NUCLEOTIDE SEQUENCE OF A COLLAGEN cDNA-FRAGMENT CODING FOR THE CARBOXYL END OF PROα1(I)-CHAINS

A. M. SHOWALTER, D. M. PESCIOTTA, E. F. EIKENBERRY, T. YAMAMOTO*, I. PASTAN*, B. DeCROMBRUGGHE*, P. P. FIETZEK and B. R. OLSEN

Department of Biochemistry, CMDNJ-Rutgers Medical School, Piscataway, NJ 08854 and **Laboratory of Molecular Biology, Division of Cancer Biology and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA*

Received 2 January 1980

1. Introduction

Recently, several groups have reported the construction of recombinant plasmids and bacteriophages containing cDNA or genomic DNA sequences coding for procollagen polypeptides [1–5]. The cloning of such hybrid molecules in *Escherichia coli* permits a rapid determination of the amino acid sequence of collagen propeptides by analysis of the nucleotide sequence of the DNA. The preparation of cloned cDNA fragments also provides a method to investigate the regulation and the arrangement of procollagen genes in the genome.

Whereas the primary structure of the amino propeptide of proα1 chains of type I procollagen has been reported [6,7] it has been more difficult to obtain adequate amounts of carboxyl propeptides for conventional amino acid sequence determination [8] and until now no sequence data for this domain of procollagen have been reported. To obtain such information as rapidly as possible, we have used a combined approach employing both conventional protein chemistry techniques and rapid nucleotide sequencing of cloned cDNA fragments.

We have reported the construction of a recombinant plasmid containing cDNA sequences coding for the carboxyl end of proα1(I)-chains [5]. Here we report on the nucleotide sequence of this cloned cDNA fragment and the amino acid sequence of a major portion of the carboxyl propeptide of chick proα(I)-procollagen chains.

Address correspondence to B. R. Olsen

2. Materials and methods

Chick proα1(I) procollagen cDNA was synthesized and cloned in *E. coli* as in [5]. The cDNA insert was isolated from the recombinant plasmid, pCOL3, by digestion with *Hin*dIII and subjected to nucleotide sequence analysis following [9]. Sequence data from various experiments were compared, synthesized into a single sequence and translated using the computer programs of [10] which were extensively revised for use on the Data General Nova 3 computer.

The carboxyl propeptide of chick type I procollagen, labeled with a mixture of ^{14}C-labeled amino acids or with D-[2-^{3}H]-mannose, was prepared and purified as in [8].

The carboxyl propeptide was cleaved with a 100–1000-fold molar excess of cyanogen bromide in 70% formic acid at 30°C for 4 h. The peptide mixture was fractioned on a 2.5 × 160 cm Sephadex G-100 column equilibrated and eluted with 0.2 M ammonium bicarbonate at room temperature. Further purification of the peptides was accomplished by chromatography on a 1.5 × 5 cm CM-cellulose column equilibrated with 6 M urea, 0.05 M sodium acetate (pH 3.8) and eluted by a 600 ml linear gradient of 0–0.3 M NaCl at room temperature.

SDS–polyacrylamide slab gel electrophoresis was done as in [11]. The molecular weights of cyanogen bromide peptides were estimated by calibrating the gels with globular proteins of known molecular weights. Those of small peptides were also calculated from amino acid analyses.

Radioactive peptides were located in slab gels by fluorography and fluorographs were quantitated by

```
200
    LYS HIS VAL TRP PHE GLY GLU THR MET SER ASP GLY PHE GLN PHE VAL TYR
--G AAG CAC GTC TGG TTC GGC GAG ACG ATG AGC GAC GGC TTC CAG TTT G7G TAC

254
GLY GLY GLU GLY CYS ASN PRO VAL VAL VAL ALA ILE GLN LEU THR PHE LEU ARG
GGC GGT GAG GGT T6C AAC CCG GT7 GRT GTC GCC ATC CAA CTG ACC TTC CTG CGC

                                          CB 3a
308
LEU MET SER THR GLU ALA THR GLN ASN VAL THR TYR HIS CYS LYS ASN SER VAL
CTG ATG TCC ACC GAG GCC ACC CAG AAC GTC ACC TAC CAC TGC AAG AAC AGC GTC

                                          CB 3b
362
ALA TYR MET ASP HIS ASP THR GLY ASN LEU LYS LYS ALA LEU LEU LEU GLN GLY
GCC TAC ATG GAC CAC GAC ACC GGC AAC CTG AAG AAG GCT CTG CTG CTC CAG GGA

                                          CB 3b
416
ALA ASN GLU ILE GLU ILE ARG ALA GLU GLY ASN SER ARG PHE THR TYR GLY VAL
GCC AAC GAG ATC GAG ATC AGG GCC GAA GGA AAC AGC CGC TTC ACC TAT GGG GTC

                                          CB 3b
470
THR GLU ASP GLY CYS THR SER HIS THR GLY ALA TRP GLY LYS THR VAL ILE GLU
ACC GAG GAT GGC TGC ACG AGT CAC ACT GGA GCA TGG GGC AAA ACA GTG ATT GAG

                                                                        |
524                                                                     |
TYR LYS THR THR LYS THR SER ARG LEU PRO ILE ILE ASP LEU ALA PRO MET ASP
TAC AAG ACG ACG AAG ACT TCG CGC CTG CCC ATC ATT GAC TTG GCT CCT ATG GAC

    CB 4
578
VAL GLY ALA PRO ASP HIS GLU PHE GLY ILE ASP ILE GLY PRO VAL CYS PHE LEU
GTT GGC GCT CCG GAC CAT GAA TTT GGC ATT GAC ATC GGC CCC GTC TGC TTT TTG

632
***
TAA ACA GGA AAA AAA AAG AAA AAG AAA AGA AAA AAA AAA AAA AAA AAA GCC CCC

686
CCA ACG CGT GAC AGG AGA GAG TAA TAA TTA TAA TAA TTA ATA AAA AAA AAA AAA

740
AAA AAA CCG GCC AAA AAT GGR AAA AAA AAA AAA AAA AAA AAA AAA AAA CCA
```

Fig.1. The nucleotide sequence and the corresponding amino acid sequence of the coding strand of the cDNA insert of pCOL3. The cysteine residues in CB3a and CB3b are underlined. The carbohydrate acceptor site in CB3a is indicated by a double line. Uncertainties in the DNA sequence are indicated by 6(G or T), 7(A or T) and R(A or G).

scanning in a Joyce-Loebl 3CS microdensitometer.

The amino acid sequences of cyanogen bromide peptides were determined by automated Edman degradation in a Beckman model 890C protein–peptide sequencer [12]. The PTH-amino acid derivatives were identified by thin-layer or by high-pressure liquid chromatography [12].

Peptides were hydrolyzed and amino acid analyses were performed as in [12].

3. Results and discussion

The procollagen cDNA insert of the plasmid pCOL3 contained 800 base pairs [5]. Cleavage of the insert with *Kpn* I-produced two fragments; one 200 basepair-fragment containing the 5′-end of the coding strand of the insert and one 600 basepair-fragment containing the 3′-end of the coding strand of the insert. The nucleotide sequence of the small *Kpn* I-fragment could not be translated into a meaningful amino acid sequence since the 3 possible reading frames contained several termination codons. It appears, therefore, that reverse transcription or the subsequent amplification of the cDNA in *E. coli* did not faithfully reproduce the proα1(I) mRNA sequence in this region. The 600 basepair *Kpn* I-fragment, however, yielded a sequence that could be translated into a plausible amino acid sequence. This nucleotide sequence and the corresponding amino acid sequence are shown in fig.1. The other 2 possible reading frames were excluded because they contained several internal termination codons.

To directly determine the amino acid sequence of the carboxy propeptide of proα1(I)-chains, we cleaved the carboxyl propeptide of type I procollagen with cyanogen bromide. When the cleavage products of the unreduced carboxyl propeptide were analyzed by slab gel electrophoresis, 5 unique peptides were identified in addition to variable amounts of partially digested material (fig.2). We have designated these peptides CB1–5 (fig.2) The structure of CB1, CB2 and CB5 will be reported in [13].

The peptide CB3 migrated as a broad band on slab gel electrophoresis with app. mol. wt 11 000–14 000 when examined without reduction. After reduction with 2-mercaptoethanol, it migrated as a single band with an app. mol. wt 7500. We conclude, therefore, that CB3 is a dimer of two peptides each having mol. wt ∿7500 linked together by at least one disulfide bond. After sequential Edman degradation of CB3, two degradation cycles showed the presence of two PTH-derivatives in about equal amounts (table 1). CB3 must therefore contain two different kinds of peptides.

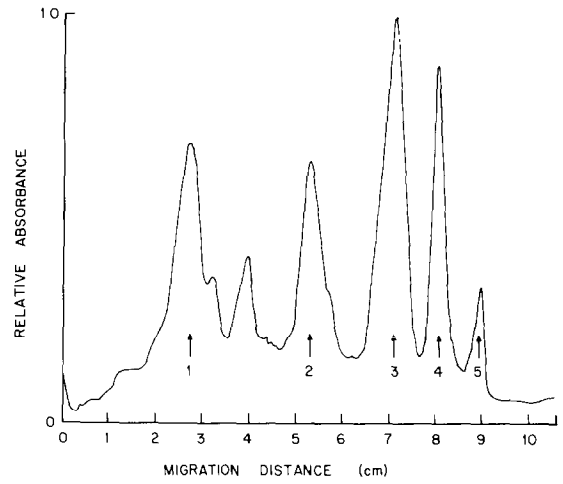

Fig.2. Densitometer tracing of cyanogen bromide peptides separated on a 15% polyacrylamide gel. The gel was stained with Coomassie blue.

Table 1
Amino acid sequence analysis of CB3

Degradation cycle	Residues identified by high pressure liquid chromatography	Residues predicted from DNA sequence	
		CB3a	CB3b
1	Asp (2.2)	Asp	Ser
2	Thr (0.8)	His	Thr
3	Not determined	Asp	Glu
4	Ala (2.0)	Thr	Ala
5	Gly (1.2)	Gly	Thr
6	Asn (1.5), Gln (1.5)	Asn	Gln
7	Not determined	Leu	Asp
8	Lys (0.3), Val (1.5)	Lys	Val
9	Lys (0.4)	Lys	Thr
10	Ala (1.8), Tyr (1.7)	Ala	Tyr
11	Leu (1.6)	Leu	His

Amount of CB3 used for analysis was not determined because of the limited amounts of peptide available and difficulties in solubilizing the peptide. Yields of PTH-derivatives recovered from the high pressure column are given as nmol in parentheses. Special precautions to quickly identify PTH-derivatives of serine and threonine were not taken. Therefore, these residues were probably not detected in cycles 1,4,5 and 9 because of their instability. Also, aqueous phases (containing the PTH-derivative of histidine) were not analyzed

Comparison of the amino acid sequence derived from the nucleotide sequence and the composite amino acid sequence obtained by Edman degradation shows that CB3 must be coded for by bases 314–574 in the cDNA. This region of the DNA contains two potential cyanogen bromide peptides which must represent the two subunits of CB3. We have designated these peptides CB3a and CB3b. The peptide CB3b (fig.1) is 68 amino acid residues long. With calculated mol. wt 7436 it is in agreement with the value obtained by gel electrophoresis. The peptide CB3a (fig.1) is only 19 amino acid residues long and is, therefore, too short to represent a peptide of mol. wt 7500. It is likely that the discrepancy between the apparent molecular weight as determined by gel electrophoresis and from the amino acid composition of CB3a can be explained by the presence of carbohydrate in CB3a.

We had shown that the carboxyl propeptides of both the proα1(I)-chains and the proα2-chains contain 2 residues of *N*-acetyl-glucosamine and 10 residues of mannose [8]. Labeling of the carboxyl propeptide with [^{3}H]mannose demonstrated that > 50% of the radioactivity incorporated into the carboxyl propeptide was recovered in CB3 [13]. This indicates that most of the mannose-rich, asparagine-linked [14] carbohydrate in the carboxyl propeptide of type I procollagen is linked to CB3. Furthermore, CB3a is the only region within the cDNA sequence that contains an acceptor site for asparagine-linked carbohydrate [15]. This sequence of Asn–Val–Thr corresponds to bases 332–340 in the cDNA (fig.1). CB3a must therefore contain the mannose-rich carbohydrate of proα1(I) carboxyl propeptides.

Gel electrophoresis after reduction of CB3 indicates that the peptide contains at least one disulfide bond. The cysteine residues in CB3a and CB3b (fig.1) must, therefore, form one of the intrachain disulfide bonds of the proα1(I) carboxyl propeptide.

The peptide CB4 is derived from the proα1(I) carboxyl propeptide [13]. Edman degradation of CB4 showed that the peptide had the amino-terminal sequence Asp–Val–Gly–Ala–Pro–Asp–Val–Tyr–Pro–Gly–Leu–Ala [13]. Amino acid analysis showed that the peptide lacked cysteine, but contained homoserine. The amino-terminal sequence of CB4 is found in the cDNA corresponding to bases 575–592. However, only the first 6 residues of CB4 are found in the DNA-derived sequence. The remaining cDNA-derived amino acid sequence, from base 593 to the TAA termination codon cannot represent part of CB4.

(i) CB4 contains homoserine; the last 13 amino acid residues coded for by the cDNA as indicated in fig.1 do not include methionine.
(ii) CB4 does not contain cysteine; the cDNA-derived sequence contains 1 residue of cysteine.
(iii) Based on amino acid analysis CB4 contains ~40 residues. This is about twice the size of the cyanogen bromide fragment the cDNA-derived sequence would generate.

It is possible that CB4 is not adjacent to CB3b, but is located elsewhere. We find this unlikely. Analysis of all the cyanogen bromide peptides derived from the carboxyl propeptide of type I procollagen shows that no other peptide has an amino-terminal sequence identical to that of CB4 [13]. A more likely possibility is that there is a deletion in the cloned cDNA following the aspartic acid codon GAC of bases 590–592. We do not know the size of this deletion, nor do we know whether the last 13 amino acid residues derived from the cDNA sequence as shown in fig.1 represent a true portion of the amino acid sequence of the proα1(I) carboxyl propeptide. Continued use of the combined cDNA–protein sequencing approach used here should allow for the elucidation of the complete primary structure of the carboxyl propeptides of procollagen.

Acknowledgements

We thank Ms Weijia Chi and Mr Thomas Troy for important technical assistance. The study was supported in part by research grant AM 21471 from the National Institutes of Health of the United States Public Health Service.

References

[1] Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F., Crkvenjakow, R., Boedtker, H. and Doty, P. (1978) Proc. Natl. Acad. Sci. USA 75, 5417–5421.
[2] Sobel, M. E., Yamamoto, T., Adams, S. L., DiLauro, R., Avvedimento, E. V., deCrombrugghe, B. and Pastan, I. (1978) Proc. Natl. Acad. Sci. USA 75, 5846–5850.
[3] Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F. and Boedtker, H. (1979) Biochemistry 18, 3146–3152.
[4] Boyd, C. D., Tolstoshev, P., Schafer, M. P., Trapnell, B. C., Coon, H. C., Kretschmer, P. J., Nienhuís, A. W. and Crystal, R. G. (1980) J. Biol. Chem. in press.

[5] Yamamoto, T., Sobel, M. E. Adams, S. L., Avvedimento, V. E., DiLauro, R., Pastan, L., deCrombrugghe, B., Showalter, A., Pesciotta, D., Fietzek, P. and Olsen, B. (1980) J. Biol. Chem. in press.
[6] Rohde, H., Wachter, E., Richter, W. J., Bruckner, P., Helle, O. and Timpl, R. (1979) Biochem. J. 179, 631–642.
[7] Hörlein, D., Fietzek, P. P., Wachter, E., Lapière, C. M. and Kuhn, K. (1979) Eur. J. Biochem. 99, 31–38.
[8] Olsen, B. R., Guzman, N. A., Engel, J., Condit, C. and Aase, S. (1977) Biochemistry 16, 3030–3036.
[9] Maxam, A. M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560–564.
[10] Staden, R. (1979) Nucleic Acids Res. 6, 2601–2610.
[11] Olsen, B. R., Hoffmann, H.-P. and Prockop, D. J. (1976) Arch. Biochem. Biophys. 175, 341–350.
[12] Pesciotta, D. M., Silkowitz, M. H., Fietzek, P. P., Graves, P. N., Berg, R. A. and Olsen, B. R. (1980) submitted.
[13] Pesciotta, D. M., Fietzek, P. P. and Olsen, B. R. (1980) in preparation.
[14] Clark, C. C. (1979) J. Biol. Chem. 254, 10798–10802.
[15] Hart, G. W., Brew, K., Grant, G. A., Bradshaw, R. A. and Lennarz, W. J. (1979) J. Biol. Chem. 254, 9747–9753.