟷Gly) (Table 1).

Most of the variations in the peptide *T4* and *T9* are more difficult to explain merely as differences between varieties. The amino acid composition of peptide *T4A* is Pro_1 Ala_3 Lys_1 and that of *T4B* is Pro_1 Ala_2 Val_1 Lys_1. No traces of phenylalanine could be found. Subtractive Edman degradation of the peptides gave the sequences shown in Table 2. Residue 38 is proline and residue 39 alanine or valine. In all other leghemoglobins residue 38 is proline (Table 1). X-Ray crystallographic studies on lupin leghemoglobin have shown that its secondary and tertiary structure is similar to that of animal myoglobin and hemoglobin chains.[24] In the hemoglobin and myoglobin sequences there is an invariable proline at the corresponding position (*cf.* Ref. 25). This proline (C2) forms a bend between helices B and C and an exchange of proline to phenylalanine would radically change the tertiary structure. It is therefore not likely that any other amino acid residue could replace proline at this position in the leghemoglobin *c* chain. The deviations at positions 39 – 40 may depend on differences in the varieties.

Peptide *T9* especially shows discrepancies between the sequences from the two laboratories (Table 1). The amino acid compositions of peptides *T9A* and *T9B* as obtained in our

Table 2. Amino acid sequences of tryptic peptides *T3A,B*, *T4A*, *T4B*, *T9A* and *T9B*. → Dansyl-Edman; ⇢ subtractive Edman. Peptide fragments obtained by hydrolysis with different enzymes are also given as well as, in parentheses, the net charge of the fragments at pH 6.5. All amino acid residues have been identified at least once as DNS-derivatives.



laboratory are as follows:
T9A Asp_1 Thr_1 Ser_2 Glu_1 Gly_2 Ala_4 Val_2 Ile_1 Leu_1 His_1 Lys_1
T9B Asp_2 Thr_1 Ser_1 Glu_1 Gly_2 Ala_4 Val_2 Ile_1 Leu_1 His_1 Lys_1

The fragments obtained by hydrolysis of the tryptic peptide *T9B* are shown in Table 2. The second amino acid in peptide *T9A* is serine and that in *T9B* is a deamidated asparagine as mentioned before. The dipeptide $\frac{Ser}{Asn}$-*Gly* (residues 79 – 80) is missing in the sequence of leghemoglobin c_2 as reported by Hurrell and Leach. In the other leghemoglobins residue 79 is serine, asparagine or threonine and residue 80 is always glycine (Table 1). At position 88 we found leucine instead of serine. The Australian group has not isolated the tryptic peptide *T9*, but this part of the sequence has been determined from automatic sequence analyses of a tryptic peptide (S – T1, residues (71 – 93) from the succinylated apoprotein. This region is not confirmed by any peptides from other hydrolysates and thus some mistakes may well have been made.

The amino acid sequence of leghemoglobin *c* from soybean cv. *Fiskeby* as reported by us differs completely at six positions from that of component *a* from the same variety. Only one point mutation is needed to change the codon for the differing amino acid residue in leghemoglobin *c* to the corresponding one in leghemoglobin *a*. The replacements at positions 114, 126 and 141 give a net charge difference of one between components *a* and *c*. which enables their separation on an ion exchanger.

The amino acid differences between leghemoglobin *a* and leghemoglobin *c* are at most twelve amino acids and at least six, depending on the alternative amino acid residues in leghemoglobin *c*. Because the components cannot yet be separated by conventional means it is impossible to say whether all of the possible varieties can be found.

REFERENCES

1. Ellfolk, N. *Endeavour 31* (1972) 139.
2. Ellfolk, N. *Acta Chem. Scand. 14* (1960) 609.
3. Appleby, C. A., Nicola, N. A., Hurrell, J. G. and Leach, S. J. *Biochemistry 14* (1975) 4444.
4. Ellfolk, N. and Sievers, G. *Acta Chem. Scand. 25* (1971) 3532.
5. Ellfolk, N. and Sievers, G. *Acta Chem. Scand. B 28* (1974) 1245.
6. Lehtovaara, P. and Ellfolk, N. *Eur. J. Biochem. 54* (1975) 577.
7. Richardson, M., Dilworth, M. J. and Scawen, M. D. *FEBS Lett. 51* (1975) 33.
8. Jegorov, C. A., Feigina, M. Iu., Kazakov, V. K., Shahparanov, M. J., Mitaleva, S. U. and Ovchinikov, Iu. A. *Bioorg. Khim. 2* (1976) 125.
9. Sievers, G., Huhtala. M.-L. and Ellfolk, N. *Acta Chem. Scand. B 31* (1977) 723.
10. Hurrell, J. G. R. and Leach, S. J. *FEBS Lett. 80* (1977) 23.
11. Sievers, G. and Ellfolk, N. *Acta Chem. Scand. 24* (1970) 439.
12. Ellfolk, N. *Acta Chem. Scand. 15* (1961) 545.
13. Ellfolk, N. and Sievers, G. *Acta Chem. Scand. 26* (1972) 1155.
14. Ellfolk, N. and Sievers, G. *Acta Chem. Scand. 27* (1973) 3986.
15. Ellfolk, N. and Sievers, G. *Acta Chem. Scand. 27* (1973) 3817.
16. Ellfolk, N. and Sievers, G. *Acta Chem. Scand. 27* (1973) 3371.
17. Ozols, J. *J. Biol. Chem. 245* (1970) 4863.
18. Offord, R. E. *Nature 211* (1966) 591.
19. Arens, A., Sund, H. and Wallenfels, K. *Biochem. Z. 337* (1963) 1.
20. Jörnwall, H. *FEBS Lett. 38* (1974) 329.
21. Crick, F. H. C. *J. Mol. Biol. 38* (1968) 367.
22. Vogel, F. *J. Mol. Evol. 1* (1972) 334.
23. Ambler, R. P. *Biochem. J. 137* (1974) 3.
24. Vainshtein, B. K., Harutyunyan, E. H., Kuranova, I. P., Borisov, V. V., Sosfenov, N. I., Pavlovsky, A. G., Grebenko, A. I. and Konareva, N. *Nature 254* (1975) 163.
25. Dayhoff, M. O., Ed., *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington 1972, Vol. 5, p. D-379.

Received February 14, 1978.