Biochem. J. (1974) **137**, 185–197
Printed in Great Britain

Triose Phosphate Isomerase from the Coelacanth

AN APPROACH TO THE RAPID DETERMINATION OF AN AMINO ACID SEQUENCE WITH SMALL AMOUNTS OF MATERIAL

By EDITH KOLB, J. IEUAN HARRIS and JOHN BRIDGEN
Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

(*Received* 13 *July* 1973)

The preparation and purification of cyanogen bromide fragments from [^{14}C]carboxymethylated coelacanth triose phosphate isomerase is presented. The automated sequencing of these fragments, the lysine-blocked tryptic peptides derived from them, and also of the intact protein, is described. Combination with results from manual sequence analysis has given the 247-residue amino acid sequence of coelacanth triose phosphate isomerase in 4 months, by using 100mg of enzyme. (Two small adjacent peptides were placed by homology with the rabbit enzyme.) Comparison of this sequence with that of the rabbit muscle enzyme shows that 207 (84%) of the residues are identical. This slow rate of evolutionary change (corresponding to two amino acid substitutions per 100 residues per 100 million years) is similar to that found for glyceraldehyde 3-phosphate dehydrogenase. The reliability of sequence information obtained by automated methods is discussed.

In determining the amino acid sequence of a protein the conventional approach has been to obtain small, usually tryptic, peptides that are relatively easy to purify and then to sequence these by one of several modifications of the phenylisothiocyanate (Edman, 1956) method. These are then overlapped by comparison with sequences of other, usually chymotryptic, peptic or thermolytic, peptides, and this approach continues to be the method of choice for proteins consisting of 100–200 amino acids (see, e.g., Ambler & Wynn, 1973; Shotton & Hartley, 1973). With longer protein chains such as glyceraldehyde 3-phosphate dehydrogenase (Davidson *et al.*, 1967; Harris & Perham, 1968), liver alcohol dehydrogenase (Jörnvall, 1970) and glutamate dehydrogenase (Landon *et al.*, 1971) the number of peptides produced by tryptic and other enzymic digestions increases to the point where their purification becomes progressively more difficult, tedious and time-consuming. Moreover a large amount of starting material is usually required to determine a complete sequence by this method. For these reasons sequence strategies based on the isolation of fewer but larger peptides obtained, for example, by specific tryptic cleavage at arginine residues after reversible blocking of lysine side chains (see, e.g., Butler *et al.*, 1969), and by cleavage at methionine residues with cyanogen bromide (Gross, 1967), have been used (see, e.g., Davidson *et al.*, 1967).

A disadvantage of this approach has been that mixtures of large peptides tend to form insoluble aggregates that are difficult to purify, but this difficulty has been largely overcome by the use of maleic anhydride (Butler *et al.*, 1969) or citraconic anhydride (Dixon & Perham, 1968) as dissociating and solubilizing reagents in peptide purification. Large peptides are also difficult to sequence by manual methods and hitherto it has usually been necessary to redigest cyanogen bromide and tryptic arginine fragments with trypsin and/or other enzymes so as to obtain smaller peptides (preferably in the range of 10–20 amino acids) that can be conveniently and reliably sequenced by manual methods such as dansyl–phenylisothiocyanate degradation (Gray, 1972). This involves additional fractionation steps that are both time-consuming and wasteful of material.

The development of automatic and quantitative sequencing methods (Edman & Begg, 1967) has now made it possible to obtain sequence information directly on proteins as well as on peptide fragments. Automated methods have thus been applied to compare *N*-terminal sequences in intact proteins (e.g. Niall, 1971; Hermodson *et al.*, 1972; Hood *et al.*, 1973; Bridgen *et al.*, 1973) and, for β-lactoglobulin, to determine the sequences of all the tryptic peptides comprising the protein chain of 162 residues (Braunitzer *et al.*, 1972). The ability to obtain the sequence of a protein entirely by automated methods is clearly an important development, but as applied to β-lactoglobulin the method still retains many of the disadvantages inherent in any approach requiring the purification of large numbers of small peptides from total tryptic (and other) digests.

We now present a more integrated and flexible approach to the problem of protein sequencing that

makes use of an automatic sequencer (Beckman 890B) to obtain sequence information directly on a representative selection of large peptide fragments as well as on the intact protein itself. Parts of the chain not readily amenable to automated sequencer analysis are sequenced as before by dansyl–phenylisothiocyanate degradation of smaller peptides produced, for example, by further tryptic cleavage, after unblocking, of citraconylated tryptic arginine peptides. It is likely, however, that peptides of this type could now be sequenced by automated 'solid phase' sequencing methods (see Laursen *et al.*, 1972; Terhorst *et al.*, 1973).

This strategy has been applied to determine the complete amino acid sequence of triose phosphate isomerase D-glyceraldehyde 3-phosphate ketol-isomerase, EC 5.3.1.1), a dimeric enzyme with a subunit consisting of about 250 amino acids. Our particular purpose has been to determine the amino acid sequence of triose phosphate isomerase from the coelacanth, the only known living survivor of the Crossopterygii, which are direct ancestors of present-day land vertebrates. Morphologically the coelacanth has remained unchanged for over 400 million years and it was of interest to determine the sequence of a coelacanth protein so that it could be compared with the sequence of the same protein from a mammalian source.

Triose phosphate isomerase from rabbit muscle (Miller & Waley, 1971; Corran & Waley, 1973) was known to contain only two methionine (and seven arginine) residues, and amino acid analysis of the coelacanth enzyme revealed that it too contained a similar number of these residues. Moreover comparison by gel electrophoresis in sodium dodecyl sulphate of the products formed by cyanogen bromide cleavage of the rabbit and coelacanth enzymes showed that the two methionine residues occupied similar positions in the two protein chains and that large peptide fragments suitable for automated analysis could be readily obtained by this means. This preliminary study suggested that the coelacanth enzyme, available only in small quantities, might be sequenced by the type of sequence strategy that we were wishing to evaluate. We now describe the experiments that have led to the determination of the sequence of the 247 amino acids that comprise the subunit of coelacanth triose phosphate isomerase in less than 4 months and by using only 100mg (4μmol of subunit) of the pure enzyme.

Materials

Triose phosphate isomerase was prepared from coelacanth (*Latimeria chalumnae*) muscle as previously described (Kolb & Harris, 1972).

Triose phosphate isomerase from rabbit muscle was obtained from Boehringer Corp. (London) Ltd., London W.5, U.K.

Trypsin ('toluene-*p*-sulphonyl-L-phenylalanine chloromethyl ketone-treated') and carboxypeptidases A and B ('di-isopropyl phosphorofluoridate-treated') were obtained from Worthington Biochemical Corp., Freehold, N.J., U.S.A.

Thermolysin was obtained from Chugai Boyaki Co. Ltd., Osaka, Japan.

Iodo[2-^{14}C]acetic acid (33mCi/mmol) was obtained from The Radiochemical Centre, Amersham, Bucks., U.K., and was diluted before use with a solution of carrier iodoacetic acid, which had been previously recrystallized from hexane to a final specific radioactivity of 1.0mCi/mmol.

Urea (Aristar grade), guanidine hydrochloride (Aristar grade) and citraconic anhydride (laboratory-reagent grade) were obtained from BDH Chemicals Ltd., Poole, Dorset, U.K.

Sequencer chemicals were obtained from Beckman Instruments Inc., Palo Alto, Calif., U.S.A.

4-Sulphonylphenyl isothiocyanate was obtained from Pearce Chemical Co., Rockford, Ill., U.S.A.

Experimental and Results

Carboxymethylation

The protein (100mg, 4μmol of subunit) at a concentration of 15mg/ml was reduced and denatured under N_2 in 50mM-Tris–HCl, pH 8.2, containing 6M-guanidine hydrochloride and 1mM-dithiothreitol, for 1h at room temperature. Iodo[2-^{14}C]acetate was then added in a twofold molar excess over total SH groups. After 2h the reagents were removed by dialysis against 0.1M-NH_3 solution. All subsequent experiments were carried out on the carboxymethylated enzyme.

Amino acid analysis

Protein or peptide samples for analysis were hydrolysed with 6M-HCl containing 0.1% (w/v) phenol in evacuated sealed tubes for 24h and, in some cases, 72h at 105°C. Quantitative analyses were performed with a Locarte amino acid analyser. The results obtained for the protein subunit and for the three CNBr fragments are given in Table 1. Apart from the low values for valine (24 for 28 in the protein subunit, and 19 for 22 in peptide CNBr III), the values obtained for the other amino acids are in good agreement with those obtained from the sequence results. The compositions of other peptides are given in Tables 2 and 3.

Protein sequence determination

The *N*-terminal sequence of the *S*-[^{14}C]carboxymethylated protein (8mg, 325nmol of subunit) was determined by automated Edman degradation in a

Table 1. *Amino acid compositions of triose phosphate isomerase and its cyanogen bromide fragments*

Protein or peptide samples were hydrolysed in evacuated glass tubes with 0.2ml of Aristar-grade HCl–water (1:1, v/v) containing 0.1% (w/v) phenol, for 24 and 72h. Cysteine was identified as *S*-carboxymethylcysteine. Tryptophan was determined in the whole protein only by the method of Edelhoch (1967). After cyanogen bromide digestion methionine was identified as homoserine.

	Composition (residues/molecule)							
	Enzyme subunit		Fragment CNBr I		Fragment CNBr II		Fragment CNBr III	
Amino acid	Analysis	Sequence	Analysis	Sequence	Analysis	Sequence	Analysis	Sequence
Carboxymethylcysteine	4.9	5	—	—	1.8	2	2.8	3
Aspartic acid	18.3	18	1.1	1	6.4	6	11.1	11
Threonine	11.0	11	—	—	2.9	3	7.9	8
Serine	12.8	13	—	—	2.8	3	9.7	10
Glutamic acid	29.5	29	—	—	5.9	6	22.5	23
Proline	9.1	10	0.9	1	4.7	5	3.7	4
Glycine	25.4	26	1.9	2	5.9	6	17.0	18
Alanine	22.4	23	0.9	1	8.6	9	11.8	13
Valine	23.8	28	1.0	1	4.8	5	19.3	22
Methionine	1.9	2	0.9	1	1.0	1	—	—
Isoleucine	11.2	12	—	—	2.9	3	8.9	9
Leucine	16.1	17	—	—	4.6	5	10.5	12
Tyrosine	4.6	5	—	—	1.9	2	2.5	3
Phenylalanine	9.7	10	1.9	2	3.9	4	4.2	4
Lysine	21.0	21	2.0	2	7.0	7	12.4	12
Histidine	3.9	4	—	—	—	—	3.5	4
Arginine	8.0	8	1.0	1	1.0	1	5.4	6
Tryptophan	4.5	5		1	—	—		4
Total residues		247		13		68		166

Beckman 890B sequencer. The standard double-cleavage with Quadrol [1.0M-*NNN'N'*-tetrakis-(2-hydroxypropyl)ethylenediaminetrifluoroacetate, pH 9.0] was used and the released phenylthiohydantoin-amino acids were identified and quantified as described by Bridgen & Secher (1973). The first 13 amino acids were obtained in good yield, but at cycle 14 a very low recovery (10% of expected) of aspartic acid was found and no further sequence information could be obtained. However, a second sequencer analysis of the protein under identical conditions but with a single very short acid cleavage at cycle 13 allowed the determination of the first 25 residues (Fig. 5), albeit with a 60% fall in yield at asparagine-15. The presence of the methionine residue at position 14 indicated that only one other methionine residue remained in the protein chain, and CNBr was therefore chosen as the primary cleavage method.

CNBr cleavage and separation of fragments of [^{14}C]*-carboxymethylated enzyme*

The protein (35mg) was dissolved in 98% (v/v) formic acid (3ml). The solution was diluted with water to 70% (v/v) formic acid, and CNBr in a two- to three-fold excess (w/w) was then added. The mixture was left at room temperature in the dark for 24h and then freeze-dried. In a preliminary small-scale experiment an attempt was made to separate the peptide mixture by gel filtration on Sephadex G-75. The sample was initially applied in 50% (v/v) formic acid and the column eluted with 5% (v/v) formic acid. This method was not successful; the larger two fragments were poorly separated from each other (as well as from partially cleaved material) and only the smaller *N*-terminal fragment (residues 1 to 14, Fig. 5) was obtained pure. Similar difficulties appear to have been encountered by Miller & Waley (1971) working with the rabbit muscle enzyme. Sodium dodecyl sulphate–polyacrylamide-gel electrophoresis showed that fragments of similar size were obtained by CNBr cleavage of the rabbit and coelacanth proteins, and in order to conserve the coelacanth enzyme alternative methods of fractionation were sought, by using commercially available rabbit enzyme. The main difficulty was the insolubility of the mixture, and a method that made use of a gradient of guanidine hydrochloride, successful in trial experiments with the rabbit enzyme, was then applied to the digest of the coelacanth enzyme. The freeze-dried mixture of peptides was dissolved in 0.2M-NH_4HCO_3, pH 8.5, containing 8M-guanidine hydrochloride and then diluted with 3 vol. of 0.2M-NH_4HCO_3.

Gel filtration was carried out on a column (2cm × 100cm), the lower two-thirds of which was filled with Sephadex G-75 and the upper third with Sephadex G-100, in 0.2M-NH_4HCO_3. Before application of the sample, seven 10ml batches of 0.2M-NH_4HCO_3 containing increasing concentrations of guanidine hydrochloride (0.4, 0.8, 1.2, 1.6, 2.0, 2.4 and 2.7M respectively) were applied to the column. The peptide mixture was applied in 3ml of 2.7M-guanidine hydrochloride in 0.2M-NH_4HCO_3 and the same buffer without guanidine hydrochloride was used for subsequent elution. Fig. 1 shows the elution profile, the fragments being numbered according to their positions in the protein chain starting at the *N*-terminus. The purity of fractions 28–54 was assessed by sodium dodecyl sulphate–polyacrylamide-gel electrophoresis (Plate 1) in the discontinuous Tris–glycine system with 0.1% (w/v) sodium dodecyl sulphate as described by Lämmli (1970). The gels contained 15% (w/v) acrylamide and 0.4% (w/v) methylenebisacrylamide. Before being run the samples were boiled for 1min with 2% (w/v) sodium dodecyl sulphate and 5% (v/v) β-mercaptoethanol. Fractions 34–40 (Fig. 1) contained pure fragment III and fractions 44–54 contained pure fragment II. Fragment I (fractions 75–80), whose sequence was already known from the sequencer results with the intact protein, was eluted with the guanidine hydrochloride. The peak directly preceding fragment 1 contained the majority of the salt. No additional peptide material was found in this peak. In fractions 29–33 fragment III was contaminated with partially cleaved material, which was eliminated by re-running the mixture on Sephadex G-75 in 0.2M-NH_4HCO_3 without guanidine hydrochloride. The sodium dodecyl sulphate–polyacrylamide-gel results indicated that the material with a molecular weight close to that of the intact protein consisted of fragments II plus III. This arose as the result of incomplete (70%) cleavage at the Met–Ile bond (residues 82–83, Fig. 5) and even after additional treatment with CNBr this bond remained intact in about 30% of the molecules.

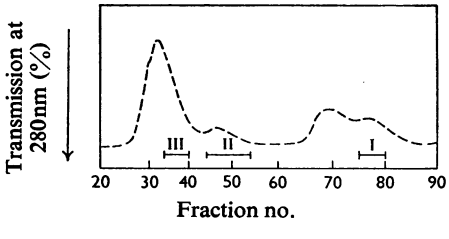


Fig. 1. *Gel filtration on Sephadex G-75/G-100 of CNBr fragments prepared from coelacanth triose phosphate isomerase*

For details see the text. Elution was with 0.02M-NH_4HCO_3 at a flow rate of 40ml/h and fractions (4ml) were collected. I, II and III refer to pure CNBr fragments (cf. Fig. 2). ----, % transmission, (LKB Uvicord).

Automated sequence determination of CNBr fragments

The sequence of fragment CNBr I (residues 1–14, Fig. 6) was known from sequencer analysis of the whole protein. Fragment CNBr II was examined in the sequencer but was found to be 90% blocked, presumably owing to the known propensity of Asn-Gly sequences to rearrange to β-imides (Ambler, 1963; Konigsberg, 1967; Bornstein & Balian, 1970), and only the partial sequence Asn-Gly-Asp-Lys-Lys-?-Leu-Gly-Glu-Leu-Ile-?-Thr-Leu-Asx-Ala could be qualitatively obtained. Fragment CNBr III (400nmol) was, however, successfully degraded through 69 residues (83–151, Fig. 5) by using the standard protein programme (Fig. 2). Residues 139, 147 and 148 could not be unambiguously identified, and these were confirmed as threonine, threonine and glutamic acid respectively by sequencer analysis of the overlapping fragment III T_3 (residues 135–189).

Digestion of carboxymethylated protein and CNBr fragments II and III with carboxypeptidases

The protein and fragments CNBr II and III (1–2mg/ml in 0.2M-NH_4HCO_3) were incubated at room temperature with 1% (w/v) carboxypeptidase A, and after 10min samples (10nmol) of each digest were removed and frozen. An equivalent amount of carboxypeptidase B was then added and further 10nmol samples were withdrawn at intervals of up

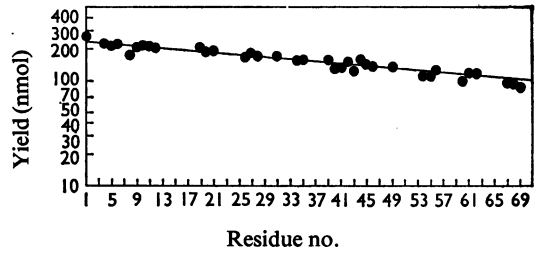
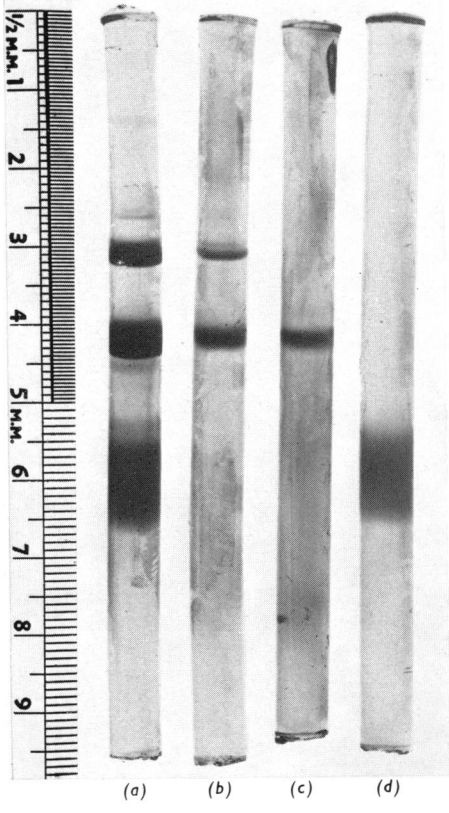


Fig. 2. *Semilogarithmic plot of absolute yield in nmol against residue number for CNBr fragment III*

Yields were calculated from g.l.c. peak heights and only those residues quantifiable by this method (alanine, glycine, valine, leucine, isoleucine, methionine, phenylalanine, tyrosine, tryptophan) are indicated. *S*-[^{14}C]-Carboxymethylcysteine was quantitatively identified at positions 4 and 44 in the fragment by scintillation counting of 10% (v/v) of each fraction. A repetitive yield of 97% was calculated from the slope of the graph. Minimum signal/'noise' ratio was 2:1. For further details see the text.



EXPLANATION OF PLATE 1

Polyacrylamide-gel electrophoresis in 0.1% (*w*/*v*) *sodium dodecyl sulphate of CNBr fragments separated by gel filtration on Sephadex G-75*/*G-100* (*Fig.* 1)

(*a*) Unfractionated mixture, (*b*) fragments CNBr III and CNBr (II+III) (fractions 29–33), (*c*) fragments CNBr III (fractions 34–40), (*d*) fragments CNBr II (fractions 44–54). For further details see the text.

to 1h and submitted to amino acid analysis. The results showed that fragment CNBr III was the *C*-terminal fragment in the protein and that the *C*-terminal sequence was -Val-Arg(Thr or Gln). The *C*-terminal residue was shown to be glutamine (identified as glutamic acid after acid hydrolysis). Fragment CNBr II gave homoserine only.

Citraconylation of CNBr fragments

Peptides CNBr II and CNBr III (600nmol) in 0.2M-sodium borate (3ml) containing 8M-urea at pH 8.5 were treated with citraconic anhydride (Dixon & Perham, 1968) added in 20μl portions during 30min to give a 40-fold excess over peptide amino groups. A pH of 8.3–8.6 was maintained by addition of 2M-NaOH and when the reaction was complete the citraconylated peptides were desalted on Sephadex G-25 in 0.2M-NH_4HCO_3.

Separation and identification of tryptic arginine peptides from fragment CNBr II

Citraconylated fragment CNBr II (600nmol) was digested with trypsin (1:80, w/w) in 0.2M-NH_4HCO_3 (2.5ml) for 4h at 37°C, and the trypsin was then inactivated by addition of di-isopropyl phosphorofluoridate to a final concentration of 0.1mM. Citraconyl groups were removed by incubation in 5% (v/v) formic acid for 8h at room temperature and, after freeze-drying, the peptide mixture was redissolved in 5mM-NH_4HCO_3 containing 1M-urea and applied to a column of DEAE-cellulose (20ml settled volume). The column was eluted with 40ml of 5mM-NH_4HCO_3 (pH 8.0) followed by a linear gradient of NH_4HCO_3, and the eluate was monitored for absorbance at 280nm and ^{14}C radioactivity. As fragment CNBr II contained only one arginine residue, two peptides were expected and, as shown in Fig. 3, two radioactive fractions (showing that both contained [^{14}C]carboxymethylcysteine) were obtained. The material in fractions 43–46 (fragment II T_2) gave *N*-terminal leucine and contained 30 amino acids, but no arginine, and the material in fractions 61–65 gave *N*-terminal aspartic acid and contained 38 amino acids including one arginine residue, showing that it was the *N*-terminal part of fragment CNBr II. These two peptides (II T_1 and II T_2, Table 2), each recovered in a yield of 30–40%, account for all the amino acids in fragment CNBr II.

Sequence analysis of peptides II T_1 and II T_2

The two peptides were each digested with trypsin (1:80, w/w) in 0.2M-NH_4HCO_3 for 4h at 37°C and the digests were fractionated on Whatman 3MM or no. 1 paper by high-voltage electrophoresis at pH 6.5 and 2.1. Peptides were stained with ninhydrin–Cd^{2+}

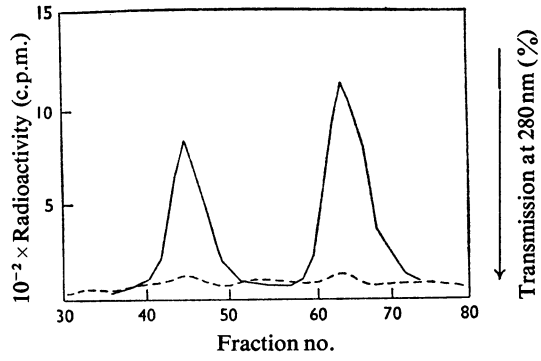


Fig. 3. *Chromatography on DEAE-cellulose of tryptic arginine fragments prepared from fragment CNBr II*

Elution was with a linear gradient (5–500mM, 100ml of each) of NH_4HCO_3. Fractions (2.5ml) were collected and 40μl samples were analysed for ^{14}C (——). - - - -, % transmission (LKB Uvicord).

(Heilmann *et al.*, 1957) and mobilities at pH 6.5 (relative to aspartic acid) were used to calculate net charges (Offord, 1966). Radioactive peptides were located by radioautography and electrophoretograms were also stained for tryptophan by using the Ehrlich reagent (Smith, 1953). Elution from the electrophoretograms was with 0.1M-NH_3. *N*-Terminal amino acids were identified by the dansyl method (Gray, 1972) and peptide sequences were determined by dansyl–Edman degradation as described by Gray (1972) and by Hartley (1970).

Peptide II T_1. This contained three lysine and one arginine residue (Table 2) and its *N*-terminal sequence, determined by sequencer analysis, was already known to be Asn-Gly-Asp-Lys-Lys-?-Leu-Gly-Glu-Leu-Ile-?-Thr-Leu-Asx-Ala. The trypsin digest gave three major peptides (II T_1a, II T_1b and II T_1c, Table 2) by electrophoresis at pH 6.5 and these contained 5, 13 and 20 amino acids respectively. The basic peptide (II T_1a) was the *N*-terminal part of peptide II T_1, the neutral peptide (II T_1b) with *N*-terminal serine occupied the middle position, and the acidic arginine-containing radioactive peptide II T_1c was *C*-terminal. Peptides II T_1b and II T_1c (available in 100–150nmol amounts) were sequenced manually by dansyl–Edman degradation, and residues not determined by sequencer analysis were thus identified.

Peptide II T_2. This contained four lysine residues and its *N*-terminal sequence was established to be Leu-Lys-Val-Asp-Pro-Lys- by dansyl–Edman degradation. Five peptides (II T_2a–II T_2e, Table 2) were isolated by electrophoresis at pH 6.5 and 2.1 after digestion with trypsin (conditions as for peptide II T_1) and sequenced by dansyl–Edman degradation. Peptides II T_2a and II T_2b could thus be shown to be

Table 2. *Amino acid compositions and electrophoretic mobilities of tryptic and thermolytic fragments from fragment CNBr II*

The letters *a*–*e* in columns 4–11 refer to tryptic lysine fragments, aligned from the *N*-terminus of the parent arginine peptides II T_1 and II T_2 respectively. The compositions predicted from the sequence results are shown in parentheses.

	Composition (residues/molecule)									
Peptide ..	II T_1	II T_2	II T_1			II T_2				
			a	*b*	*c*	*a*	*b*	*c*	*d*	*e*
Residue numbers	15–52	53–82								
Mobility at pH 6.5 (Asp = −1.0) ..			+0.37	0	−0.37	+0.65	0	0	+0.55	−0.25
Carboxymethyl-cysteine	0.9 (1)	0.9 (1)			0.8 (1)			0.9 (1)		
Aspartic acid	4.1 (4)	2.2 (2)	2.0 (2)	1.1 (1)	1.2 (1)		1.0 (1)	1.0 (1)		
Threonine	1.8 (2)	0.8 (1)		0.9 (1)	0.8 (1)					0.9 (1)
Serine	0.9 (1)	1.8 (2)		0.8 (1)					0.8 (1)	0.8 (1)
Glutamic acid	4.0 (4)	2.1 (2)		2.0 (2)	2.2 (2)			1.1 (1)		1.0 (1)
Proline	2.8 (3)	1.7 (2)			2.9 (3)		0.8 (1)			0.9 (1)
Glycine	3.5 (3)	3.2 (3)	1.3 (1)	1.1 (1)	1.1 (1)			1.2 (1)		2.2 (2)
Alanine	5.1 (5)	3.9 (4)		2.0 (2)	2.8 (3)			2.0 (2)		2.0 (2)
Valine	1.9 (2)	3.1 (3)			1.9 (2)		1.1 (1)	0.9 (1)	1.1 (1)	
Homoserine		1.0 (1)								0.8 (1)
Isoleucine	2.0 (2)	1.1 (1)		0.9 (1)	0.9 (1)					0.9 (1)
Leucine	3.5 (4)	0.9 (1)		2.5 (3)	0.8 (1)	1.0 (1)				
Tyrosine	0.9 (1)	0.8 (1)			1.0 (1)			0.9 (1)		
Phenylalanine	2.1 (2)	2.0 (2)			1.9 (2)			1.1 (1)		0.9 (1)
Lysine	3.0 (3)	4.0 (4)	2.1 (2)	1.1 (1)		1.0 (1)	1.0 (1)	1.0 (1)	1.0 (1)	
Arginine	1.0 (1)				1.0 (1)					
Total	38	30	5	13	20	2	4	10	3	11

N-terminal, and II T_2e, with *C*-terminal homoserine, to be *C*-terminal in peptide II T_2. The decapeptide II T_2c and the tripeptide II T_2d were not overlapped, and these fragments were positioned by reference to the sequence of the rabbit muscle enzyme (Corran & Waley, 1973).

These results establish the complete sequence of the 68 residues in fragment CNBr II.

Separation and identification of tryptic arginine fragments from fragment CNBr III

Citraconylated fragment CNBr III was digested with trypsin (1:80, w/w) in 0.2M-NH_4HCO_3 for 4h at 37°C. The digest was fractionated on a column (2.5cm × 100cm) of Sephadex G-50 and eluted with 0.2M-NH_4HCO_3 (Fig. 4). End groups of relevant

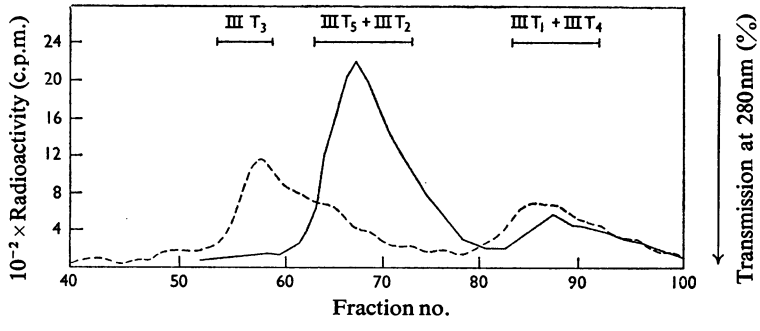


Fig. 4. *Gel filtration on Sephadex G-50 of tryptic arginine fragments prepared from fragments CNBr III*

Elution was with 0.2M-NH_4HCO_3 at 30ml/h. Fractions (3ml) were collected and 40μl samples were analysed for ^{14}C (——). – – – –, % transmission (LKB Uvicord).

fractions were determined by dansylation after removal of citraconyl groups. Fractions 54–58 (peptide III T_3) with a single end group, glutamic acid, were pooled and subjected to amino acid analysis (Table 3). This peptide was subsequently completely sequenced in the sequencer (see below). End-group analysis of fractions 63–73 indicated the presence of two partially resolved peptides with *N*-terminal isoleucine and histidine respectively. These were further resolved by re-running on the same Sephadex G-50 column. Sequencer analysis (see below) showed that the components were peptide III T_2, whose sequence was known, and a second peptide (III T_5) whose structure was determined by mixture analysis. Fractions 83–92 also contained two peptides, with *N*-terminal lysine and isoleucine. Electrophoresis of this mixture at pH 6.5 gave a neutral, fluorescent, tryptophan-containing peptide (III T_4) with *N*-terminal lysine. The second component was virtually insoluble in pH 6.5 buffer and was identified as peptide III T_1 by comparison with the sequencer results from fragment CNBr III. Peptide III T_4 was shown to contain 16 amino acids (Table 3) and its sequence was determined by dansyl–Edman degradation, except that tryptophan was not positively identified as the second residue, and that amide groups were not assigned. Digestion with thermolysin [2% (w/w) in 0.2M-ammonium acetate containing 5mM-$CaCl_2$, pH 8.5, for 3h at 37°C] followed by two-dimensional electrophoresis at pH 6.5 and 2.1 gave four major peptide fragments, whose electrophoretic mobilities and sequences are given in Table 3. Peptide III T_4a was a basic dipeptide containing *N*-terminal lysine and tryptophan, confirming that the latter was the second residue in the sequence. The electrophoretic mobilities of peptides III T_4c and III T_4d indicated that amide groups were absent, and the third and fourth residues in the neutral peptide III T_4b were identified as glutamic acid and asparagine respectively by g.l.c. of the phenylthiohydantoin derivatives. Fractions 113–116 (peptide III T_6) contained free glutamine only.

The position of peptides III T_1, III T_2 and III T_3 in the sequence were defined by comparison with the sequencer result from CNBr fragment III (Fig. 5) whereas peptides III T_6 and III T_5 were positioned from the carboxypeptidase results. This resulted in a unique position for peptide III T_4. This order was confirmed by homology with the rabbit muscle enzyme (Fig. 6).

Automated sequencing of tryptic arginine fragments

The fragments that remained to be examined in this way were the tryptic peptides III T_3 and III T_5.

Table 3. *Amino acid compositions and electrophoretic mobilities of tryptic arginine and thermolytic fragments from fragment CNBr III*

The letters *a–d* in columns 4–7 are the fragments produced by thermolysin digestion of tryptic arginine peptide III T_4. The compositions calculated from the sequence results are shown in parentheses.

	Composition (residues/molecule)					
Peptide	III T_3	III T_4	III T_4			
			a	*b*	*c*	*d*
Residue numbers	135–189	190–205				
Mobility at pH 6.5 (Asp = −1.0)	..	0	+0.52	0	−0.47	−0.54
Carboxymethylcysteine	—	—				
Aspartic acid	2.0 (3)	2.0 (2)		1.0 (1)		1.0 (1)
Threonine	3.7 (4)	0.9 (1)			0.9 (1)	
Serine	2.6 (3)	2.1 (2)			0.7 (1)	1.2 (1)
Glutamic acid	8.8 (9)	2.1 (2)		1.1 (1)	1.0 (1)	
Proline	2.2 (2)					
Glycine	4.9 (5)					
Alanine	4.9 (5)	1.1 (1)				0.9 (1)
Valine	6.2 (8)	3.1 (3)			1.1 (1)	1.1 (1)
Isoleucine	2.4 (3)					
Leucine	2.5 (3)	1.2 (1)		0.9 (1)		
Tyrosine	0.9 (1)					
Phenylalanine	0.9 (1)					
Tryptophan	(2)	(1)				
Lysine	4.0 (4)	2.0 (2)	1.0 (1)	1.0 (1)		
Histidine	(1)					
Arginine	1.0 (1)	1.1 (1)				
Total	55	16	1	4	4	4

Because of their length, 55 and 43 residues respectively, these peptides were dissolved in 1.0M-dimethylallylaminetrifluoroacetate buffer, pH9.0 (300μl), and treated with 4-sulphonylphenylisothiocyanate (20×molar excess over amino groups, at 45°C, for 2h) in screw-top tubes under N_2. This procedure increases the polarity of the peptide and minimizes losses during solvent extractions (Braunitzer *et al.*, 1971). The derivatized peptides were introduced directly into the sequencer cup and dried under vacuum. The strategy used for the degradations was based on that described by Niall (1971) whereby the polarity of the residual peptide is continually monitored and the programme varied accordingly so as to minimize extractive losses. By these principles the 1.0M-Quadrol buffer was first replaced by 0.2M-Quadrol (1M-Quadrol–propan-1-ol–water, 5:12:8, by vol.) and then by the volatile 1.0M-dimethylallylaminetrifluoroacetate, pH9.0, so that the ethyl acetate wash could be omitted completely. To facilitate removal of the volatile buffer the sequencer was modified to allow a continuous stream of N_2 through the cup on all high-vacuum steps. Details of the programmes are shown in Table 4. In an initial run on peptide III T_3 the dilute-Quadrol programme was used beyond residue 17 and the peptide then washed out of the cup. Peptide III T_5 was contaminated with peptide III T_2. Fortunately the latter peptide was washed from the cup faster than fragment III T_5 and its sequence could not be identified after about 15 cycles of degradation. In the four positions (residue numbers 209, 211, 218 and 239) where the proposed identification coincided with that which would have been expected from peptide II T_2, the absolute yield of phenylthiohydantoin at that step was calculated (Bridgen & Secher, 1973). Correlation with the phenylthiohydantoin yields at steps immediately before and after these positions showed no abnormally low yields.

By these procedures peptide III T_3 (55 residues) was sequenced through to the *C*-terminal arginine residue, and peptide III T_5 (42 residues) was sequenced to within two residues of the *C*-terminal arginine. Lysine residues at positions 187 and 244 could not be unambiguously identified, but these assignments agree with the amino acid analyses of these fragments.

This completed the sequence analysis of the 166-residue fragment CNBr III and, since the order of the three CNBr fragments was already known, the complete 247-residue sequence of triose phosphate isomerase from coelacanth muscle could be established (Fig. 5).

Quantity of material used

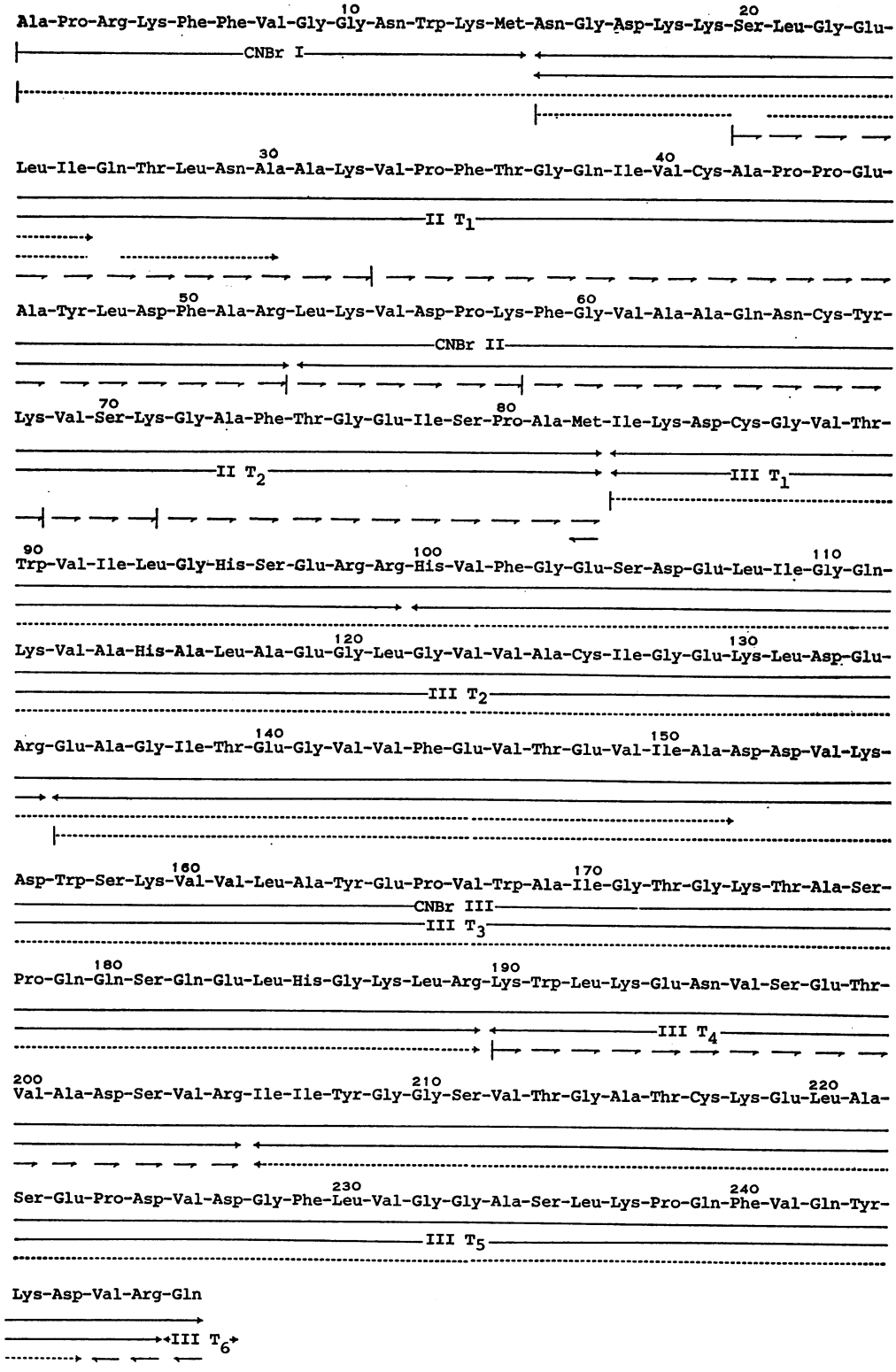
Carboxymethylated triose phosphate isomerase was submitted to cleavage with CNBr in batches of about 30–35mg (1.0–1.5μmol of subunit) and the recovery of pure fragments after gel filtration on Sephadex G-75/G-100 was about 40%. Citraconylation of CNBr fragments II and III was carried out with 500–600nmol of peptide, and tryptic arginine fragments were recovered in yields of 30–70% after chromatography on DEAE-cellulose and gel filtration on Sephadex G-50. Sequencer analysis of purified CNBr and tryptic arginine peptides was

Table 4. *Coupling buffers and solvent-wash times used for automated degradation of peptides III T_3 and III T_5*

Peptide	Residue no.	Buffer	Benzene wash time(s)	Ethyl acetate wash time(s)	No. of cleavage reactions per step
III T_3 (300 nmol)	1–10	1M-Quadrol	300	600	2
	11–17	0.2M-Quadrol	300	350	1
	18–50	Dimethylallylamine	300	—	1
	51–53	Dimethylallylamine	150	—	1
	54–55	Dimethylallylamine	—	—	1
III T_5 (200 nmol)	1–11	1M-Quadrol	300	600	2
	12–19	0.2M-Quadrol	300	350	1
	20–38	Dimethylallylamine	300	—	1
	39–41	Dimethylallylamine	—	—	1

Fig. 5. *Amino acid sequence of triose phosphate isomerase from coelacanth muscle*

The CNBr and tryptic arginine fragments are shown as solid lines. Residues identified by sequencer analysis (- - - -), by dansyl–Edman degradation (——›), by carboxypeptidase digestion (‹——). Serine at position 20 and glutamine at position 26 could not be identified during sequencer analysis of fragment CNBr II. Overlaps were not obtained between residues 58 and 59, residues 68 and 69 and residues 71 and 72 (see the text). Residues are numbered as for the rabbit muscle enzyme (Corran & Waley, 1973) and include a deletion at position 3 in the coelacanth enzyme.


Ala-Pro-Arg-Lys-Phe-Phe-Val-Gly-Gly-Asn-Trp-Lys-Met-Asn-Gly-Asp-Lys-Lys-Ser-Leu-Gly-Glu-
CNBr I
Leu-Ile-Gln-Thr-Leu-Asn-Ala-Ala-Lys-Val-Pro-Phe-Thr-Gly-Gln-Ile-Val-Cys-Ala-Pro-Pro-Glu-
II T1
Ala-Tyr-Leu-Asp-Phe-Ala-Arg-Leu-Lys-Val-Asp-Pro-Lys-Phe-Gly-Val-Ala-Ala-Gln-Asn-Cys-Tyr-
CNBr II
Lys-Val-Ser-Lys-Gly-Ala-Phe-Thr-Gly-Glu-Ile-Ser-Pro-Ala-Met-Ile-Lys-Asp-Cys-Gly-Val-Thr-
II T2
III T1
Trp-Val-Ile-Leu-Gly-His-Ser-Glu-Arg-Arg-His-Val-Phe-Gly-Glu-Ser-Asp-Glu-Leu-Ile-Gly-Gln-
Lys-Val-Ala-His-Ala-Leu-Ala-Glu-Gly-Leu-Gly-Val-Val-Ala-Cys-Ile-Gly-Glu-Lys-Leu-Asp-Glu-
III T2
Arg-Glu-Ala-Gly-Ile-Thr-Glu-Gly-Val-Val-Phe-Glu-Val-Thr-Glu-Val-Ile-Ala-Asp-Asp-Val-Lys-
Asp-Trp-Ser-Lys-Val-Val-Leu-Ala-Tyr-Glu-Pro-Val-Trp-Ala-Ile-Gly-Thr-Gly-Lys-Thr-Ala-Ser-
CNBr III
III T3
Pro-Gln-Gln-Ser-Gln-Glu-Leu-His-Gly-Lys-Leu-Arg-Lys-Trp-Leu-Lys-Glu-Asn-Val-Ser-Glu-Thr-
III T4
Val-Ala-Asp-Ser-Val-Arg-Ile-Ile-Tyr-Gly-Gly-Ser-Val-Thr-Gly-Ala-Thr-Cys-Lys-Glu-Leu-Ala-
Ser-Glu-Pro-Asp-Val-Asp-Gly-Phe-Leu-Val-Gly-Gly-Ala-Ser-Leu-Lys-Pro-Gln-Phe-Val-Gln-Tyr-
III T5
Lys-Asp-Val-Arg-Gln
III T6

carried out with 200–400nmol of peptide. Yields of tryptic lysine peptides purified by paper electrophoresis after redigestion of unblocked arginine peptides were 30–50% and the sequences of these peptides, determined by manual dansyl–Edman degradation involving up to 20 successive degradative cycles, was accomplished with 100–150nmol of peptide. Recoveries are based on the quantitative amino acid analysis of purified peptides, and, for [^{14}C]carboxymethylcysteine-containing peptides, on radioactivity.

Discussion

In commencing this work we were faced with the problem that only 100mg of pure coelacanth enzyme was available. Moreover at the outset the extent to which it was likely to be homologous with the rabbit enzyme was not known. The results of amino acid analysis and of cleavage with CNBr on a small scale soon indicated that key residues, such as methionine, had been conserved in approximately homologous positions and that the strategy that we wished to employ for the sequence analysis could therefore be tested in the first instance on commercially available rabbit muscle enzyme. This preliminary work showed that the three fragments produced by CNBr at the two methionine residues (cf. Miller & Waley, 1971) could be purified adequately by gel filtration on Sephadex and that tryptic arginine fragments prepared from citraconylated CNBr fragments could also be purified in satisfactory yields. Sequencer analysis of the carboxymethylated coelacanth protein indicated a methionine residue at position 13 (position 14 in the rabbit enzyme), and this confirmed the choice of CNBr as the initial method of cleavage since, with only one other methionine residue in the chain, the three primary fragments could be readily aligned from the *N*-terminus. Tryptic digestion of lysine-blocked peptides was chosen as the second cleavage method because peptides with *C*-terminal arginine retain a positive charge under the conditions of the sequencer analysis and are therefore less likely to be washed out of the reaction cup during the degradation procedures. Moreover in this case citraconylation was preferred to maleylation because the necessary arginine peptides could be purified either on Sephadex, or, in the unblocked form, on DEAE-cellulose, so that removal of the blocking group under milder conditions was an advantage.

Extended sequencer analysis of both the intact protein and of CNBr fragment II was frustrated by the presence of the Asn-Gly sequence (residues 15–16). Owing to a presumed rearrangement from an α- to a β-peptide linkage, and the consequent fall in yield, only limited information could be obtained with these materials, and this type of rearrangement as well as the tendency of some *N*-terminal glutamine residues to cyclize to pyrrolid-2-one-5-carboxylic acid, could be a major problem in automated (as well as in manual) sequence analysis. The use of a single, very short acid cleavage step at the preceding residue alleviates but does not completely solve this problem. The degradation of peptide CNBr III through 69 residues was achieved by use of very high-purity material. This fragment of 166 residues was of an optimal length and contained a sufficient number of charged residues to minimize extractive losses, but not so long that the phenylthiohydantoin-amino acid background, which presumably arises by non-specific cleavage along the polypeptide chain, was sufficiently extensive to create anomalies in the identification.

The major problem with the smaller arginine fragments was programme design. To obtain complete sequences on peptides of this length (around 50 residues) it is necessary to strike a constant balance between degradative efficiency and the times of the solvent washes that will remove peptide material from the cup. Complete sequence determination can only be achieved by continuous monitoring of the run and alteration of the programme according to the number of charged and hydrophobic residues remaining in the chain. In this we were helped considerably by the receipt from Dr S. G. Waley of the then unpublished sequence for the rabbit muscle enzyme (Corran & Waley, 1973). Comparison of the amino acid compositions of coelacanth peptides with sequences of corresponding parts of the rabbit chain allowed us to predict the charge distribution along the chain and hence avoid preliminary degradative experiments or the time-consuming examination of every phenylthiohydantoin-amino acid fraction as it emerged from the machine. Thus only one fragment, III T_3, was prematurely washed from the cup. Determination of the sequence of peptide III T_5 by mixture analysis of fragments III T_2 and III T_5 proved easier than expected, particularly when the hydriodic acid-hydrolysis results were correlated with the g.l.c. data. Apparently the contaminating fragment III T_2 was being washed out faster than the major fragment III T_5. It is doubtful, however, whether this technique could be applied to mixtures of unknown peptides beyond about ten residues.

By these methods about 75% of the sequence of coelacanth triose phosphate isomerase was obtained by automated methods. The remaining 25% was determined by manual techniques, and the complete sequence of 247 residues could then be assembled. The reliability of any sequence determined by automated means will primarily depend on the reliability of the phenylthiohydantoin-amino acid identification at each step. G.l.c., an extremely sensitive detection method, is capable of detecting fractions of a nmol of each derivative. However, conventional gas–liquid chromatographs use flame-ionization detectors, where the peak height is approximately proportional

to the size of the side chain, and so residues such as glycine, proline and alanine give only a poor response. Serine and threonine suffer from the added disadvantage that their phenylthiohydantoin derivatives are unstable, readily giving rise to dehydro and oxidized forms. T.l.c. suffers from not being quantitative and can generally only be used for the first 20–30 residues of each sequence determination. Re-

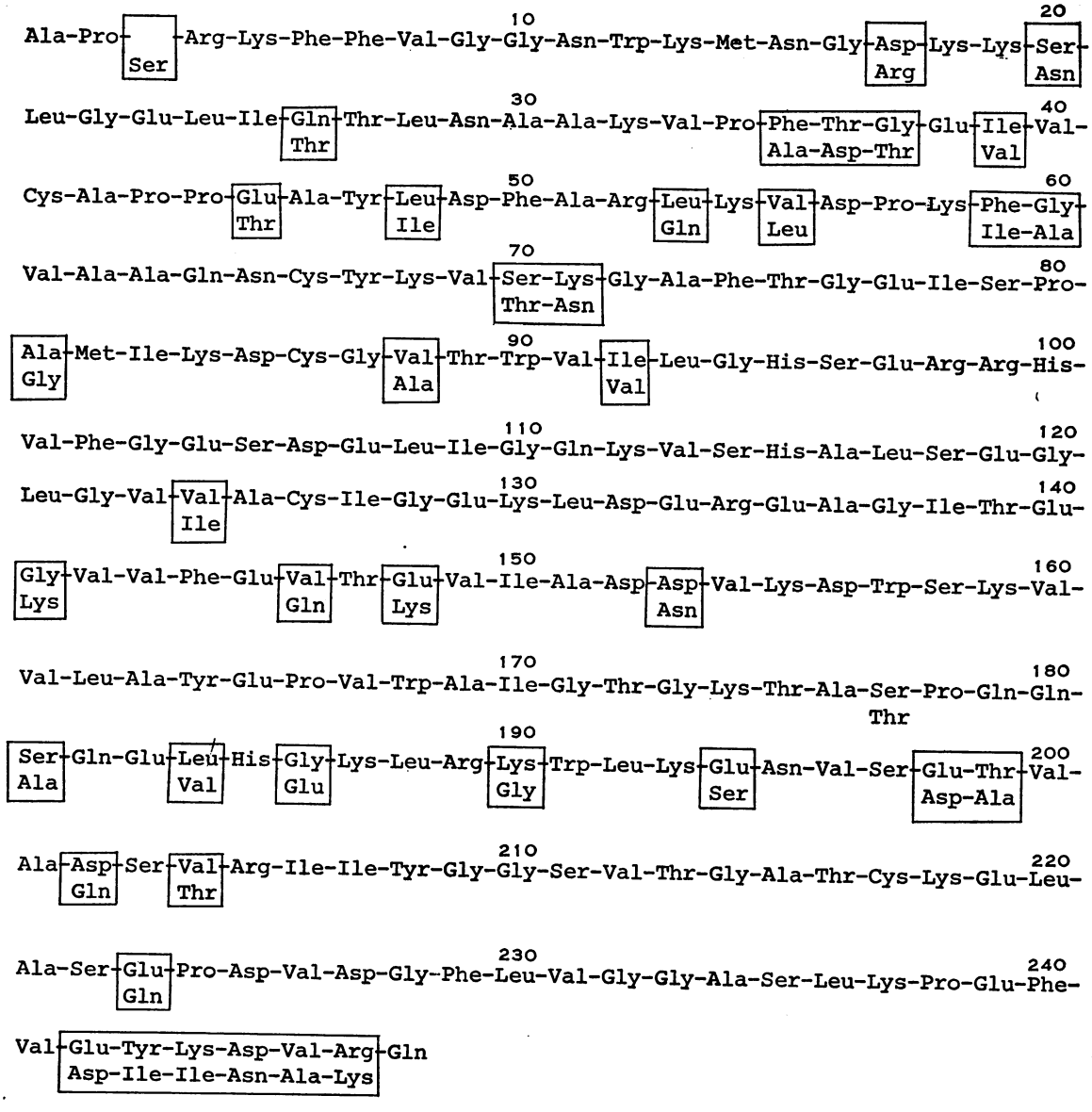


Fig. 6. *Comparison of the amino acid sequence of triose phosphate isomerase from coelacanth and rabbit muscle*

The sequences are identical except as indicated in the boxed positions below the continuous line of the coelacanth sequence. Residue 3 in the coelacanth enzyme is shown as a deletion in order to maximize the sequence homology between the two species.

generation of the original amino acid by hydrolysis with HI proceeds at about 60% efficiency, so that at least 10nmol of phenylthiohydantoin amino acid must be hydrolysed. In this study all the basic amino acids from the sequencer analyses were identified in this way. It should be noted that serine phenylthiohydantoin is converted into alanine during hydrolysis, and these two must therefore be differentiated beforehand by g.l.c. Methionine phenylthiohydantoin and tryptophan phenylthiohydantoin also suffer hydrolytic destruction, but fortunately these two phenylthiohydantoins are readily identified by g.l.c. analysis. Glutamine and asparagine are deamidated during hydrolysis and this deamidation can also occur on old g.l.c. columns. In general, assignment of amides is probably more reliable than by manual methods, but becomes less reliable as the run proceeds. The only reliable way of identifying cysteine is by scintillation counting of a ^{14}C-labelled, usually carboxymethyl or aminoethyl, derivative.

The general applicability of this approach will depend on (*a*) access to an automatic sequencer, and (*b*) the ability to produce and to fractionate fragments of suitable size for automated analysis. The number of useful methods for specific cleavage of protein chains to yield large fragments is regrettably small, but reports of enzymes specific for lysine (Wingard *et al.*, 1972) and glutamic acid residues (Houmard & Drapeau, 1972) may prove to be useful additions. In determining the sequence of coelacanth triose phosphate isomerase we were helped by the presence of relatively few methionine and arginine residues and by the unexpectedly high degree of homology with the rabbit muscle enzyme; but hindered by the presence of the Asn-Gly sequence close to the *N*-terminus of the protein.

Of the 247 residues that may be compared in the rabbit and coelacanth enzymes (Fig. 6) 207 (84%) are identical, and on the assumption (see Dayhoff, 1969) that the two species diverged approximately 400 million years ago this indicates an evolutionary rate of two amino acid substitutions per 100 residues per 100 million years. The morphology of the coelacanth appears to have remained unchanged during this period, but little is yet known about its detailed biochemistry. Comparison of two amino acid sequences of one protein is insufficient to provide a reliable estimate either for the time of evolutionary divergence of vertebrates or for the rate of molecular evolution of triose phosphate isomerase. It is, however, of interest that the slow rate of evolution found for triose phosphate isomerase corresponds very closely to that found for glyceraldehyde 3-phosphate dehydrogenase, the next enzyme in the glycolytic pathway, as shown by comparison (Jones & Harris, 1972) of the amino acid sequence of enzyme from three distantly related (pig, lobster and yeast) species.

We thank Dr. R. Acher (University of Paris) for the generous gift of coelacanth muscle, Dr. S. G. Waley for a pre-publication copy of the amino acid sequence of rabbit muscle triose phosphate isomerase, and Mr. F. Northrop for assistance with the automated sequencer analyses.

References

Ambler, R. P. (1963) *Biochem. J.* **89**, 349–378
Ambler, R. P. & Wynn, M. (1973) *Biochem. J.* **131**, 643–675
Bornstein, P. & Balian, G. (1970) *J. Biol. Chem.* **245**, 4854–4856
Braunitzer, G., Schrank, B., Ruhfus, A., Petersen, S. & Petersen, V. (1971) *Hoppe-Seyler's Z. Physiol. Chem.* **352**, 1730–1732
Braunitzer, G., Chen, R., Schrank, B. & Stangl, A. (1972) *Hoppe-Seyler's Z. Physiol. Chem.* **353**, 832–834
Bridgen, J. & Secher, D. S. (1973) *FEBS Lett.* **29**, 55–57
Bridgen, J., Kolb, E. & Harris, J. I. (1973) *FEBS Lett.* **33**, 1–3
Butler, P. J. G., Harris, J. I., Hartley, B. S. & Leberman, R. (1969) *Biochem. J.* **112**, 679–689
Corran, P. H. & Waley, S. G. (1973) *FEBS Lett.* **30**, 97–99
Davidson, B. E., Sajgo, M., Noller, H. F. & Harris, J. I. (1967) *Nature (London)* **216**, 1181–1185
Dayhoff, M. O. (1969) *Atlas of Protein Sequence and Structure*, vol. 4, National Biomedical Research Foundation, Silver Spring, Md.
Dixon, H. B. F. & Perham, R. N. (1968) *Biochem. J.* **109**, 312–413
Edelhoch, H. (1967) *Biochemistry* **6**, 1948–1954
Edman, P. (1956) *Acta Chem. Scand.* **10**, 761–768
Edman, P. & Begg, G. (1967) *Eur. J. Biochem.* **1**, 80–91
Gray, W. R. (1972) *Methods Enzymol.* **25**, 333–344
Gross, E. (1967) *Methods Enzymol.* **11**, 238–255
Harris, J. I. & Perham, R. N. (1968) *Nature (London)* **219**, 1025–1028
Hartley, B. S. (1970) *Biochem. J.* **119**, 805–822
Heilmann, J., Barollier, J. & Watzske, E. (1957) *Hoppe-Seyler's Z. Physiol. Chem.* **309**, 219–220
Hermodson, M. A., Ericsson, L. H., Titani, K., Neurath, H. & Walsh, K. A. (1972) *Biochemistry* **11**, 4493–4502
Hood, L., McKean, D., Farnsworth, V. & Potter, M. (1973) *Biochemistry* **12**, 741–749
Houmard, J. & Drapeau, G. R. (1972) *Proc. Nat. Acad. Sci. U.S.* **69**, 3506–3509
Jones, G. M. T. & Harris, J. I. (1972) *FEBS Lett.* **22**, 185–189
Jörnvall, H. (1970) *Eur. J. Biochem.* **16**, 41–49
Kolb, E. & Harris, J. I. (1972) *Biochem. J.* **130**, 26P
Konigsberg, W. (1967) *Methods Enzymol.* **11**, 461–469
Lämmli, U. K. (1970) *Nature (London)* **227**, 680–683
Landon, M., Langley, T. J. & Smith, E. L. (1971) *J. Biol. Chem.* **246**, 3807–3816
Laursen, R. A., Horn, M. J. & Bonner, A. G. (1972) *FEBS Lett.* **21**, 67–70
Miller, J. C. & Waley, S. G. (1971) *Biochem. J.* **122**, 209–218
Niall, H. D. (1971) *J. Agr. Food Chem.* **19**, 638–644

Offord, R. E. (1966) *Nature (London)* **211**, 591–593
Shotton, D. M. & Hartley, B. S. (1973) *Biochem. J.* **131**, 643–675
Smith, I. (1953) *Nature (London)* **171**, 43–44
Terhorst, C., Möller, W., Laursen, R. & Wittmann-Liebold, B. (1973) *Eur. J. Biochem.* **34**, 138–152
Wingard, M., Matsueda, G. & Wolfe, R. S. (1972) *J. Bacteriol.* **112**, 940–949