Nucleic Acids Research

Isolation and characterization of a human collagen α1(I)-like gene from a cosmid library

Elisabeth H.Weiss, Kathryn S.E.Cheah*, Frank G.Grosveld, Hans Henrik M.Dahl+, Ellen Solomon* and Richard A.Flavell

Laboratory of Gene Structure and Expression, National Institute for Medical Research, Mill Hill, London NW7 1AA, and *Laboratory of Somatic Cell Genetics, Imperial Cancer Research Fund Laboratories, P.O. Box 123, Lincoln's Inn Fields, London WC2A 3PX, UK

ABSTRACT

        We have isolated a human collagen α1(I)-like gene from a cosmid library.   The clone which contains 37kb of human DNA has been shown to contain this gene by DNA sequencing, hybrid arrest and hybrid selection assays and Northern blot hybridizations.   The collagen gene sequence extends through most of the cloned DNA and must, therefore, be at least 35kb in length.

INTRODUCTION

        Collagen is one of the most important and abundant proteins of vertebrate connective tissue.   At least 5 genetically distinct types of collagen have been identified and each tissue expresses one, or several, of these types.   The types of collagen are assembled from 9 different chains en-coded by distinct genes (1, 2).   Nothing is known of the copy number or extent of polymorphism of these genes, and the number of genes in this family may indeed be much greater than 9.   It is clear however that the regulation of the different collagen genes is basic to the development and differentia-tion of vertebrates.

        The most abundant and best studied of the vertebrate collagens is Type I, a heteropolymer of two α1(I) chains and one α2(I) chain (1, 2). In vivo these molecules are synthesized as large precursor pro-α chains which associate to form procollagen.   Genomic clones for the entire chick α2(I) (3, 4) and part of the sheep α2(I) gene (5) have been isolated and reveal several extraordinary features.   First, 5kb of coding sequence are contained within a gene at least 38kb in length.   Second, the coding sequence is inter-rupted by 50 intervening sequences, the largest number found in a gene to date.   Third, the size of the coding regions is most frequently 54bp, or 18 amino acids, in length.   Others are 108 (2 x 54), 99($\left[2 \times 54\right]$ -9), 45 (54-9).   On the basis of these findings 54bp has been proposed as the

size of the primordial collagen gene (6, 7).

We report here the isolation and characterisation of a human collagen α1 (I)-like gene as a cosmid recombinant. This clone contains 37kb of human DNA and may contain the entire gene.

MATERIALS AND METHODS

The general procedures for the screening of cosmid libraries, growth and analysis of cosmid recombinants, nick translation and blot-hybridizations were performed as described previously (15). DNA sequences were determined as in Maxam and Gilbert(24).

Hybrid-arrested translation of procollagen mRNA

Total polyA$^+$ RNA was prepared from approximately 10^9 human skin fibroblasts as previously described (8). DNA (5μg) was dissolved in deionized formamide, heated to 90°C for 5 min., cooled and hybridized with 2μg polyA$^+$ RNA in 80% formamide, 0.4M NaCl, 20mM PIPES pH6.4 for 3h at 48°C. The hybridization mixture was diluted 1 in 3 with ice-cold water and divided into 2 aliquots, one of which was heated at 90°C for 30 sec.

The RNA-DNA hybrids were precipitated with 2.5 vols. ethanol, redissolved and translated in a messenger-dependent reticulocyte cell-free system (9) under conditions optimal for procollagen mRNA (10). The trans-lation products were analysed on SDS-polyacrylamide gels (10%) as described previously (10).

Hybrid-selected translation of mRNA

The EcoRI fragments of cosH col.I the chicken α1(I) cDNA probe and a chicken actin cDNA probe (gift of D. Cleveland, Johns Hopkins University, Baltimore, USA) were each linked separately to diazobenzyloxymethyl (DBM) paper discs essentially as described by Alwine et al. (12) and Goldberg et al. (13). These fragments were prehybridized for 1-2h at 37°C in a mixture containing 50% formamide, 0.9M NaCl, 20mM PIPES pH6.4, 1mM EDTA, 0.2% SDS, 0.5μg/μl calf liver tRNA (Boehringer Mannheim) and 0.5μg/μl polyA (Sigma). Human fibroblast total polyA$^+$ RNA (0.2μg/μl hybridization volume) was hybrid-ized to these DNA fragments for 6h at 37°C in the same mixture, except polyA was excluded. At the end of the hybridization period the mRNA mixture was removed, diluted with ice-cold water and precipitated with 2.5 vols. ethanol. The DBM paper discs were washed twice for 10 min at 37°C in prehybridization buffer; 6 times with 1 x SSC (150mM NaCl, 15mM tri-sodium citrate) 0.2% SDS for 5 mins each at 37°C; twice in 10mM Tris-HCl pH7.4, 2mM EDTA for

5 min.,each at 37°C.   The mRNA hybridized to the DNA was eluted in 20mM
PIPES pH6.4, 1mM EDTA, 0.2% SDS for 1-2h at 37°C, diluted 1 in 3 with ice-
cold water and precipitated with ethanol.   The bound and unbound mRNAs
were translated in the reticulocyte lysate cell-free system and the trans-
lational products analyzed in SDS-polyacrylamide gels as before.

RNA blotting (Northern) experiments

Fibroblast polyA[+] RNA was denatured in 50% formamide, 6% formal-
dehyde, 20mM morpholinepropanesulphonic acid (MOPS) pH7.0, 5mM sodium
acetate and 1mM EDTA, electrophoresed in a 0.8% agarose gel containing 2.2M
formaldehyde and blotted onto nitrocellulose as described by Crain et al.,
(14).   The filters were prehybridized at 42°C for 3-4h in 50% formamide,
5 x SSC, 16.6mM phosphate buffer (pH7.2), 0.1% sodium pyrophosphate, 0.02%
Ficoll (Pharmacia), 0.02% polyvinylpyrrolidone (Pharmacia), 20µg/ml nuclease-
free bovine serum albumin, 100µg/ml denatured salmon sperm DNA (Sigma), 10%
dextran sulphate (Pharmacia).   Where the DNA probe was known to contain
repetitive sequences, human DNA at 30µg/ml was included in the prehybridiza-
tion and hybridization mixtures.   The EcoRI fragments of cosH col.1, pCg54
and the actin cDNA clone were labelled with ^{32}P-dCTP (Amersham, U.K. specific
activity 3000Ci/mmol) by nick-translation and hybridized to the RNA blots for
18-24h at 42°C in the same buffer as for prehybridization.   After hybridiza-
tion the filters were washed 4 times for 5 min. each in 2 x SSC,0.1% SDS
at 20°C and twice for 15 min. each in 0.1 x SSC, 0.1% SDS at 50°C.   The
filters were dried and exposed to Kodak X-Omat S film at -70°C.   The sizes
of the mRNA molecules which hybridized to labelled probe were determined
from a calibration curve made from the mobilities of fragments of ribosomal
RNA of known molecular weight obtained from ppp(A2'p)$_2$A-treated mouse cells
(gift of R.H.Silverman, ICRF, London).

RESULTS

Isolation of a human collagen gene cosmid

We have previously described the construction of a human cosmid
library using the vector pJB8 (15).   We screened this library using as probe
the chicken procollagen α1(I) cDNA clone pCg54 (16).   Several positive
colonies were picked and after preliminary analysis, one clone, cosH col.1 was
selected for further work.

The restriction map for this cosmid is shown in Fig. 1.   The regions
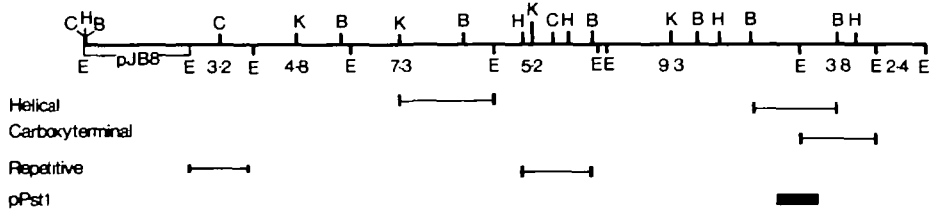of the cosmid that contain the helical- and C terminal-coding segments of the

Fig. 1.   Restriction map of cosH col.1.
The EcoRI sites at the ends of the insert were generated during the construc-
tion of the human cosmid library using the vector pJB8 (15).   The 1.1kb PstI
fragment was isolated (pPstl) for DNA sequence analysis; its location is
indicated.   The location of sequences, coding for helical and C-terminal
region or containing repetitive sequences are indicated.
(B) BamHI;   (C) ClaI;   (E) EcoRI;   (H) HindIII;   (K) KpnI.
The numbers refer to the size of the EcoRI fragments in kilobase pairs.

gene were localized by hybridization of the relevant portions of the chicken

α1(I) cDNA plasmid (see Fig. 2).   In this way, the C terminal region was

localized to a 3.8kb EcoRI fragment (E3.8) of the cloned DNA, while the helical

segments were localized in   E7.3, E9.3 and E3.8.   These are placed

on the map of Fig. 1; this shows that the C-terminal segment is near one

end of the clone, while the helical segments extend through most of the

remainder of the human DNA.   Since the chicken cDNA probe used only covers

the 3' terminal regions of the helical segments, the extent of the helical

segments on the cosmid may well be an underestimate.   Regions containing

highly repetitive sequences were localized by hybridizing EcoRI digested

cosH col.1 with ^{32}P nick-translated human placental DNA.   Both the E5.2,

which is located within the collagen gene, and the 5' terminal fragment E3.2

hybridize strongly to this probe (Fig. 2).

CosmidH col.1 specifically binds human procollagen α1(I) mRNA

To determine which collagen gene we had isolated, we assayed for

the ability of the cloned DNA to hybridize specifically to procollagen mRNA.

To do this, we first subcloned a PstI fragment (pPstl) which hybridized

strongly to the chicken pCg54 collagen α1(I) cDNA plasmid.   Total fibroblast

polyA$^+$ RNA was hybridized to pPstl and the mRNA-DNA hybrids translated in a

messenger dependent reticulocyte lysate cell-free system.   The cell-free

translation of procollagen α1(I) mRNA was specifically arrested by hybridiza-

tion of mRNA to pPstl, but when the mRNA-DNA hybrid was disrupted by heating

to 90°C a polypeptide corresponding to procollagen α1(I) chain was synthesized.

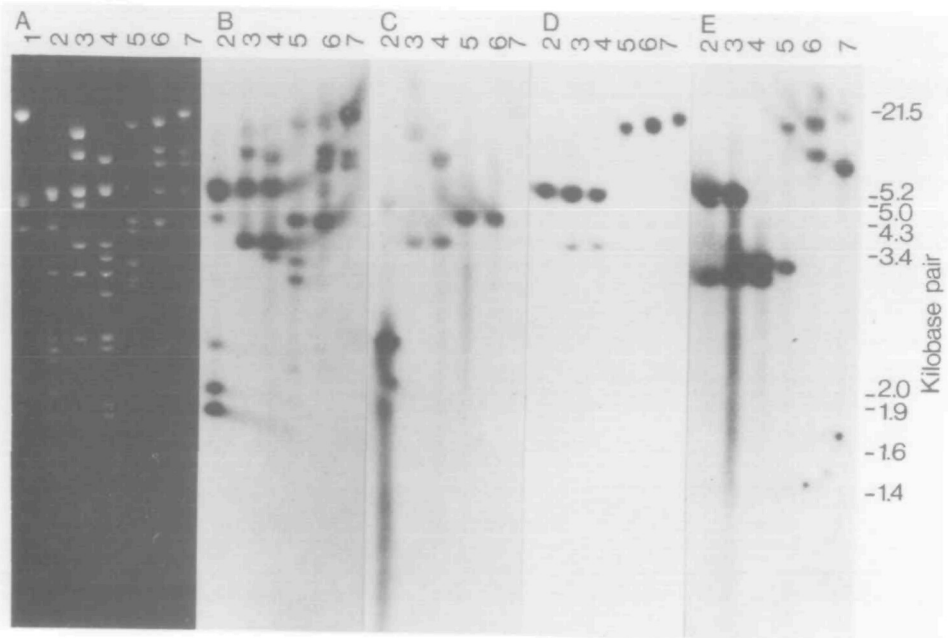The cell-free synthesis of the procollagen α2(I) chain was unaffected by

Fig. 2a. Analysis of cosH col.1 by Southern blot hybridization.
A   Ethidium bromide staining
B,C,D,E   Autoradiographs
B   Hybridized with ^{32}P-labelled Kpn-HindIII fragment of pCg54 which
           corresponds to essentially the whole insert.
C   Hybridized with ^{32}P-labelled 630bp fragment of pCg54 coding for
           helical region (see Fig. 2b).
D   Hybridized with ^{32}P-labelled 360bp fragment of pCg54 coding for
           C-terminal region (see Fig. 2b).
E   Hybridized with ^{32}P-labelled human placental DNA.
track 1: λ x HindIII/EcoRI
      2:   cosH col.1 digested with BamHI and EcoRI
      3:   cosH col.1 digested with EcoRI
      4:   cosH col.1 digested with KpnI and EcoRI
      5:   cosH col.1 digested with BamHI and KpnI
      6:   cosH col.1 digested with BamHI
      7:   cosH col.1 digested with KpnI
The 5.2kb band detected by the 360bp probe and the Kpn HindIII fragment of
pCg54 is the cosmid vector pJB8 which hybridizes to traces of contaminating
pBR322 from the pCg54 preparation.

hybridization to pPst1 (data not shown). In addition, a hybrid selection experi-
ment was performed; pPst1 was bound to DBM paper and fibroblast polyA[+] mRNA
was hybridized to this immobilized DNA. As controls, pCg54 and a chicken
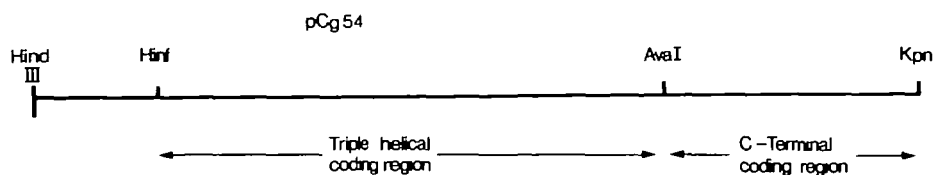actin cDNA clone were used. The hybridized RNA was then eluted and trans-

Fig. 2b. Schematic map of pCg54 showing the fragments used as probes.

lated. Fig. 3 shows that again the synthesis of pro α1(I), but not pro α2(I) collagen is directed by the mRNA bound to both pPstl and pCg54.

The DNA sequence of a segment of cosHG identifies an α(I)-like collagen gene

To establish the identity of the collagen gene in cosH col.I, we determined the DNA sequence of the subclone pPstl. The sequencing strategy and the general structure of this region is shown in Fig. 4 and the DNA sequence is shown in Fig. 5. The sequence shows the presence of 5 exons of a collagen gene. We aligned the polypeptide sequence predicted from this with the sequence of collagen α1(I) of both chicken and calf (17, 18). The human sequence shows homology with both sequences from residues 838 to 937 (Fig. 6). There is no identifiable region of similarity with collagen α2(I).

Comparison of amino acids 838-937 from the chick and calf α1(I) chains reveals about 90% homology in the non-glycine residues. Our clone shows only 60-65% homology with these sequences. Greater conservation across species might be expected for a particular chain type and we therefore cannot exclude that we have cloned a collagen gene other than α1(I). cosH col.1 cannot however contain the α1(III) gene expressed by fibroblasts because our DNA sequence does not correspond with the published amino acid sequence of the human α1(III) chain (25) cosH col.1 also cannot contain the α1(II) gene because it does not hybridize with chick cartilage polyA$^+$ mRNA, which includes α1(II) mRNA (data not shown), nor does it show homology with this region of the calf α1(II) chain (W. Butler, personal communication). We therefore propose that cosH col.1 is either the human α1(I) gene, a closely related α1(I)-like gene, or pseudogene (see Discussion). For the remainder of this article we shall refer to the gene cloned in cosH col.1 as the human collagen α1(I) gene.

The gene structure of this region is very similar to that previously described for the chicken and sheep collagen α2(I) genes, since the three complete exons are all multiples of 54 nucleotides (54, 108 and 54 nucleotides respectively (3, 4, 5)).
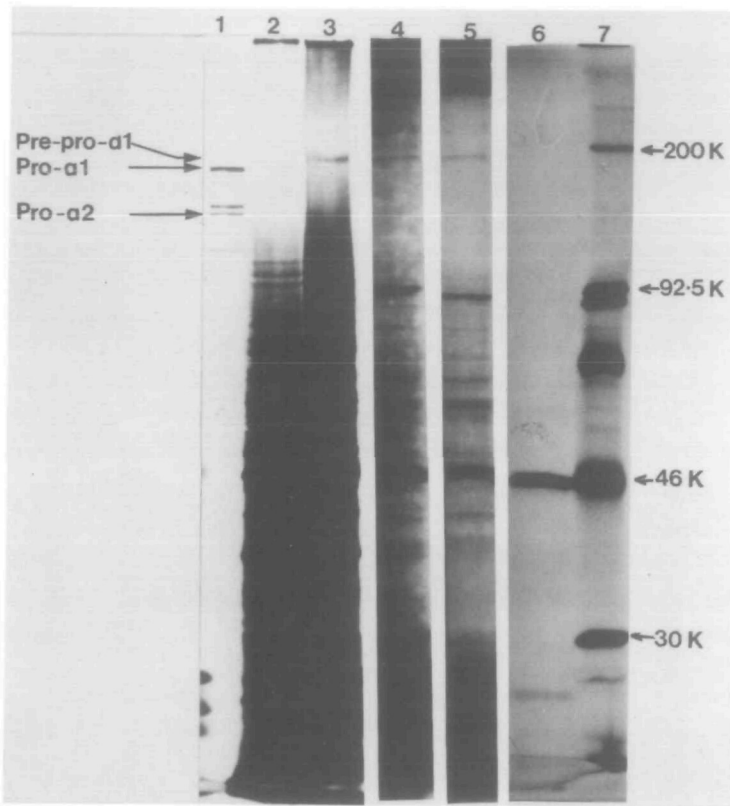
Fig. 3. Hybrid-selected translation of human fibroblast mRNA.
Fluorograms of SDS polyacrylamide gels of translation products labelled with
^{35}S-methionine.   Tracks; 1, pro-α1(I) and pro-α2(I) standards; 2, trans-
lation products of residual mRNA not selected out by the different DNA
fragments;   3, translation products of fibroblast mRNA;   4, products of pPst1
selected mRNA;   5, products of pCg54 (chicken α1(I) cDNA clone) - selected
mRNA;   6, translation products of actin cDNA-selected mRNA;   7, polypeptide
molecular weight markers (Amersham Radiochemicals U.K.).
Tracks 1-3 and 6-7 represent shorter exposure times than tracks 4 and 5
because of the high incorporation of radioactivity in the former.   Longer
exposure of track 6 does not show any polypeptide of similar molecular size
to pro-α1(I) or pro-α2(I) chains, but does show the products up to 92,000MW
seen in tracks 4 and 5.   These are also observed in the absence of added
mRNA (not shown).
The products of cell-free translation of procollagen mRNA do not undergo the
post-translational modification of glycosylation and hydroxylation which occur
in vivo.   Therefore unhydroxylated ^{14}C -proline labelled type I pro-α
chains were used as standards.   The translation products do not comigrate
exactly with these standards because they are synthesized as pre-proforms
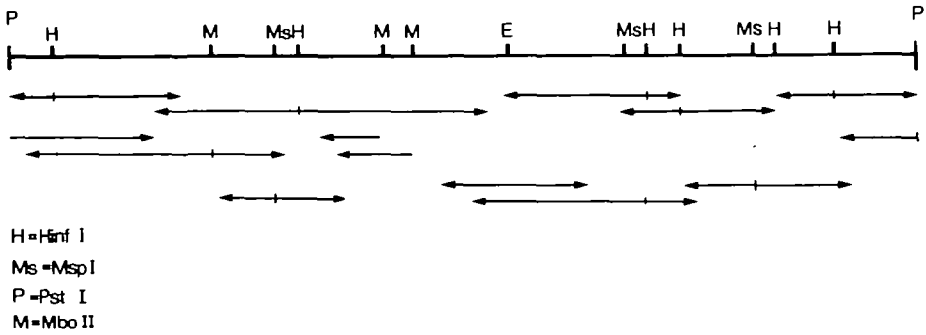but their collagenous identity has been established previously (10).

H = Hinf I
Ms = Msp I
P = Pst I
M = Mbo II

Fig.4 . Map of the 1.1kb PstI fragment (pPst1) showing the position of
         individual sequences obtained.
The straight arrows give the 5'-3' direction and length of the fragment
sequenced as in Maxam and Gilbert (24).

At least 35kb of cosH col.1 is collagen α1(I) gene
         The sizes of the sheep and chicken collagen α2(I) genes are
30-38kb long (5, 3).   Since cosH col.1 is 38kb and sequences corresponding
to the carboxy terminus are located close to the vector moiety (Fig. 2), we


```
            10        20        30        40        50        60        70        80        90       100
         GGTGAACCTC GACGACACCT GACCAGTGAG ACCCCCTGGG CTCGCCCTCA TTGGCGACAG GGGCCCTGTG AGTCTCTGTG CTGGGTCACG AAGGACAAGC

           110       120       130       140       150       160       170       180       190       200
         CCCAGTCAGG GCCTCGCAGA AGGGGCGCGC ACCGCTGCCC GACAGGCCAA ACCCTAGGTA CAATGGCAAG GTTCTCGGG AGACAGACGG GCATAGAGAC

           210       220       230       240       250       260       270       280       290       300
         CAAGCGCTGC TTCTGGAAGC AGGAGGGAAA CTTCGCTGCG AAACTTTGGC TTCAAAGTGT GAGTGAGTTG CCCAGAACAG GAGAGGCCTG CGCTTCTGAG

           310       320       330       340       350       360       370       380       390       400
         AGGGGCTGCG GGAGCAGAGG GGGAGGTGCA CAGAGGACAG CTCTAGGTGC GTTCTTGTTT CACTTTGTCC AGGGAAGCCC GGGTGCTGAT GGCCCCCCTG

           410       420       430       440       450       460       470       480       490       500
         GGCACAGATGG GGCTGCTGCCA GTGAAGGTGA GTGTGTGGTG TGTGTGTGTG CACTGGGTTG GGGAGGACAT TGCCTGGGGG CTGACAGGTG ACGTGGGGGT

           510       520       530       540       550       560       570       580       590       600
         GGCAGGTTGG AACAAGTGTG ATGTGAGGGT AGAAGGACGT TGTGTTGGTG TGTGTTGTGG AAGATTGTTG TGTGAGGGTG AGACGTGTGT GGTGACAGGG

           610       620       630       640       650       660       670       680       690       700
         TAATGGTGGT GAAACGGGTG GTGTGGGAGG TGGTGGAACG GGTGGGGGGG GTGGGTGGGG TGGGGGGGGT GGTGGAACTG GGAAGGAAGG AGAGAGAGGA

           710       720       730       740       750       760       770       780       790       800
         GAAGGTGTAA GTATGGTGGA ATTGAGTAAA AGGGGGGTTG GGGTGGGGGG TGGGGGTGAG GGAGTTGGTG GGTTTTGGGA GTGTGTGGAC TGAGGAGGAC

           810       820       830       840       850       860       870       880       890       900
         TGGGGGGAGA TGGAGAGGAG GGGGGGACGT GTGGTGGTTT TGTGTAGGGT GGGGTGAGTG TGTGGTGAGG GTGGACAAGG GGGGATGGGA GGGTGAGGAC

           910       920       930       940       950       960       970       980       990      1000
         GAGGTGGAGG GGGGGGGAATG GAGGTGAGTA TGGAAGTGTG GGGGACGGAG TGGGGAGGAG GGATAGGGTG GGAGGGGAGG GAGGGTGGAG GTGGTTGGTG

          1010      1020      1030      1040      1050      1060      1070      1080      1090      1100
         GGGTGGAAGGC GTGTGGTTGG GGGGATTGGT GAGGTTGGGT GGGAGAGGGAG GGGGGTGGTG TGGTGGGGGT GAGGTGAGTG AATGGGTGTG TGTGTTGTTG

          1110      1120      1130      1140      1150      1160      1170      1180      1190
         GGAGGGTGGT GAAGGGGGGA GAGGTGAGAA AGGAGAGGGT GGAGAGGGTG GGGAGAGAGG GGTGAAGGGA GAGGGTGGGT TGAGTGGGG
```


Fig.5 .   The DNA sequence of the 1.1kb PstI fragment of cosH col.1
The underlined regions are the collagen exons.   The position of introns and
exons was deduced by comparison of the DNA sequence translated protein with
the known sequence of other collagens and by aligning introns on the basis
of the GT---AG rule.

```
                831                 838
pPstl           G E P G R E   G S P G A D G P P G R D G A A G V K
chick α1(I)       A             A     E   A                   P
bovine α1(I)      S             A     E   S               S P   A


                856
pPstl           G N R G E T G A V G A P G T P G P P G S P G P A G P T
chick α1(I)       D           P A   P     A     A     A       V     A
bovine α1(I)      D           P A   P     A     A     A       V     A


                                    892
pPstl           G K Q G D R G E A   G A Q G P M G P S G P A G A R G J Q
chick α1(I)         N           T       P A     A     P             P A
bovine α1(I)        S           T       P A     I     V             P A


                910
pPstl           G P Q G P R G D K G E A G E P G E R G L K G H R G F T G
chick α1(I)                    T       Q     D       M
bovine α1(I)                 Z T     Z Z     B       I                 S
```

Fig.6 .
A comparison of the amino acid sequence encoded by the collagen exons in subclone pPstl with the published sequences of chicken and bovine collagen α1(I). The standard single letter amino acid code is used. The gaps represent the boundaries of the respective exons and the numbers in the sequence refer to the first amino acid of each exon. The sequence of the chicken and bovine collagens is shown only where they differ from the human sequence.

considered it possible that this clone contained the entire human collagen α1(I) gene.

We have used the chicken cDNA clone pCg54, or the pPstl subclone of the human collagen α1(I) gene, described here as labelled probes in Northern blots of human fibroblast polyA$^+$ mRNA. Both probes detect two mRNAs of 6.9kb and 5.7kb. We presume that one or both of these species is the human procollagen α1(I) mRNA (Fig. 7; see refs. 19 and 20). We have used this Northern assay to determine the boundaries of the collagen α1(I) gene on our cosmid clone. To do this, Northern blots of human fibroblast polyA$^+$ mRNA were hybridized with each respective EcoRI fragment of cosH col.1(Fig. 7). All EcoRI fragments detect the 6.9kb and 5.7kb putative collagen α1(I) mRNAs except the 5' terminal EcoRI + ClaI double digest fragment. Though the hybridization with E3.2 is weak (e.g. Fig. 7) we reproducibly observed positive hybridization to the putative procollagen mRNAs. It appears, therefore,
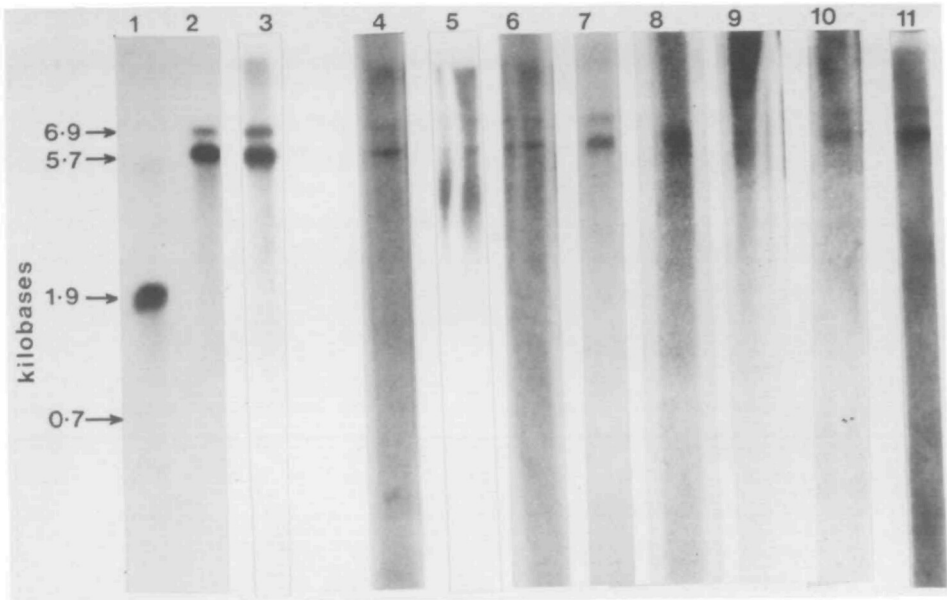
Fig.7 .   Hybridization of ^{32}P-labelled EcoRI fragments of cosH col.1 to
           mRNA in a Northern blot.
The positions of the EcoRI fragment within cosH col.1 are shown in Fig. 1.
^{32}P-labelled DNA fragments used as probes in tracks 1-11 were as follows:
1, actin cDNA clone; 2, pCg54; 3, E9.3; 4, E7.3; 5, E5.2; 6, E4.8;
7, E3.8; 8, E3.2; 9, 5' fragment of E3.2 digested with ClaI; 10, E2.4;
11, pPstl.   The lengths of the mRNA species which hybridized to labelled
probes are indicated and were determined from molecular weight markers of
fragments of RNA (see Materials and Methods).

that the collagen α1(I) gene extends throughout essentially the entire human

DNA insert in cosH col.1.   The same experiment was performed using the hybrid

selection assay.   All the EcoRI fragments with the exception of the 5'

terminal EcoRI + ClaI fragment select collagen α1(I) mRNA, as shown by this

assay (not shown).

DISCUSSION

        We describe here the isolation of the human collagen α(I) gene

and present a number of arguments for its identity.   First, the cloned DNA

sequence hybridizes specifically to human collagen α1(I) mRNA as shown by

hybrid arrest and hybrid selection experiments.   Second, the DNA sequence of

a segment of this gene is homologous to the corresponding sequence of calf

and chicken collagen α1(I).   It cannot be excluded that this gene is a

collagen pseudogene, but this seems unlikely since we have used cosH col.1
and the pPstl subclone in Southern blots of human DNA.   These probes detect
the DNA fragments predicted by the map of cosH col.1 (data not shown).   In
addition, the two terminal fragments of cosH col.1 are replaced by two new
EcoRI fragments.   Since the cosmid library was generated by partial digestion
of human DNA with MboI, the 5' and 3' terminal EcoRI fragments are truncated
(by MboI cleavage) in the cosmid clone.   From the data described here, we
conclude that the collagen α1(I) gene that we have cloned is large; it must
be at least 35kb as judged by our hybrid selection and Northern hybridization
analyses.   It is possible that the entire human collagen α1(I) gene is
present on this clone, since the 5' terminal fragment fails to hybridize to
collagen α1(I) mRNA.   This fragment, however, may represent an intron and
the remainder of the collagen α1(I) gene could lie outside the DNA regions
that we have cloned.   Further work will elucidate this matter.   Clearly
the fact that most of this large gene is present on this clone re-emphasizes
the advantage of cosmid cloning for large genes.

          Five exons of the human collagen α1(I) gene have been sequenced,
three of which are complete, and the general arrangement of this gene appears
similar to that previously described for the sheep and chicken α2(I) genes
(3-5).   Of the three complete exons, two are 54bp separated by one of 108bp.
Of the remaining exons not completely sequenced, one is larger than 54bp.
A unit of 54bp has been postulated as the size of the primordial collagen
gene and is based on the structure of the chicken α2(I) gene (6).   The exon
sizes so far determined in the human α1(I) gene correlate well with this hypo-
thesis, as does the mouse α1(I) gene (21).   The 108bp exon may have evolved
by the elimination of an intron and the subsequent fusion of two exons of
54bp.

          Comparison of the structure of the human α1(I) gene with the
chicken α2(I) and mouse α1(I) genes reveals similarities.   Particularly
interesting is the fact that in the human gene the length of the exons and
their position with respect to the protein sequence are the same, not only in
the mouse α1(I) gene (21), but also in the chicken α2(I) gene (4).   Thus, in
the chicken α2(I), mouse α1(I) and human α1(I) genes, amino acids 892-909 and
838-855 (exons no. 8 and 10 in ref 4) occur as two separate 54bp exons (Figs.
5 and 6 and refs. 4, 6 and 21).   Similarly, amino acids 856-891 in the
chicken α2(I) is thought to occur either as two 54bp exons or in one exon of
108 because this region has not been completely sequenced (4, 6).   In both
the mouse and human α1(I) genes these amino acids occur as an exon of 108bp.

Amino acids 910-945 are encoded in a 108bp exon in the chicken α2(I) (6).
This exon is incompletely sequenced in the human α1(I), but is larger than
54bp and could, therefore, be also 108bp.   The fact that the position and
sizes of these exons is the same in the α1(I) and α2(I) genes (at least for
those regions examined) is interesting, since these genes are believed to
have diverged some 700 million years ago (see ref. 7).   This result is remin-
escent of a similar conservation in the α- and β-globin genes which diverged
in a similar time period.   It is noteworthy that, in the region sequenced,
the human collagen α1(I) gene seems to have shorter introns than the chicken
collagen α2(I) gene.   A similar relatively compressed structure has also
been found in this region of the mouse α1(I) gene.   Since approximately 5%
of the gene is contained in 1kb of DNA, it may be expected that if the rest of
the gene were organised likewise the human α1(I) gene would be approximately
20kb in length.   However, the DBM selection and Northern blot analyses
(Fig. 7) suggest that the gene is at least 35kb long.   Since the exons
sequenced in the human gene lie in the 3' half of the gene and the density
of exons in this region of the α2(I) gene is higher than in the 5' half (4),
this apparent difference may not be so exaggerated.

Studies on the cDNA clones for the chicken proα(I) chain reveal
a clear predominance of the codons CGC for glycine, CCC for proline and GCC
for alanine (23).   In the mouse, collagen α1(I) gene, however, T is preferred
in the third position for these amino acids (21).   In the human α1(I) gene,
T is also preferred in the third position for alanine (8/11) but C(13/34) and
A(13/34) are equally preferred for glycine;   and T(7/15) and C(6/15) for
proline.   In the chicken α2(I) gene T is preferred for glycine, T and C for
alanine and there was little preference for proline (22).   It would appear,
therefore, that there is no clear evidence for a common interspecies or inter-
chain codon usage for the collagen α1(I) and α2(I) genes.

The isolation of a clone containing a substantial, if not comp-
lete, portion of the human collagen α1(I) gene has made possible the first
determination of some human collagen amino acid sequences.   In addition,
this clone makes possible the study of inherited diseases in man where the
functioning of this collagen gene is disturbed, as well as detailed studies
of collagen gene linkage.   The α1(I) collagen gene has previously been
assigned to chromosome 7 on the basis of phenotypic analysis (22).   Using
our DNA probes, we have recently confirmed this location.

REFERENCES
1.   Bornstein, P. and Sage, H.   (1980) Ann. Rev. Biochem. 49, 957-1003.
2.   Cheah, K. S. E. and Grant, M. E.   (1982) In: Collagen in Health
        and Disease.   Weiss, J. B. and Jayson, M. I.V. (Eds.) Churchill
        Livingstone, Edinburgh, in press.
3.   Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V. E., Sullivan, M.,
        Pastan, I. and deCrombrugghe, B. (1980).   Proc. Natl. Acad. Sci.
        U.S.A., 77, 7059-7063.
4.   Wozney, J., Hanahan, D., Tate, V., Boedtker, H. and Doty, P. (1981)
        Nature 294, 129-135.
5.   Schafer, M. P., Boyd, G. D., Tolstoshev, P. and Crystal, R. G. (1980)
        Nucl. Acids Res. 8, 2241-2253.
6.   Yamada, Y., Avvedimento, V. E., Mudryj, M., Ohkubo, H., Vogeli, G.,
        Irani, M., Pastan, I. and deCrombrugghe, B. (1980) Cell 21,
        689-696.
7.   Solomon, E. and Cheah, K. S. E. (1981) Nature 291, 450-451.
8.   Cheah, K. S. E., Grant, M. E. and Jackson, D. S.   (1979) Biochem.
        Biophys. Res. Comm. 91, 1025-1031.
9.   Pelham, H. R. B. and Jackson, R. J. (1976) Eur. J. Biochem. 67,
        249-256.
10.   Cheah, K. S. E., Grant, M. E. and Jackson, D. S. (1979) Biochem. J.
        182, 81-92.
11.   Cleveland, D. W., Lopata, M. A., MacDonald, R. J., Cowan, N. J.,
        Rutter, W. J. and Kirscher, M. W. (1980) Cell 20, 95-105.
12.   Alwine, J. C., Kemp, D. J. and Stark, G. R. (1977) Proc. Natl. Acad.
        Sci. U.S.A., 74, 5350–5354.
13.   Goldberg, M. L., Lifton, R. P., Stark, G. R. and William, J. G. (1980)
        In: Methods in Enzymology, vol. 68, Wu, R. (Ed.), Academic Press,
        New York, pp. 207-220.
14.   Crain, W. R. Jr., Durica, D. S. and van Doren, K. (1981) Mol. Cell
        Biol. 1, 711-720.
15.   Grosveld, F. G., Dahl, H. H. M., deBoer, E. and Flavell, R. A. (1981)
        Gene 13, 227-237.
16.   Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F.
        and Boedtker, H. (1979) Biochemistry 18, 3146-3152.
17.   Fietzek, P. P. and Kühn, K. (1976) In: Int. Rev. Connective Tissue

Res. $\underline{7}$, 1-60.
18. Galloway, D. (1982) In: Collagen in Health and Disease. Weiss, J. B. and Jayson, M. I. V. (Eds.) Churchill Livingstone, Edinburgh, in press.
19. Rave, N., Crkvenjakov, R. and Boedtker, H. (1979) Nucl. Acids Res. $\underline{6}$, 3559-3568.
20. Adams, S. L., Alwine, J. C., deCrombrugghe, B. and Pastan, I. (1979) J. Biol. Chem. $\underline{254}$, 4935-4938.
21. Monson, J. M. and McCarthy, B. J. (1981) Recombinant DNA $\underline{1}$, in press.
22. Sykes, B. and Solomon, E. (1978) Nature $\underline{272}$, 548-549.
23. Fuller, F. and Boedtker, H. (1981) Biochem. $\underline{20}$, 996-1006.
24. Maxam, A. M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. U.S.A. $\underline{74}$, 560-546.
25. Seyer, J. M. and Kang, A. H. (1981) Biochem. $\underline{20}$, 2621-2627.