Complete characterization of the human IGF-I nucleotide sequence isolated from a newly constructed adult liver cDNA library

Y. Le Bouc[†], D. Dreyer, F. Jaeger, M. Binoux[†] and P. Sondermeyer*

Transgene S.A., 11 rue de Molsheim, 67000 Strasbourg and [†]*INSERM U. 142, Hôpital Trousseau, 75012 Paris, France*

Received 2 December 1985

A full-size cDNA sequence coding for insulin-like growth factor I (IGF-I) was isolated from a human liver library. For the construction of this bank, a new method was developed which anneals dG-tailed cDNA with a synthetic adaptor 5′-AATTCCCCCCCCCCC-3′ followed by ligation into the EcoRI site of a lambda immunity-insertion vector. Based on the sequence analysis of the complete IGF-I messenger we concluded that the protein is synthesized as a precursor containing a signal peptide of 22 or 25 amino acid residues. In addition, the sequence extended in the 3′-direction and showed the presence of multiple polyadenylation sites in the IGF-I message.

Insulin-like growth factor I *cDNA* *Cloning* *λ vector* *Adaptor molecule*

1. INTRODUCTION

The main purpose in developing this new method has been the cloning of the cDNA coding for human IGF-I which thus far, at least in our hands, has eluded the isolation probably due to the low abundance of the messenger RNA. Screening of a newly constructed liver cDNA bank resulted in the isolation of a series of IGF-I candidates with several interesting features concerning the sequence in the 5′- and 3′-region of the messenger RNA.

Current procedures [1,2] for the construction of cDNA libraries in λ vectors involve essential steps which can be technically demanding and tend to diminish both the cloning efficiency and the average length of cDNA inserts. The method described here has been an attempt to circumvent most of the enzymatic reactions by using a synthetic adaptor molecule 5′-AATTCCCCCCCCCCC-3′ which can be annealed with dG-tailed cDNA and subsequently ligated into the *Eco*RI site of a λ cI-insertion vector.

2. MATERIALS AND METHODS

2.1. *RNA isolation and cDNA synthesis*

RNA was extracted using the guanidium-salt procedure [3] and enriched in poly(A)$^+$ RNA on oligo(dT) cellulose. First and second strand synthesis and S$_1$-digestion were done as described [4,5].

2.2. *Construction and screening of cDNA libraries*

Terminal transferase and tailing buffer were obtained from BRL and used as indicated by the supplier adding 100 ng cDNA per 30 μl final reaction volume. Length of the tails was calculated from the trichloroacetic acid incorporation with [^{3}H]-dGTP after incubation at 37°C with 5 U enzyme for 3 min. Reaction was stopped by adding 5 mM EDTA and heating 5 min at 65°C. A typical annealing experiment started with 100 ng cDNA in 30 μl tailing buffer, to which 60 μl TEN (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 100 mM NaCl) and 10

* To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50 © 1986 Federation of European Biochemical Societies

µl of a 1 µM solution of the phosphorylated oligomer 5′-dAATTCCCCCCCCCCC-3′ were added. This oligonucleotide and subsequent ones used for screening the libraries were synthesized on a silica support following described procedures [6]. Annealing mixture was heated for 2 min at 65°C and incubated for 16 h at 50°C.

Ligations were done at 4°C for 6–24 h by adding an equal volume of ligation buffer (130 mM Tris-HCl, pH 7.6, 20 mM $MgCl_2$, 2 mM spermidine, 20 mM DTT, 1 mM ATP), 5 µg *Eco*RI-digested λNM607 [7] and 4 U T_4 DNA ligase (Boehringer). DNA was alcohol precipitated and dissolved in 10 µl TEN. In vitro packaging was done as described in [8,9]. Libraries were amplified on *E. coli* POP101 [10], run on a CsCl gradient and stored as such at 4°C. Recombinants were screened by plating 2×10^4 pfu per dish (∅130 mm) on a lawn of *E. coli* POP101 and preparing filter replicas according to Benton and Davis [11].

3. RESULTS

3.1. *cDNA synthesis and insertion into λNM607 (schematically presented in fig.1)*

The initial steps for the construction of cDNA libraries closely resemble previously described protocols [5,12]. A size fractionation of the double-stranded cDNA on sucrose gradients was always included and turned out to be essential for maximizing the average length of the inserts obtained after transfer into the λ vector. Ends of the double-stranded cDNA were extended with 10–15 dG-residues using terminal transferase as described in section 2.

To integrate the cDNA within the unique *Eco*RI restriction site in λNM607, a synthetic adaptor was annealed with the dG-tailed cDNA and subsequently ligated with the *Eco*RI cohesive ends of the λ arms, thereby restoring both *Eco*RI sites. Even a few missing base pairs between the 5′-end of the cDNA and the 3′-end of the adaptor will not affect the stability of the recombinant molecule which is held together by dG-dC stretches of at least 10 nucleotides.

Phage DNA was packaged in vitro and recombinants were positively selected by plating on *E. coli* PEP101 [10]. Average amounts of 1×10^5 recombinants per ng cDNA were obtained. Background levels after ligation of the vector in the presence of an excess adaptor but without cDNA never exceeded 2×10^4 pfu/µg of vector on *E. coli* POP101.

3.2. *cDNA sequence coding for the IGF-I precursor*

A human liver cDNA library was screened for IGF-I with an oligomer complementary to the region coding for amino acids 18–24 of the mature protein [13].

Twelve positive signals were detected after screening 400000 recombinants. Nine of them

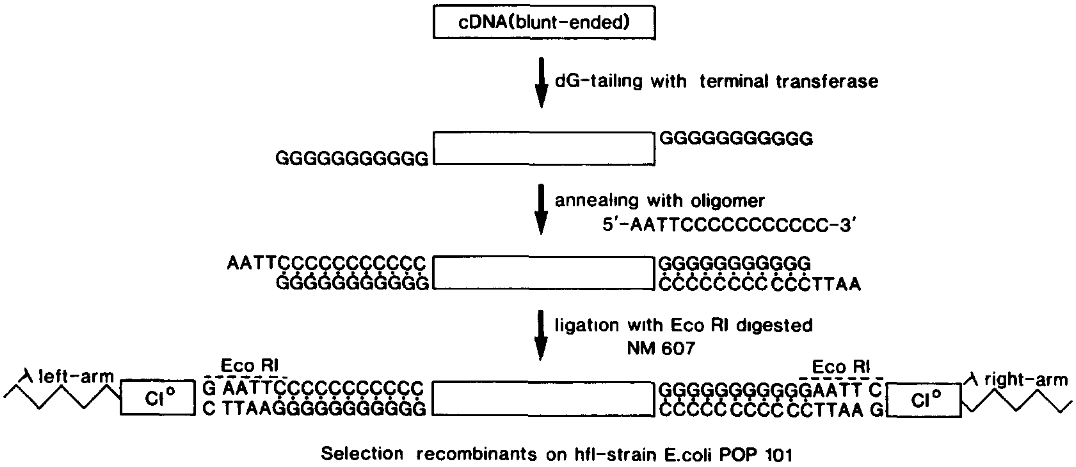
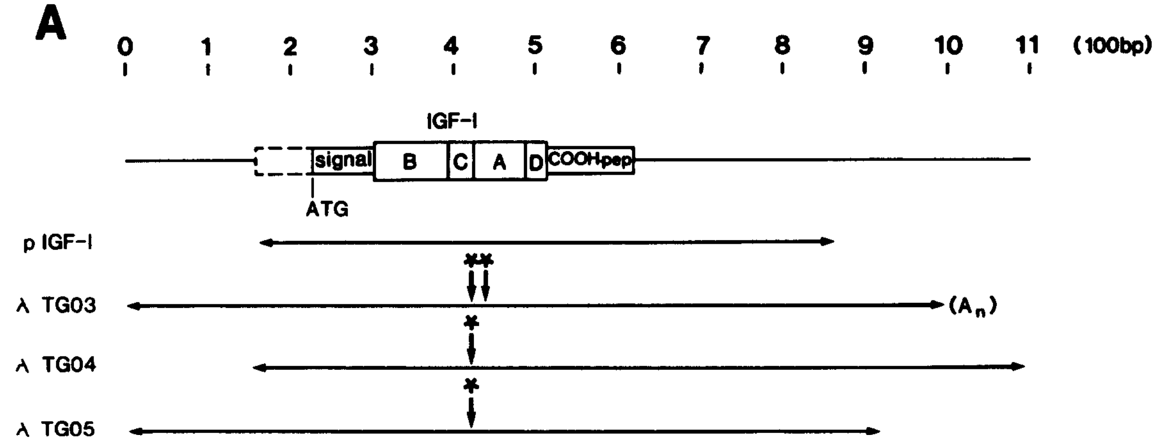


Fig.1. Schematical presentation of the tailing-adaptor procedure used for integration of blunt-ended cDNA into the *Eco*RI site of the cI gene of λNM607.

were confirmed by cross-hybridization in a later stage to contain IGF-I specific sequences. The average size of the *Eco*RI inserts of these candidates was 800 bp. After restriction analysis, 3 of them (λTG03, λTG04 and λTG05) were selected for further characterization by sequencing in M13 derivatives [14].

The sequence of clone λTG03, containing an *Eco*RI insert of 1073 bp, is given in fig.2. The 5′-end was extended with 139 bases compared to



B

```
                                                                   ┌→λTG05
                                                GAATTCCCCCCCCCCCCTTCAGAGCAGATAGAGCCTGCGCAATGG

45                   75                   105                  135                      ┌→λTG04
 AATAAAGTCCTCAAAATTGAAATGTGACATTGCTCTCAACATCTCCCATCTCTCTGGATTTCTTTTTGCTTCATTATTCCTGCTAACCAATTCATTTTCAGACTTTGTACTTCAGAAGCA

165                           195                           225
 ATG GGA AAA ATC AGC AGT CTT CCA ACC CAA TTA TTT AAG TGC TGC TTT TGT GAT TTC TTG AAG GTG AAG ATG CAC ACC ATG TCC TCC TCG
 Met Gly Lys Ile Ser Ser Leu Pro Thr Gln Leu Phe Lys Cys Cys Phe Cys Asp Phe Leu Lys Val Lys Met His Thr Met Ser Ser Ser
(-48)                                                                                      (-25)       (-22)
255                           285                           315
 CAT CTC TTC TAC CTG GCG CTG TGC CTG CTC ACC TTC ACC AGC TCT GCC ACG GCT GGA CCG GAG ACG CTC TGC GGG GCT GAG CTG GTG GAT
 His Leu Phe Tyr Leu Ala Leu Cys Leu Leu Thr Phe Thr Ser Ser Ala Thr Ala Gly Pro Glu Thr Leu Cys Gly Ala Glu Leu Val Asp
                                                                    (-1)(+1)
345                           375                           405                                                       *
 GCT CTT CAG TTC GTG TGT GGA GAC AGG GGC TTT TAT TTC AAC AAG CCC ACA GGG TAT GGC TCC AGC AGT CGG AGG GCG CCT CAG ACA GGC
 Ala Leu Gln Phe Val Cys Gly Asp Arg Gly Phe Tyr Phe Asn Lys Pro Thr Gly Tyr Gly Ser Ser Ser Arg Arg Ala Pro Gln Thr Gly

435                  *        465                           495
 ATC GTG GAT GAG TGC TGC TTC CGG AGC TGT GAT CTA AGG AGG CTG GAG ATG TAT TGC GCA CCC CTC AAG CCT GCC AAG TCA GCT CGC TCT
 Ile Val Asp Glu Cys Cys Phe Arg Ser Cys Asp Leu Arg Arg Leu Glu Met Tyr Cys Ala Pro Leu Lys Pro Ala Lys Ser Ala Arg Ser

525                           555                           585
 GTC CGT GCC CAG CGC CAC ACC GAC ATG CCC AAG ACC CAG AAG GAA GTA CAT TTG AAG AAC GCA AGT AGA GGG AGT GCA GGA AAC AAG AAC
 Val Arg Ala Gln Arg His Thr Asp Met Pro Lys Thr Gln Lys Glu Val His Leu Lys Asn Ala Ser Arg Gly Ser Ala Gly Asn Lys Asn

615                           645                           675
 TAC AGG ATG TAG GAA GAC CCT CCT GAG GAG TGA AGA GTG ACA TGC CAC CGC AGG ATC CTT TGC TCT GCA CGA GTT ACC TGT TAA ACT TTG
 Tyr Arg Met ***

705                  735                  765                  795
 GAACACCTACCAAAAAATAAGTTTGATAACATTTAAAAGATGGGCGTTTCCCCCAATGAAATACACAAGTAAACATTCCAACATTGTCTTTAGGAGTGATTTGCACCTTGCAAAAATGGT

825                  855                  885                  915
 CCTGGAGTTGGTAGATTGCTGTTGATCTTTTATCAATAATGTTCTATAGAAAAGAAAAAAAAAATTATATATATATATATATCTTAGTCCCTGCCTCTCAAGAGCCACAAATGCATGGGT

945                  975                  1005 ┌→ poly A- λTG03    1035
 GTTGTATAGATCCAGTTGCACTAAATTCCTCTCTGAATCTTGGCTGCTGGAGCCATTCATTCAGCAACCTTGTCTAAGTGGTTTATGAATTGTTTCCTTATTTGCACTTCTTTCTACACA

1065                 1095                 1125                 1155
 ACTCGGGCTGTTTGTTTTACAGTGTCTGAGGGGGGGGGGGAATTC
```

Fig.2. (A) Positions of the *Eco*RI inserts in λTG03, λTG04 and λTG05, relative to the putative mRNA coding for IGF-I. Also indicated is the previously reported sequence for this gene (pIGF-I [13]). Differences in comparison with this sequence are labelled with an asterisk. (B) The assembled sequence of the messenger coding for IGF-I as deduced from the data obtained with the independent isolates λTG03, λTG04 and λTG05. The 5′-end corresponds to the sequence found in λTG03 and λTG05 while the 3′-end corresponds to λTG04. The region coding for a mature IGF-I protein is underlined. The positions for some of the relevant amino acids are given in parentheses, number +1 corresponding to the N-terminal Gly-residue of the mature protein.

the previously published IGF-I sequence [13]. A stop signal at nucleotide position 69 essentially excluded any initiation of translation upstream of the 5′-proximal ATG at codon position −48.

The region coding for the precursor of IGF-I was found to be identical to the originally published data except for a silent T→C transition at position 434 present in at least 3 independent IGF-I candidates (see fig.2). In addition, λTG03 contained a single base mutation at position 453 resulting in a CTC-triplet coding for leucine instead of the original phenylalanine. However, this last variant was not confirmed when analyzing the sequence of λTG05 and it was therefore concluded that this mutation was introduced during the cDNA cloning procedure.

The most striking feature in the 3′-non-translated part of the sequence was the length of this region compared to previously reported results. In λTG03 we sequenced an additional 118 bp before running into a poly(A) track. The sequence of a second IGF-I clone λTG04, extended again 87 nucleotides further at the 3′-end and must even have continued in the original messenger RNA since no dA residues were found before the start of the dG-tail.

4. DISCUSSION

The results described here show the feasibility and reliability of this new cDNA cloning procedure, which has been a successful attempt to simplify the construction of cDNA banks. The method uses a synthetic adaptor 5′-AATTCCCC-CCCCCCC-3′ which can be annealed with dG-tailed cDNA and ligated at the 5′-end with the *Eco*RI cohesive ends of the λ vector. After in vivo packaging, recombinants can be selected positively on a hfl-strain *E. coli* POP101 due to insertional inactivation of the cI gene. The overall efficiency of this procedure, which varies between 20000 and 200000 recombinants per ng double-stranded cDNA, is at least as good as has been reported using classical linkers [2].

In addition to the advantages already indicated, this method generates inserts of a longer average size since many of the enzymatic incubations, such as methylation, linker-polymerization and restriction digests, are essentially avoided. Also, artefactual rearrangements due to the insertion of multiple cDNA fragments in a single vector molecule can easily be recognized in the sequence analysis by the presence of dG-tracks before and after each *Eco*RI restriction site. However, the major advantage turned out to be the very low level of background recombinants using a single-stranded and non-complementary adaptor molecule. The previous protocol based on linker-polymerization and cleavage with a restriction enzyme easily tends to generate high numbers of false positives due to traces of linker fragments present in the final ligation.

The isolation of several IGF-I cDNA clones indicates the efficiency of a λ vector compared to currently used plasmid cloning vehicles. From our results, the frequency of the IGF-I messenger in human liver tissue was estimated to be no more than 2×10^{-5} which corresponds with data reported before [13]. Sequencing of λTG03, containing an *Eco*RI insert of 1073 bp, extended the previously reported IGF-I cDNA at the 5′-end. The sequence of this region, which was confirmed by a second isolate λTG05, showed a TGA-stop at nucleotide position 69 (fig.2) and no additional ATG-codons to be considered as potential translation initiators. Therefore, we conclude that initiation must occur at 1 of 3 methionines at position −48, −25 or −22. Analyzing the hydrophobic pattern of the different putative signal sequences suggested [13] that the second methionine is the most likely position for translational initiation. This is further supported by the homology with the signal sequence of preprorelaxin, a protein which like IGF-I belongs to the insulin gene family [15]. However, a strong argument for initiation at the third methionine (position −22) was found when comparing the 5 preceding bases with the consensus sequence emerging from a systematic analysis of translational start sites in eukaryotic mRNAs [16]. The nucleotides in this region gave 4 out of 5 matches with the CCACC consensus while the sequence preceding the second methionine only corresponded at position −3.

Considering these arguments, it is tempting to speculate that in fact both ATG-codons at position −25 and −22 may have served as translation initiators leading to 2 types of IGF-I precursors with a small difference in the length of their signal peptides.

Comparison of the coding sequence with previously reported data revealed a polymorphism in the gene coding for IGF-I. A variant in the third base of the Gly-codon (position 42) was found in independent IGF-I cDNAs thereby excluding erroneous transcription during preparation of the cDNA.

With respect to the sequence in the 3′-non-translated region, a number of remarks can be made when comparing our data with the sequence as reported before [13]. First of all, 2 of the cDNA λTG03 and λTG04, contained an additional 118 bases. The previously indicated end of the message at position 888 (fig.2) indeed shows an almost contiguous poly(A) stretch of 15 nucleotides which is also found in the cDNA clones we described here. However, the sequence of both our cDNA clones continued further to the 3′-end in an identical way. We therefore assume that this A-rich region and presumptive end of the IGF-I message as proposed before [13], has served for the reverse transcriptase to initiate the oligo(dT) primed synthesis of the first cDNA strand. A more likely end for the IGF-I message was found in λTG03 which contained a poly(A) of 52 bases followed by 11 dG-residues (fig.2). However, the usual preceding AATAAA polyadenylation signal was not found except maybe from the ACTAAA sequence 37 bp before the poly(A). The absence of a strong polyadenylation signal may have been the reason why the sequence of a second candidate, λTG04, continues even further ending abruptly after an additional 87 bases. Systematic studies, using in vitro manipulated DNA [17], indeed showed that different positions for polyadenylation can be used during maturation of a messenger RNA. The gene coding for IGF-I turns out to be an example where the primary RNA transcript can be processed at multiple positions. Apparently, in the absence of distinct polyadenylation signals related sequences can be recognized with relatively low efficiency. Interestingly enough, comparable results have been reported with the RNA coding for IGF-II representing at least 4 species some of which may have differed considerably in the length of the 3′-non-translated region [18]. Whether these are partially processed intermediates of the primary transcript or whether they are actually translated cannot be decided yet.

ACKNOWLEDGEMENTS

The authors are grateful to the chemists of Transgene for supplying the synthetic oligonucleotides. We thank N. Monfrini and F. Daul for preparation, and P. Meulien for critical reading of the manuscript. Our many discussions with A. Pavirani have been very helpful and were highly appreciated by the authors. We thank J.P. Lecocq for continued interest in this work.

REFERENCES

[1] Young, R.A. and Davis, R.W. (1983) Science 222, 778–782.

[2] Schwarzbauer, J.E., Tamkin, J.W., Lemischka, I.R. and Hynes, R.O. (1983) Cell 35, 421–431.

[3] Chirgwin, J.M., Przbyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry 18, 5294–5299.

[4] Kupper, H., Keller, W., Kurz, C., Forss, S., Schaller, H., Frauze, R., Strohmaier, K., Marquardt, O., Zaslavsky, V.G. and Hofschneider, P.H. (1981) Nature 289, 555–559.

[5] Efstratiadis, A. and Villa-Komaroff, L. (1979) in: Genetic Engineering (Stelow, J.K. and Hollaender, A. eds) vol. I, p.15, Plenum, New York.

[6] Kohli, V., Balland, A., Wintzerith, M., Sauerwald, R., Staub, A. and Lecocq, J.P. (1982) Nucleotide Acids Res. 10, 7439–7448.

[7] Murray, N.E., Brammar, W.J. and Murray, K. (1977) Mol. Gen. Genet. 150, 53–61.

[8] Hohn, B. and Murray, K. (1977) Proc. Natl. Acad. Sci. USA 74, 3259–3262.

[9] Sternberg, N., Tiemeier, D. and Enquist, L. (1977) Gene 1, 225–280.

[10] Lathe, R. and Lecocq, J.P. (1977) Virology 83, 204–280.

[11] Benton, W.D. and Davis, R.W. (1977) Science 196, 180–182.

[12] Williams, J.G. (1981) in: Genetic Engineering (Williamson, R. ed.) vol I, p.2, Academic Press, New York.

[13] Jansen, M., Van Schaik, F.M.A., Ricker, A.T., Bullock, B., Woods, D.E., Gabbay, K.H., Nussbaum, A.L., Sussenbach, J.S. and Van de Brande, J.L. (1983) Nature 306, 609–611.

[14] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.

[15] Dayhoff, M.O. (1978) Atlas of Protein Sequence and Structure, National Biochemical Research Foundation.

[16] Kozak, M. (1984) Nucleotide Acids Res. 12, 857–872.

[17] Mason, P.J., Jones, M.B., Elkington, J.A. and Williams, J.G. (1985) EMBO J. 4, 205–211.

[18] Bento Soares, M., Ishii, D.N. and Efstratiadis, A. (1985) Nucleotide Acids Res. 13, 1119–1134.