Cloning and characterization of five overlapping cDNAs specific for the human proα1(I) collagen chain

Mon-Li Chu, Jeanne C.Myers, Michael P.Bernard, Juy-Fang Ding and Francesco Ramirez

Departments of Biochemistry and Obstetrics and Gynecology, University of Medicine and Dentistry of New Jersey-Rutgers Medical School, Piscataway, NJ 08854, USA

Received 6 July 1982; Revised and Accepted 13 September 1982

ABSTRACT

We report the cloning of five overlapping cDNAs bearing sequences specific for the human proα1(I) collagen chain. Poly-A RNA enriched for collagen sequences was purified from normal human fibroblasts and used as template to synthesize double stranded cDNA. The cDNA was inserted into the Eco RI site of pBR 322 by blunt-ending and dG:dC tailing. The clones were screened by colony hybridization using the original RNA population and the resulting five positive clones subjected to restriction endonuclease mapping analysis and DNA sequencing. These overlapping clones cover from residue 247 in the α chain to part of the 3' end untranslated region of the proα1(I) mRNA for a total of 3400 nucleotides.

INTRODUCTION

Type I collagen is the most abundant among the five different types of collagen identified up to date. This protein found in skin, bones and tendon is a trimer composed of two α1(I) chains and one α2(I) chain. These polypeptides are synthesized as precursor molecules (procollagen) with N-terminal and C-terminal propeptides which are cleaved extracellularly after the secretion of the assembled trimeric protein (1). At least three different groups of inherited connective tissue disorders have been directly associated with the biosynthesis of Type I collagen: Osteogenesis Imperfecta, Marfan and Ehlers-Danlos syndrome. The biochemical evidence suggests that this spectrum of altered phenotypes may have been caused at different steps of the biosynthesis of collagen either involving the expression of the constitutive genes or of the enzymes responsible for the post-translational modifications of the protein (2).

In order to investigate the several questions related to the structure and function of this important class of protein in man,

0305-1048/82/1019-5925$ 2.00/0

we have recently cloned a cDNA bearing specific sequences for the human proα2(I) chain (3). We have used this probe for the chromosomal assignment of the corresponding gene and found it to be localized on chromosome 7 (4) at about 7q22 (Henderson, A., et al., manuscript in preparation). Furthermore, we have isolated from a phage library overlapping genomic clones covering almost the entire human proα2(I) gene (5). In this paper we report the isolation and characterization of five overlapping cDNAs specific for the human proα1(I) chain, the other polypeptide of Type I. These clones (Hf164, Hf404, Hf590, Hf677, Hf784) span from amino acid residue 247 in the α chain to the 3' end untranslated region of the proα1(I) mRNA covering, therefore, more than 75% of the sequences of the human α1(I) chain.

MATERIALS AND METHODS

Cloning Procedure

The construction, identification and characterization of the proα1(I) cDNA clones was performed essentially as previously reported for the cloning of the proα2(I) cDNA (Hf32) (3). All experiments were carried out according to the NIH guidelines on Recombinant DNA Research.

Nucleic Acid Purification, Blotting and DNA Sequencing

RNA and DNA purification, Southern and Northern hybridization were performed as previously described (3). DNA sequencing was performed according to the chemical modification procedure of Maxam and Gilbert (6).

RESULTS AND DISCUSSION

Total poly-A RNA was isolated from cultured normal fibroblasts by proteinase K treatment and oligo(dT) cellulose chromatography. The RNA was enriched for collagen sequences by fractionation on a methyl-mercury sucrose gradient (3). The degree of enrichment was estimated by gel electrophoresis of the translation product from a rabbit reticulocyte lysate (7) with and without prior treatment with bacterial collagenase (8). The RNA was used as template in the presence of purified AMV reverse transcriptase (9) to synthesize double stranded cDNA which was inserted in the Eco RI site of pBR 322. This was achieved by

tailing 50 ng of cDNA with oligo(dG) and annealing it to 100ng of pBR 322, digested with Eco RI, blunt ended with reverse transcriptase and tailed with oligo (dC). The chimeric molecules were used to transform the E. coli strain RR1 in the presence of $MnCl_2$ and $CaCl_2$ (10). Two thousand recombinants were obtained with a transformation efficiency of 0.2%. Eight hundred clones were grown on nitrocellulose filters laid on agar plates followed by in situ chloramphenicol amplification prior to the screening (11). The colonies were lysed by treatment with 0.5M NaOH followed by neutralization with 1M Tris pH 8.0, 1.5M NaCl. The filters were air dried, baked for one hour in a vacuum oven and hybridized against 5ng/ml of cDNA (specific activity $7x10^7$ cpm/μg) synthesized from the original collagen enriched poly-A RNA population. The resulting positive clones were grown, the DNA purified and used as labelled probe to hybridize against Northern blotted total poly-A RNA from normal fibroblasts. Using this approach, we were able to estimate the size of the RNA sequences complimentary to the positive cDNAs and from the intensity of the hybridizing bands their quantitative representation in the RNA population. In Figure 1 is shown the pattern obtained with Hf32, the proα2(I) probe. Using this cDNA, two major bands were seen identifying two mRNAs respectively 6200 and 5700 nucleotides in length and a minor 5500 nucleotides in length. The same pattern was observed when polysomal RNA was used in the Northern blotting analysis excluding the possibility that they represent preferential accumulation of precursor molecules.
The isolation and detailed analysis of the human proα2(I) collagen gene has now demonstrated that these multiple transcripts are encoded from the same gene and vary for the length of their 3' end untranslated region (5, 12, 13, Myers, J. C., et al., manuscript in preparation). Five of the positive clones under investigation hybridized in the size range predicted for a collagen mRNA and with an intensity equal to the proα2(I) mRNA. In Figure 1 is shown the pattern of one of these clones (Hf164). Like the proα2(I) probe, these clones hybridize with a multiple pattern which identifies two major mRNA species 7200 and 5900 nucleotides in length present in both total poly-A and polysomal RNA. This result suggests that also the proα1(I) gene transcri-

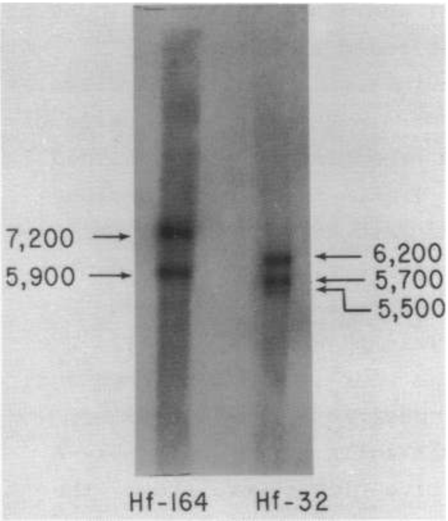

Figure 1. Northern blotting analysis of total poly-A RNA isolated from human fibroblasts and hybridized to ^{32}P labelled Hf164 (left) and Hf32 (right). RNA markers were run on parallel slot and visualized by ethidium bromide staining.

bes more than one mRNA and we now have genetic evidence supporting this idea (13). It should be noted that the size of the proα1(I) and proα2(I) mRNAs is clearly different, indicating no cross-hybridization between the two species. The five clones (Hf164, Hf404, Hf590, Hf677, Hf784) were analyzed by restriction endonuclease mapping and Southern blotting. Using these two criteria, several unique sites were found to be in common and a map of the overlapping segments of the five clones was derived (Figure 2). The Eco RI site in pBR 322 was restored in all the clones after insertion of the target cDNA into the vector. The size of the clones was variable: 0.4 kb (Hf164), 0.5 kb (Hf590), 1.0 kb (Hf784), 1.8 kb (Hf404 and Hf677).

The final evidence that these clones were indeed overlapping and encoding for the proα1(I) chain was derived from direct DNA sequencing. Selected sections of the clones were sequenced and the data obtained confirmed the scheme depicted in Figure 2. Unfortunately, very little is known about the primary structure of the human proα1(I) chain, but there is a high degree of similarity between species for the homologous chain (14). Thus, we compared our sequences with those derived from the protein analysis of the calf proα1(I) chain (15) and of a chick proα1(I) cDNA (pCg 54) (16). In figures 3 and 4 are shown the data obtained from the sequencing analysis of 5' and 3' ends of the two largest

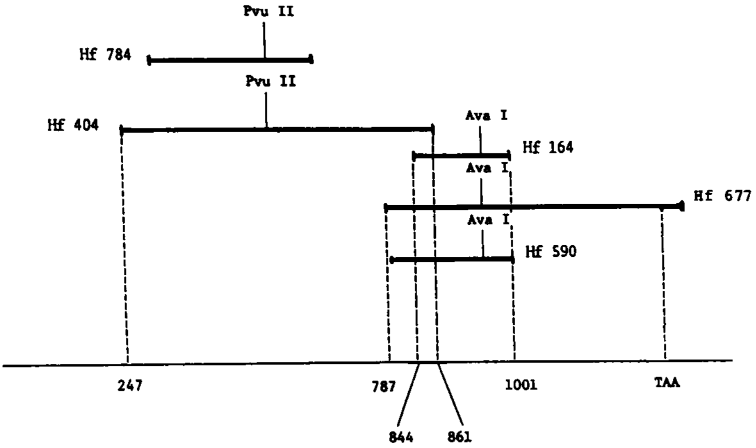

Figure 2. Composite map of the overlaps of the proα1(I) clones. The numbers at the bottom indicate the relative position of the clones with respect to the amino acid residues of the α1(I) chain. Also indicated is the position of the termination codon (TAA).

clones (Hf404, Hf677). The derived amino acid sequences are compared, when possible, with the homologous chain from the other two species. Hf404 extends from residue 247 to residue 861 in the α chain. Hf677 spans from residue 787 to 270 nucleotides into the 3' end untranslated region of the proα1(I) mRNA. The two clones, therefore, overlap for 222 nucleotides and together cover roughly 3400 nucleotides of the proα1(I) mRNA, including 75% of the α chain, the telopeptide and the C-terminal propeptide. A more detailed map of the two clones is shown in Figure 5. Several points should be noted. First, there is a very high degree of similarity between the amino acid sequences of the human, calf and chicken proα1(I) chain (90%); the same homology is observed also at the nucleotide level in the coding sequences between the human and the avian gene. Second, this similarity in the proα1(I) mRNA dramatically changes after the termination codon (TAA), unlike the human and chick proα2(I) mRNA which show a high degree of conservation of the nucleotide sequences also in the 3' end untranslated region of the mRNA (17). The 3' end untranslated region of the chick proα1(I) mRNA is 73% AT-rich, while the same portion of the human proα1(I) mRNA is only 32%

```
   |--a-->       247         250
         Hf 404 [GGC CCT CCT GGT CCC AAG GGT AAC AGC GGT GAA CCT GGT GCT CCT GGC AGC AAA GGA GAC ACT GGT
                [Gly Pro Pro Gly Pro Lys Gly Asn Ser Gly Glu Pro Gly Ala Pro Gly Ser Lys Gly Asp Thr Gly
 Calf  proα1(I)  Gly Pro Pro Gly Pro Lys Gly Asn Ser Gly Glu Pro Gly Ala Pro Gly Asn*Lys Gly Asp Thr Gly

                                                             280
                [GCT AAG GGA GAG CCT GGC CCT GTT GGT GTT CAA GGA CCC CCT GGC CCT GCT GGA GAG GAA GGA AAG
                [Ala Lys Gly Glu Pro Gly Pro Val Gly Val Gln Gly Pro Pro Gly Pro Ala Gly Glu Glu Gly Lys
                 Ala Lys Gly Glu Pro Gly Pro Thr*Gly Ile*Gln Gly Pro Pro Gly Ala*Ala Gly Glu Glu Gly Lys

                                                     300                                             312
                [CGA GGA GCT CGA GGT GAA CCC GGA CCC ACT GGC CTG CCC GGA CCC CTT GGC GAG CGT GGT GGA CTT
                [Arg Gly Ala Arg Gly Glu Pro Gly Pro Thr Gly Leu Pro Gly Pro Leu Gly Glu Arg Gly Gly Leu
                 Arg Gly Ala Arg Gly Glu Pro Gly Pro Ser*Gly Leu Pro Gly Pro Pro*Gly Glu Arg Gly Gly Pro*

   <--b--|       821                             830                                     840
         Hf 404 [CCC ATG GGC CCC CCT GGA TTG GCT GGA CCC CCT GGT GAA TCT GGA CGT GAG GGG GCT CCT GGT GCC
                [Pro Met Gly Pro Pro Gly Leu Ala Gly Pro Pro Gly Glu Ser Gly Arg Glu Gly Ala Pro Gly Ala
 Calf  proα1(I)  Pro Met Gly Pro Pro Gly Leu Ala Gly Pro Pro Gly Glu Ser Gly Arg Glu Gly Ala Pro Gly Ala
 Chick proα1(I)  Pro Met Gly Pro Pro Gly Leu Ala Gly Pro Pro Gly Glu Ala*Gly Arg Glu Gly Ala Pro Gly Ala

                         845                                                         860
                [GAA GGT TCC CCT GGA CGA GAC GGT TCT CCT GGC GCC AAG GGT GAC CGT GGT GAG ACC
                [Glu Gly Ser Pro Gly Arg Asp Gly Ser Pro Gly Ala Lys Gly Asp Arg Gly Glu Thr
                 Glu Gly Ser Pro Gly Arg Asp Gly Ser Pro Gly Ala Lys Gly Asp Arg Gly Glu Thr
                 Glu Gly Ala*Pro Gly Arg Asp Gly Ala*Ala*Gly Pro*Lys Gly Asp Arg Gly Glu Thr
```

Figure 3. DNA sequencing of the $\xmapsto{a}$ and $\xleftarrow{b}|$ fragment of Hf404. The amino acid sequences derived are shown in comparison with the known primary structure of the α1(I) chains of calf (15) and chicken (16). The numbers indicate the amino acid residues of the α1(I) chain. The asterisks indicate the residues differing from those present in the human proα1(I). The sequences underlined indicate the sequenced overlaps between Hf404 and Hf677.

AT-rich. Third, it has been found in both species that in the proα2(I) mRNA, there is a preference for U in the third base position of the codon for Gly, Pro and Ala (16, 17). In the chick proα1(I) mRNA, there is preference of U for Gly, C for Pro and equal percentage of C and U for Ala in the third nucleotide (16). In the human proα1(I) mRNA U is the preferred nucleotide for both Gly and Ala, while the Pro codon has an almost identical third nucleotide preference for C and U. Overall, the five clones extend from residue 247 in the α chain domain of the protein to the 3' end untranslated region of the messenger RNA, covering a total of 3400 nucleotides and they provide the starting material for the determination of the still unknown primary structure of the human proα1(I) collagen. It is interesting to note that the large size of the two proα1(I) transcripts implies that the untranslated regions account for 20 to 40% of the entire

```
|—c→            787                                                 800

Hf 677         [GGT CAG AGA GGA GAG AGA GGC TTC CCT GGT CCT CTT GGC CCC TCT GGT GAA CCT GGC AAA
               [Gly Gln Arg Gly Glu Arg Gly Phe Pro Gly Pro Leu Gly Pro Ser Gly Glu Pro Gly Lys
Calf  proαl(I)  Gly Gln Arg Gly Glu Arg Gly Phe Pro Gly Leu*Pro*Gly Pro Ser Gly Glu Pro Gly Lys
Chick proαl(I)   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -

                                                                    820

               [CAA GGT CCC TCT GGA GCA AGT GGT GAA CGT GGT CCC CCC GGT CCC ATG GGC CCC CCT GGA
               [Gln Gly Pro Ser Gly Ala Ser Gly Glu Arg Gly Pro Pro Gly Pro Met Gly Pro Pro Gly
                Gln Gly Pro Ser Gly Ala Ser Gly Glu Arg Gly Pro Pro Gly Pro Met Gly Pro Pro Gly
                 -   -   -   -   -   -   -  Gly Glu Arg Gly Pro Pro Gly Pro Met Gly Pro Pro Gly

                                                            838                             846

               [TTG GCT GGA CCC CCT GGT GAA TCT GGA CGT GAG GGG GCT CCT GGT GCC GAA GGT TCC CCT
               [Leu Ala Gly Pro Pro Gly Glu Ser Gly Arg Glu Gly Ala Pro Gly Ala Glu Gly Ser Pro
                Leu Ala Gly Pro Pro Gly Glu Ser Gly Arg Glu Gly Ala Pro Gly Ala Glu Gly Ser Pro
                Leu Ala Gly Pro Pro Gly Glu Ala*Gly Arg Glu Gly Ala Pro Gly Ala Glu Gly Ala*Pro


|—d→

Hf 677         [GAA TTC GGC TTC GAC GTT GGC CCT GTC TGC TTC CTG TAA ACT CCC TCC ATC CCA ACT TGG
               [Glu Phe Gly Phe Asp Val Gly Pro Val Cys Phe Leu  -   -   -   -   -   -   -   -
Chick proαl(I)  Glu Phe Gly Phe Asp Val Gly Pro Val Cys Phe Leu TAA ACA GGA AAA AAA AAG AAA AAG

                CTC CCT CCC ACC CAA CCA ACT TTC CCC CCA ACC CGG AAC AGA CAA GCA ACC CAA ACT GAA
                 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
                AAA AGA AAA AAA AAA AAA AAA AAG CCC CCC CAA CGC GTG ACA GGA GAG AGT AAT AAT TAT

                CCC CCT CAA AAG C
                 -   -   -   -  -
                AAT AAT TAA TAA A
```

Figure 4. DNA sequencing of the $\xmapsto{c}$ and $\xmapsto{d}$ fragment of Hf677. The amino acid sequences derived are shown in comparison with the known primary structure of the αl(I) chains of calf (15) and chicken (16). The numbers indicate the amino acid residues of the αl(I) chain. The asterisks indicate the residues differing from those present in the human proαl(I). The sequences underlined indicate the sequenced overlaps between Hf677 and Hf404, and the termination codon. Underlined (═════) is also the canonical sequence preceeding in the chick proαl(I) mRNA the poly-A addition site, not present in the same position in the human pro 1(I) mRNA.

mRNA. This peculiar feature has been already observed in other eucaryotic mRNAs, where the length of the non-coding regions equals or even exceeds that of the coding sequences (18, 19). The future isolation and analysis of the human proαl(I) collagen gene will allow us to determine the exact nature of the length heterogeneity of these monogenic transcripts.

The data obtained in these studies clearly show that the five overlapping clones encode for the human proαl(I) chain and exclude that the collagen αl-like gene recently isolated from a cosmid library by cross-hybridization is the human proαl(I) gene (20). The authors have also suggested that, according to their

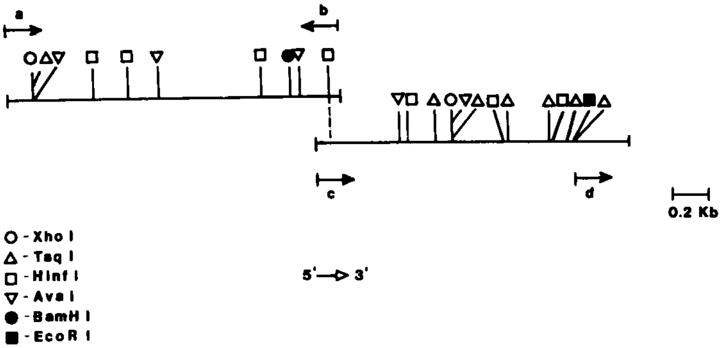

Figure 5. Restriction endonuclease map of Hf404 (left) and Hf677 (right). The symbols $\vdash\xrightarrow{a}$ $\vdash\xrightarrow{c}$ $\vdash\xrightarrow{d}$ indicate the 5' end label of sense strand used for DNA sequencing (Figure 4-5). The symbol $\xleftarrow{b}\dashv$ indicates the 3' end label of antisense strand used for DNA sequencing (Fig. 3-4). The sense of transcription is indicated by the arrow at the bottom of the figure. At the lower right hand corner in the figure is the scale reference expressed in kilobases.

preliminary data, this gene is located on chromosome 7 syntenic to the human proα2(I) gene. Similar experiments using our probes have unequivocably excluded this assignment and they have placed the human proα1(I) gene on chromosome 17 (21). Therefore, the observed nucleotide divergency of these two interdependent, coexpressed genes is probably a reflection of their differential localization in the genomic complement as is the case for the and β globin genes (22)

Sequences analysis of these cDNA clones have allowed for the first time the determination of some of the primary structure of the human proα1(I) chain. Comparative analysis with the other human and chick cDNAs will provide further information about the evolution of this important class of proteins. These probes will also make possible the future isolation of the corresponding gene from phage libraries as it has already been done for the collagen genes of other species (23-26). This is a necessary step for the understanding of the mechanisms involved in those human inherited disorders in which the proα1(I) collagen gene is believed to be directly involved.

ACKNOWLEDGEMENTS

We are very grateful to Dr. D. J. Prockop for his enthusiastic support and for many useful discussions and helpful suggestions. This work was supported by grants from NIH (AM-16516) and the Lalor Foundation.

REFERENCES

1. Prockop, D.J., Kivirikko, K.I., Tuderman, L. and Guzman, N. (1979) The New England Journal of Med. 301, 13-23.
2. McKusick, V.A. (1979) Heritable Disorders of Connective Tissue, Mosby, St. Louis, Missouri.
3. Myers, J.C., Chu, M.L., Faro, S.H., Clark, W.J., Prockop, D.J. and Ramirez, F. (1981) Proc. Natl. Acad. Sci. USA 78, 3516-3520.
4. Junien, C., Weil, D., Myers, J.C., Van Cong, N., Chu, M.L., Foubert, C., Gross, M.S., Prockop, D.J., Kaplan, J.C. and Ramirez, F. (1982) Am. J. Hum. Genet. 34, 381-387.
5. Myers, J.C., Sangiorgi, F.O., Chu, M.L., Ding, G., Prockop, D.J. and Ramirez, F. (1981) Feder. Proceed. 40, 1650.
6. Maxam, A. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560-569.
7. Pelham, H.R. and Jackson, R.J. (1976) Eur. Journ. Bioch. 67, 247-256.
8. Monson, J.M. and Goodman, H.M. (1978) Biochem. 17, 5172-5178.
9. Myers, J.C., Ramirez, F., Kacian, D.L., Flood, M. and Spiegelman, S. (1980) Anal. Bioch. 101, 88-96.
10. Peacock, S.L., McIver, C.M. and Monahan, J.J. (1981) Bioch. Biophys. Acta 655, 243-250.
11. Hanahan, D. and Meselson, M. (1980) Gene 10, 63-67.
12. Myers, J.C., Chu, M.L. and Ramirez, F. (1982) J. Cell. Bioch. 6, 318.
13. Chu, M.L., Myers, J.C. and Ramirez, F. (1982) Feder. Proceed. 41, 852.
14. Bornstein, P. and Sage, H. (1980) Ann. Rev. Biochem. 49, 957-1003.
15. Hofmann, H., Fietzek, P.P. and Kuhn, K. (1978) J. Mol. Biol. 125, 137-165.
16. Fuller, F. and Boedtker, H. (1981) Biochemistry 20, 996-1006.
17. Bernard, M.P., Myers, J.C., Chu, M.L., Ramirez, F., Eikenberry, E.F., and Prockop, D.J. Biochemistry, in press.
18. Moormann, R.J.M., van der Velden, H.M.W., Dodemont, H.J., Andreoli, P.M., Bloemendal, H., and Schoenmakers, J.G.G. (1981) Nucl. Ac. Res. 9, 4813-4822.
19. Kakidani, H., Furutani, Y., Takahashi, H., Noda, M., Morimoto, Y., Hirose, T., Asai, M., Inayama, S., Nakanishi, S., and Numa, S. (1981) Nature 298, 245-249.
20. Weiss, E.H., Cheah, K.S.E., Grosveld, F.G., Dehl, H.H., Solomon, E. and Flavell, R.A. (1981) Nucl. Ac. Res. 10, 1981-1984.
21. Huerre, C., Junien, C., Weil, D., Chu, M.L., Morabito, M., Foubert, C., Myers, J.C., Van Cong, N., Gross, M.S., Prockop, D.J., Boue, A., Kaplan, J.D., de la Chapelle, A., and Ramirez, F., Proc. Natl. Acad. Sci. USA, in press.
22. Ramirez, F., Mears, J.G. and Bank, A. (1980) Mol. Cell. Bioch. 31, 133-145.

23. Ohkubo, H., Vogeli, G., Mudryi, M., Avvedimento, V.E., Sullivan, M., Pastan, I., and de Crombrugghe, B. (1980) Proc. Natl. Acad. Sci. USA 77, 7059-7063.
24. Boyd, C.D., Tolstoshev, P., Schafer, M., Trapnell, B.C., Coon, H.C., Kretschmer, P.J., Nienhuis, A.W. and Crystal, R.G. (1980) J. Biol. Chem. 255, 3212-3220.
25. Wozney, J., Hanahan, D., Tate, V., Boedtker, H., and Doty, P. (1981) Nature 294, 129-135.
26. Monson, J.M. and McCarthy, J.B. (1981) DNA 1, 59-69.