J. Biochem. Mol. Biol. Vol. 28, No. 5 pp. 402~407 September 30, 1995

Analysis of Partial cDNA Sequence from Human Fetal Liver

Jae Wha Kim, Jae Chan Song, In Ae Lee, Younghee Lee, Myoung Soo Nam[§], Yoonsoo Hahn[1], Jae Hoon Chung[1] and In Seong Choe*

Molecular and Cellular Biology Research Group, Korea Research Institute of Bioscience and Biotechnology, KIST, Taejon 306-600, [1]Department of Biological Sciences, KAIST, Taejon 305-701, Korea

(Received April 29, 1994)

Abstract: Single-run Partial cDNA sequencing was conducted on 1,592 randomly selected human fetal liver cDNA clones of Korean origin to isolate novel genes related to liver functions. Each partial cDNA sequence determined was analyzed by comparing it with the databases, GenBank, Protein Information Resource (PIR) and SWISS-PROT Protein Sequence Data Bank. From a set of 1,592 cDNA clones reported here, 1,433 (90.0% of the total) were informative cDNA sequences. The other 159 clones were identified as DNA sequences which had originated from the cloning vector. Among 1,433 informative partial cDNA sequences, 851 (59.3%) clones were revealed to be identical to known human genes. These known genes have been classified into 225 different kinds of genes. In addition, 340 clones (23.7%) showed various degrees of homology to previously known human genes. Ninety four (6.6%) clones contained various repeated sequences. Twenty four (1.7%) partial cDNA sequences were found to have considerable homology to known genes from evolutionarily distant organism such as yeast, rice, Arabidopsis, mouse and rat, based on database matches, whereas 124 (8.7%) had no significant matches. Human homologues to functionally characterized genes from different organisms could be classified as candidates for novel human genes of similar functions. Information from the partial cDNA sequences in this study may facilitate the analysis of genes expressed in human fetal liver.

Key words: cDNA, human fetal liver, random sequencing.

The human genome is estimated to consist of 50,000 to 100,000 genes, only a few hundred of which have been characterized with respect to their structure and function. Several approaches have been adopted to isolate novel genes. Sequencing of cDNA, reverse transcribed from mRNA, is a very effective way to reveal genetic information, because the coding sequences of genes represent the vast majority of the information content of the genome, but is only 3% of the genomic DNA (Adams *et al.*, 1991). Recently, the random isolation of more than 3,000 partial cDNA sequences from a set of human brain cDNA libraries has been reported (Adams *et al.*, 1991; 1992). Partial sequences of cDNA clones are informative in analysis of gene function because putative amino acid sequences could be readily predicted from nucleotide sequences of cDNA. An active gene, *fosB*, was identified by this strategy (Itoh *et al.*, 1994). Partial sequences of cDNA are also useful in mapping large DNA fragments because a cDNA represents a single copy DNA (Burglin *et al.*, 1992). These unique partial cDNA sequences can be designed as expressed sequence tags (ESTs). ESTs can be localized along chromosomes and used for isolation of new genes (Sedlacek *et al.*, 1992).

The liver is one of the important organs in the human body and estimated to have approximately 500 different functions, mainly metabolic processing (Mathews *et al.*, 1990). Other than metabolic processing, haematopoiesis is an important and very active function in human fetal liver. Haematopoiesis is a continuous process that generally maintains a steady state in which the production of mature blood cells equals their loss. Steady-state levels described by haematopoiesis are maintained by cytokines, the agents of cell-cell communication (Jan Klein, 1990).

The construction of a large-scale collection of partial cDNA sequences from human fetal liver tissue and homology searches in databases would facilitate the discovery of novel genes of interest involved in various

[§]Present address: Faculty of Agriculture, Hokkaido University, Sapporo 060, Japan

*To whom correspondence should be addressed.
Tel: 82-42-860-4180, Fax: 82-42-860-4593

liver functions. Single-run sequencing of partial cDNA clones from human fetal liver has been initiated, but not enough data has been generated (Saito *et al.*, 1994; Huber *et al.*, 1993). To construct a collection of partial cDNA sequences, 1,592 partial cDNA sequences were randomly selected and their nucleotide sequences have been determined. In this paper, we report the result of analysis of a set of 1,433 informative partial cDNA sequences that could be identified and their functions predicted based on database matches.

Materials and Methods

Construction of a cDNA library

Total RNA was purified from liver tissue of a 26 week old human fetus of Korean origin by the acid guanidine phenol chloroform (AGPC) method described by Chomczynski *et al.* with minor modification (Chomczynski *et al.*, 1987). Cell lysis was performed with glass rod and mesh (No. 60). The integrity of total RNA prepared by this method was examined by formaldehyde agarose gel electrophoresis. Poly(A)$^+$ RNA was isolated from total RNA using oligo-dT cellulose affinity chromatography (Sambrook *et al.* 1989). The cDNA libraries were constructed using the ZAP-cDNA cloning kit (Stratagene, La Jolla, CA) according to the manual provided by the manufacturer (Stratagene). Five μg of poly(A)$^+$ RNA was primed with an oligo-dT primer for the synthesis of cDNA. After ligation of *Eco*RI adaptors to the cDNA, it was digested with *Xho*I, and finally ligated into an *Eco*RI, *Xho*I-cut λ ZAP vector. Ligated DNA was packaged *in vitro*, using a Gigapack II Gold packaging system (Stratagene). The cDNA library contained 8.5×10^7 primary plaques, and was amplified to a titer of 7×10^{10} pfu/ml. The cDNA library was stored in aliquots of 1 ml at 4℃.

Isolation of cDNA clones

Mass excision of phagemid from the ZAP express vector (Stratagene) was performed by the infection of 10^6 cells of XLOLR strain with 10^7 pfu of ExAssist helper phage. Single plaques were cored from the agar plate and transferred to each well of a 96-well microtiter plate containing 50 μl of LB medium with tetracycline. The microtiter plates were incubated overnight without shaking. The glycerol stock was made by adding an equal volume of 40% sterile glycerol into each well. Single strand DNA was isolated from the excised double strand phagemid by the infection of M13 helper phage according to the user's manual.

DNA sequencing

Dideoxy chain termination reactions were performed with ^{35}S-dATP (Amersham) and Sequenase Kit Version 2.0 (USB, Cleveland, USA). The oligonucleotide complementary to SK primer sequence (CGCTCTAGAACTAGTGGATC; Korea Biotech. Co., Taejon, Korea) was used to determine specifically the 5′-end nucleotide sequences of each cDNA clone. After sequencing reactions, samples were run on 6% polyacrylamide/7 M urea gel in 1×TBE buffer at constant voltage of 1800 V.

Computer analysis

Partial cDNA sequences were examined for similarities to the GenBank nucleic acid database release 82, 1994 using an IBM compatible computer interfaced via TCP/IP to a geri 4680 main computer. cDNA sequences were also translated in all six reading frames, and each translation was compared with the protein sequence database, Protein Information Resource (PIR) release 41, 1994 and SWISS-PROT release 29, 1994. GenBank, PIR and SWISS-PROT searches were conducted with modified programs for nucleotide (BLASTN) and peptide (BLASTX) comparisons which permitted many query sequences to be automatically searched. The BLAST programs (Altschul *et al.*, 1990) contain a rapid database searching algorithm that searches for local areas of similarity between two sequences and then extends the alignments on the basis of defined match and mismatch criteria. After the BLASTN search, cDNA sequences which were not matched or had low homology were compared against GenBank by FASTA to determine if significant matches were missed due to the lower sensitivity of BLASTN used in the database search.

Results and Discussion

Construction of the cDNA library and sequencing

In single-run cDNA sequencing, it is important to use a cDNA library of high quality, reflected in the number of independent clones, the integrity of base sequences and their size (Kim, *et al.*, 1993; Park, *et al.*, 1993; Hoog *et al.*, 1991). The cDNA library used in this experiment is composed of 8.5×10^7 independent clones. All of the 10^4 plaques of the λ ZAP XRII cDNA library except 8 were white plaques according to a color test using X-gal and IPTG. Among white plaques, ninety were randomly selected and tested for cDNA inserts. Two clones contained no inserts and one had an insert DNA shorter than 200 base pairs (bp). The size distribution of cDNA inserts cloned in vector DNAs ranged from 0.2 to 2.4 kb with an average of 0.7 kb. These results confirmed that over 96% of the cDNA inserts cloned could give useful information by random sequencing. cDNA inserts were unidirectionally

Table 1. Accuracy of nucleotide sequencing by chain termination reaction was tested by comparing the nucleotide sequences of the clones containing human serum albumin gene

No. of clones compared	Average length	Aligned bases	Mismatched and ambiguous bases	Deletions	Accuracy (%)
49	176	8624	62	3	99.2

cloned into a λ ZAP XRII cDNA cloning system to give nucleotide sequence information on the 5′-side of cDNA when the single-stranded cDNA inserts were sequenced. Nucleotide sequences from 90 to 289 with average of 176 bases were determined to carry out database searches. As the accuracy of query nucleotide sequences was an important factor for database search, we examined the accuracy by comparing the cDNA sequences from multiple copies of serum albumin gene cloned in this experiment with the nucleotide sequences of serum albumin from GenBank database. Three deletions, fifty seven mismatches and five ambiguous bases were found among the 8,624 nucleotide sequences compared, as shown in Table 1. The estimated error rate was below 1%. A total of 1,592 partial cDNA sequences were determined.

Classification of partial cDNA sequences

From a set of 1,592 partial cDNA clones sequenced in this experiment, 1,433 were informative complementary DNA sequences. The other 159 clones were identified to be vector DNA sequences or clones containing cDNA sequences shorter than 60 bp. Composition of single-run cDNA sequence data is shown in Table 2. On the basis of database searches, the 1,433 informative cDNA sequences were classified into five groups as shown in table 2 based on the criteria described below.

Eight hundreds fifty one of the sequences (59.3% of the informative clones) consisted of matches to previously known human DNA sequences and designated as identified clones. The clones in this group satisfied the following criteria. Probability specified by Altschul *et al.* (1990) was below 10^{-5}. Nucleic acid and/or amino acid sequences matched with more than 60% identity and the number of matched nucleotides were longer than 60 bases or 20 amino acids. Gaps were not introduced into database search processes. The clones with high homologies to house keeping genes including ribosomal genes, mitochondrial genes, and tissue specific genes, globin genes, the α-fetoprotein gene, the transferrin gene, and the fetal liver type cytochrome p-450 gene were included in this group. Four different kinds of globin genes were found in the cDNA library. The ratio between beta-globin and γ-globin genes was over 1 to 8 (17 clones: 145 clones) consistent with the fact of using fetal liver material for the construction of the cDNA library (Mathews, 1990). The partial cDNA sequences in this group which showed homologies between 60~90% as shown in Table 3 may need attention. They are expected to be useful in the isolation of novel genes that might have similar functions in liver tissue (Pearson *et al.*, 1988).

Table 2. Classification of sequences obtained by random sequencing according to the information by data base comparisons

Classification	No. of clones	Frequency (%)
Identified clone	851	59.3
Unidentified clone	340	23.7
Other	24	1.7
No matched clone	124	8.7
Repeat	94	6.6
Total	1433	100

Three hundreds fourty (23.7%) clones in this group had low degrees of homology to known human genes and were designated "unidentified clones". The homologies of the clones were less than 60%, or the length of matching nucleotide sequences was less than 60 base pairs. These clones should be distinguished from the other set of 124 (8.7%) cDNA sequences that had no significant matches at all. Because some proteins share very low homologies except evolutionary conserved consensus sequences, these cDNA sequences would well be compared with a protein motif database such as the PROSITE, to search for novel genes of specific liver functions (Koch *et al.*, 1991; Pawson *et al.*, 1992).

Twenty four clones (1.7%) matched non-human entries in GenBank and protein databases and were designated as "other species". cDNA sequences in this group showed high homologies of more than 70% with the cDNA sequences from other species including mouse A10 protein, rat developmentally regulated intestinal protein (OCDI-5) and rat ribosomal proteins L34 and L5. No human homologous sequence to these genes has yet been reported. These clones are possibly a basis for isolating new human genes that have the same or similar functions in the liver.

Ninety four clones (6.6%) matched repetitive elements of the human genome and were named "repeating sequences". Alu sequence, L1 sequence, di- and tri-nucleotide repeats were included in this group. The clones containing repeating sequences could not be characterized because the number of corresponding entries in the databases which included these repeating sequences was too large. For the characterization of these sequences,

Table 3. A list of cDNA clones showing homologies between 60 to 90% to known cDNA sequences of human origin

Clone I.D.	Locus	Nucleotide similarity	Gene
18C03	HUMHP604A	82/92 (89%)	chaperonin (HSP60)
27H05	HSMRNAOXY	70/103 (67%)	oxytocin receptor
25G11	S67309S1	50/77 (64%)	estrogen receptor
P02A06	HSANGG5	100/134 (74%)	angiotensinogen
3H11	HSCD59EX	66/96 (68%)	CD59
4A04	HUMCD19A	60/84 (71%)	CD19
24A11	HUMYB1A	80/140 (76%)	Y box binding protein-1
23H10	HSCPPOX	87/87 (89%)	coprox gene for coproporphyrinogen oxidase
24D03	HSHLAF	89/100 (89%)	HLA-F gene for human leukocyte antigen F
24A12	HUMCTI	106/119 (89%)	erythrocyte membrane protein
15H11	HUMHUSIIIA	40/46 (86%)	acrosin-trypsin inhibitor
25A11	HUMMYLCB	67/72 (86%)	non-muscle myosin alkali light chain
14F04	HSVMYCLC2	53/66 (80%)	ventricular myosin I
21C02	HSEWS	80/96 (83%)	EWS
20F01	T11257	130/145 (89%)	ADP-ribosylation factor 3
26E09	HSRR2SS	75/84 (89%)	ribonucleotide reductase small subunit
3H03	HUMRETPIGB	130/146 (89%)	retinal pigment epithelium
5F04	HSMP21HOM	125/152 (82%)	P21
19E07	HUMCPSI	132/138 (89%)	carbamyl phosphate synthetase I
9F08	HSSELPM	97/110 (88%)	selenoprotein P
13B05	HUMZNF7	85/110 (77%)	zinc-finger protein 7 (ZFP7)
18C09	HSJUND	109/142 (76%)	jun-D mRNA (onco gene)
27A05	HSG6PDGEN	74/103 (72%)	glucose-6-phosphate
22G04	HUMTOPPG2	50/71 (70%)	topoisomerase I pseudogene 2
10A07	HUMNPM	57/59 (72%)	nucleophosmin
10D04	HUMSERG	54/75 (72%)	serglycin
26E06	HSLYSOZY	40/64 (62%)	lysozyme
21D10	HUMMMTVPOL	57/84 (67%)	mammary tumor virus pol
18C11	S45332	57/85 (67%)	erythropoietin receptor
P01F07	HSGSA1R	91/122 (74%)	coupling protein G(s) alpha subunit (alpha-S1)
P01A10	HSEAP	87/97 (89%)	Epstein-Barr virus small RNAs (EBERs)
P03H01	HSU02310	61/75 (81%)	fork head domain (FKHR)
P03C09	HUMANK	78/87 (89%)	erythroid ankyrin
23G01	HSTNFABX	43/62 (69%)	tumor necrosis factor

the repeating sequence was removed from each cDNA partial sequence determined and the remaining nucleotide sequence of each partial cDNA sequence was subjected to database search. Only 14 clones (15%) out of 94 sequences gave further information needed for the characterization (data not shown). The nucleotide sequences of the 3'-end of the cDNA inserts may give useful information for the identification of the clones in this group. 124 clones (8.7%) did not have any significant matches at all and were designated "no match".

Redundancy of partial cDNA sequences

In previous complementary DNA sequencing studies, redundancies of cDNA sequences represented by partial cDNA sequences have been reported. The most abundant clones in human brain cDNA libraries were beta-actin (0.6%) and myelin basic protein genes (0.5%) as reported by Adams *et al.* (1991 and 1992). Cytochrome b and elongation factor 1α genes and the groups of cDNA sequences that formed contigs of both genes constituted 11% of a mouse testis cDNA library (Hoog *et al.*, 1991). The redundancy of cDNA sequences in this study are shown in Table 5. Globin genes (17.4% of total clones) were the most common cDNA clones. Globin genes were composed of α-globin (6.0%), β-globin (1.2%), γ-globin (10.1%) and δ-globin (0.1%) genes. The abundant presence of globin messages in liver tissue was expected because four major alternate forms

Table 4. A list of cDNA clones which show various degrees of homologies to the nucleotide sequences of other species than human

Clone I.D.	Locus	Nucleotide similarity	Gene
18H10	MMEIF4AII	121/151 (80%)	*M. musculus* eIF-4AII
17G05	MMP311AA	60/70 (85%)	*M. musculus* P311
5A03	PFASANTM	48/54 (78%)	*P. yoelii* merozoite surface antigen
14D12	MUSNUABPRO	73/96 (76%)	*Mus musculus* nucleic acid binding protein (hnRNP X)
7F01	DDIPRKNGPK	48/72 (66%)	Dictyostelium discoideum protein kinase
18B11	MUSTCTEX	58/67 (87%)	Mouse tctex-1
18H05	IRRPL35	43/67 (64%)	Rat ribosomal protein L35
16F05	YSCMTCG	52/90 (57%)	*S. cerevisiae* mitochondrion DNA
20B05	BTMLRQSMR	76/97 (78%)	*B. taurus* MLRQ subunit of NADH
20G07	DDCOXV	58/98 (59%)	*D. discoideum* coxV gene for cytochrome c
24F08	MUSSLPSEXA	43/62 (69%)	Mouse sex-limited protein (SlpA)
26A04	RNMRLCB	63/69 (91%)	Rat smooth muscle myosin RLC-B
23E09	RRPS19	72/75 (96%)	Rat ribosomal protein S19
20B12	RRRPL34	138/151 (91%)	Rat ribosomal protein L34
2G03	RATRPL35AA	89/100 (89%)	Rat 60S ribosomal subunit protein L35
21B11	RNUNR	93/105 (85%)	Rat unr
P01E11	RNRPL5	49/55 (89%)	Rat ribosomal protein L5
P01D06	YSCMTCG	55/87 (63%)	*S.cerevisiae* mitochondrion DNA
P01E04	MMU05264	63/91 (69%)	Mus musculus C3H gp49B
P03C11	XELRNPA3A	161/166 (96%)	*Xenopus laevis* ribonucleoprotein
P02A08	MUSKA10X	57/65 (87%)	Mouse A10
P02D02	RNRPS14	69/75 (92%)	Rat ribosomal protein S14
P02E06	RATPRSTTNC	67/100 (67%)	Rat prostatein C3 subunit
P03D05	RATOCI5	77/91 (84%)	Rat developmentally regulated intestinal protein

Table 5. Redunancy of nucleotide sequences of the clones selected randomly from fetal liver cDNA library

Redundancy	Kinds of genes	No. of clones	Proportion (%)
1	136	136	15.9
2	16	32	3.8
3	12	36	4.3
5	1	5	0.6
6	1	6	0.7
8	4	32	3.8
11	3	33	3.9
13	1	13	1.5
17	1	17	2.0
35	1	35	4.1
86	1	86	10.1
108	1	108	12.7
145	1	145	17.0
	179	684	80.4
Mitochondrial gene	11	117	13.7
Ribosomal gene	35	50	5.9
	225	851	100

of globin have been displayed according to the developmental stages of fetus. Another common clone was the albumin gene (7.5%).

851 of the putatively identified clones were classified into 225 different kinds of genes and the genes having cDNA redundancies of more than 3 numbered 663 (46.2%). In any random cDNA sequencing study, clone redundancy is a major problem in identifying novel genes(Adams *et al*,. 1991). Several strategies to reduce redundancy have been probed. Keith *et al.* (1993) have constructed a cDNA library from mRNAs that were selected for low redundancy according to the developmental stages. As another strategy, Hoog tried a differential cDNA screening. A mixture of labeled cDNA prepared from four different tissues was used as a probe to mark the abundantly expressed cDNA clones, and the clones that failed to hybridize were selected for sequencing. With this strategy, the probability of finding new genes increased up to 7 times that of random sequencing (Hoog *et al.*, 1991).

In addition to the above two strategies, an equalized cDNA library would be useful to increase the probability of finding new genes by cDNA random sequencing

(Ko, 1990; Sasaki, *et al.*, 1994). In this case the redundancy would be reduced because each gene would be represented by equal frequency in cDNA library.

With enough partial cDNA sequences from fetal liver tissue of Korean origin, which can be utilized as hybridization probes, at hand, it is possible to facilitate searching out novel genes by cDNA random sequencing. cDNA clones that have homologies to the partial cDNA sequences which are highly or moderately represented in the set of 1,592 cDNA sequences could be eliminated through a prescreening process using Southern hybridization.

References

Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J, M., Utterback, T. R., Nagle, J. W., Fields, C. and Venter, J. C. (1992) *Nature* **355**, 623.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulous, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. and Ventor, J. C. (1991) *Science* **252**, 1651.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403.

Burglin, T. R. and Barnes, T. M. (1992) *Nature* **357**, 376.

Hoog, C. (1991) *Nucleic Acids Res.* **19**, 6123.

Huber, P. A., Redwood, C. S., Avent, N. D., Tanner, M. J. and Marston, S. B. (1993) *J. Muscle Res. Cell Motil.* **14**, 385.

Ito, K., Matsubara, K. and Okubo, K. (1994) *Gene* **140**, 295.

Klein, J. (1990) Immunology pp. 8-27 Blackwell Scientific Publication Co., Boston.

Keith, C. S., Hoang, D. O., Barret, B. M., Feigelman, B., Nelson, M. C., Thai, H. and Baysdorfer, C. (1993) *Plant Physiol.* **101**, 329.

Kim, C. W., Markiewicz, P., Lee, J. J., Schierle, C. F. and Miller, J. M. (1993) *J. Mol. Biol.* **231**, 960.

Ko, M. S. H. (1990) *Nucleic Acids Res.* **18**, 5705.

Korn, B., Sedlacek, Z., Manca, A., kioschis, P., Konecki, D., Lehrach, A. and Poustka, A. (1992) *Hum. Molecular Genetics* **1**, 235.

Koch, C. A., Anderson, D., Moran, M. F., Ellis, C. and Pawson, T. (1991) *Science* **252**, 668.

Mathews, C. K. (1990) Biochemistry (Mathews, C. K. and van Holde, K. E. eds.) pp. 779-812. The Benjamine/Cummings Publishing Co., Redwood City.

Park, Y. S., Kwark, J. M., Kwon, O. Y., Kim, Y. S., Lee, D. S., Cho, M. J., Lee, H. H. and Nam, H. G. (1993) *Plant Physiol.* **103**, 359.

Pawson, T. and Gish, G. D. (1992) *Cell* **71**, 359.

Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci.* USA **85**, 2444.

Chomczynski, P. and Sacchi, N. (1987) *Anal. Biochem.* **162**, 156.

Saito, H., Nishikawa, A., Gu, J., Ihara, Y., Soejima, H., Wada, Y., Sekiya C., Niikawa, N. and Taniguchi, N. (1994) *Biochem. Biophys. Res. Commun.* **198**(1), 318.

Sasaki, Y. F., Ayusawa, D. and Oishi, M. (1994) *Nucleic Acids Res.* **22**, 987.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular cloning-a labolatory manual 2nd ed. pp. 7.26-7.29 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Stratagene (1993) Instruction manual for ZAP-cDNA synthesis kit.