LETTER

Large-Scale Concatenation cDNA Sequencing

Wei Yu, Björn Andersson, Kim C. Worley, Donna M. Muzny, Yan Ding, Wen Liu, Jennifer Y. Ricafrente, Meredith A. Wentland, Greg Lennon,[1] and Richard A. Gibbs[2]

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030; [1]Human Genome Center, Lawrence Livermore National Laboratories, Livermore, California 94550

A total of 100 kb of DNA derived from 69 individual human brain cDNA clones of 0.7–2.0 kb were sequenced by concatenated cDNA sequencing (CCS), whereby multiple individual DNA fragments are sequenced simultaneously in a single shotgun library. The method yielded accurate sequences and a similar efficiency compared with other shotgun libraries constructed from single DNA fragments (>20 kb). Computer analyses were carried out on 65 cDNA clone sequences and their corresponding end sequences to examine both nucleic acid and amino acid sequence similarities in the databases. Thirty-seven clones revealed no DNA database matches, 12 clones generated exact matches (≥98% identity), and 16 clones generated nonexact matches (57%–97% identity) to either known human or other species genes. Of those 28 matched clones, 8 had corresponding end sequences that failed to identify similarities. In a protein similarity search, 27 clone sequences displayed significant matches, whereas only 20 of the end sequences had matches to known protein sequences. Our data indicate that full-length cDNA insert sequences provide significantly more nucleic acid and protein sequence similarity matches than expressed sequence tags (ESTs) for database searching.

[All 65 cDNA clone sequences described in this paper have been submitted to the GenBank data library under accession nos. U79240–U79304.]

The characterization of expressed sequences in the Human Genome Project has largely relied upon expressed sequence tags (ESTs) that are generated from single DNA sequence reads from the termini of cloned cDNAs. Although ESTs are powerful tools for gene discovery, characterization of gene expression, and localization of certain homologies to genomic fragments (Adams et al. 1991, 1992; Khan et al. 1992; Hillier et al. 1996), they only contribute limited information about each clone because of their short length and generally low data quality. In contrast, full-length and high-quality sequence of the cDNA clones can potentially provide more accurate database comparisons and definitive alignments for determining gene structures, and can enhance the ability of different computer programs to predict gene function. In addition, full-length cDNA insert sequences may provide more information, compared to ESTs, for annotation and interpretation of the rapidly accumulating genomic sequences in the public databases.

To facilitate the rapid, accurate, and efficient sequencing of cDNAs, we developed the concatenation cDNA sequencing (CCS) procedure, where multiple short DNA molecules (1.0–5.0 kb) are first ligated to form long DNA fragments, or concatemers, that are then randomly sheared and sequenced (Andersson et al. 1996, 1997). The boundaries between individual cDNAs are recognized at the stage of computer editing by virtue of the restriction endonuclease recognition sites that were utilized when the clones were first isolated. These sites are electronically ''cut'' prior to computer assembly, and as a result each cDNA sequence is ultimately identified as an individually assembled contiguous sequence (contig). CCS represents an efficient alternative to oligonucleotide walking and deletion library construction and has been applied previously in this laboratory for analysis of a single concatenation library from 16 cDNAs with a total length of 36 kb (Andersson et al. 1997). The upper size limits for a concatenation library that can be sequenced efficiently is not known.

In this study CCS has been applied for the generation of high-quality cDNA sequence on a large scale, using 69 different cDNAs from the Soares 1NIB infant brain library (Soares et al. 1994), totaling >100 kb in length. Computer analyses were performed on these cDNA clone sequences and their

[2]Corresponding author.
E-MAIL agibbs@bcm.tmc.edu; FAX (713) 798-5741.

corresponding ESTs to examine nucleic acid and amino acid sequence similarities in the database, and to evaluate the possible advantages of obtaining full-length cDNA insert sequences.

RESULTS

Concatenation cDNA Sequencing

The strategy of CCS has been described previously in detail (Andersson et al. 1996, 1997). The cDNA inserts studied here ranged from 0.7–2.0 kb in size, and the total length of the 69 cDNA clones was 105.4 kb. The sequencing included 1548 reads generated by forward sequencing from the M13 universal priming site, and 138 clone end sequences with an average length of 400 bp generated by forward and reverse sequencing from the original cDNA cloning vector, Lafmid BA. An additional 107 reads were also generated from the original clones, utilizing 85 specific oligonucleotide primers for gap closure and complete double-stranded coverage. The resulting sequences were edited to the established community standard of ≥99.99% accuracy.

Of a total 1793 sequences obtained for this project, 122 reads (6.8%) were not included in the assembly because of poor quality, and 106 (5.9%) were attributed to M13 vector sequences. The total number of reads required per kilobase of complete sequence was 17.0, and the number of primers used per kilobase was 0.81. The average redundancy of these cDNA sequences was ~6.7, and the distribution of reads was uniform across each sequence for most of the cDNA clones, with less than five containing overrepresentation of termini sequences (Andersson et al. 1997). The clone end sequences included in this project helped to identify each clone, as the corresponding paired 5′ and 3′ end sequences appeared in the same contig during assembly, confirming that every contig faithfully represented a single cDNA clone insert.

Computer Search for DNA Sequence Similarities in the Database

DNA sequence similarities were identified using the BLASTN via the BCM Search-Launcher (Smith et al. 1996) and were considered statistically significant when $P < 0.00001$ was identified with at least 100 bases displayed in the alignments. Identical sequences were observed between clones P3 and P5, J4 and J6, F6 and G6, and D8 and E8, respectively, leaving 65 different clones for computer analysis. Two clones, B6 and G5, lacked poly(A) tails. Four clones, O2, P4, L4, and A5, carrying *Alu* repetitive elements, were also included for the analysis, as some proteins appear to contain translated *Alu* sequences (Gish and States 1993). Of the 65 full-length clone insert sequences used for DNA similarity searches, 37 sequences were not matched to any known genes. The remaining 28 were matched to genes known previously from humans and non-mammalian organisms such as yeast and *Drosophila*. The 28 matched sequences included 12 exact matches (≥98% identity) and 16 nonexact matches (57%–97% identify) to database entries.

Although 44,608 ESTs corresponding to 25,701 1NIB cDNA library clones have been deposited in the database by Washington University EST project sponsored by Merck & Co. (Hillier et al. 1996), we elected to use our own end sequences in the assemblies. This avoided any possible discrepancy between each clone and its presumed corresponding ESTs. These problems can arise from well-to-well cross-contamination events, or from simply a lack of either 5′ or 3′ ESTs in the database, and were observed when searching for the published end sequences corresponding to clones 2F, 3M, 4K, 5E, 7P, and 8D.

The 130 corresponding end sequences derived from the 65 clones were also used to search DNA similarities (excluding dbEST), and the results were compared to those obtained from the full-length clone insert sequences (Table 1). Data from 45 clones displayed no match to nucleic acid sequences in the database, whereas 12 clones generated exact matches and 8 nonexact matches by either 5′- or 3′-end sequences. Therefore, the searching yielded identical results in the category of exact matches, for both end DNA sequences and full-length clone sequences. In the nonexact match category, however, the end sequences failed to match 8 of the 16 entries identified by full-length data.

Computer Search for Protein Sequence Similarities in the Database

Searches for protein sequence similarity were performed using the BLASTX from the BCM Search-Launcher (Smith et al. 1996) with both full-length clone insert sequences and ESTs. Protein sequence similarities were considered statistically significant with $P < 0.00001$ and with at least 10 amino acids displayed in the alignments. Of 65 clone sequences used for protein similarity searching, 38 cDNA full-length clone insert sequences were not assigned to any known protein function, whereas 27 displayed

Table 1. Comparison of Clone and End Sequences for DNA Similarity Matches

Clone[a]	Length (bp)	Match	Sequence		
			full length (acc. no.)	similarity (%)	end match
A1*	1876	serine/threonine protein kinase	X70764	54	–
B1	1410	—			–
C1	1632	—			–
D1	1554	human mRNA for L2TR-1	D38496	99	+
F1	881	—			–
G1	1601	—			–
H1*	1834	tissue-type plasminogen activator enhancer	Z48484	67	–
I1*	1680	*Drosophila* fat protein gene	M80537	57	–
L1	1615	—			–
N1	1456	—			–
P1	1606	glycerol-3-phosphate dehydrogenase	U36310	99	+
D2	1838	opioid-binding cell adhesion molecule	L34774	99	+
E2	1600	—			–
F2	1203	—			–
G2	1179	*B. malayi* 63-kD antigen	J03266	66	+
J2	1261	mouse X11 protein	L34676	91	+
K2	1218	—			–
L2	1493	—			–
M2	1488	human tumor necrosis factor receptor	M85145	76	+
N2	1706	—			–
O2	1430	human oligodendrocyte myelin glycoprotein	L05367	84	+
P2	1470	—			–
A3	1346	human deoxyhypusine synthase	L39068	100	+
B3	1584	—			–
J3	1255	human Zic protein	D76435	100	+
K3*	1768	thymidylate syntase	D00596	66	–
M3	1579	human DNA with a HBV insertion site	M15772	100	+
N3	1536	—			–
P3/P5	1414	apurinic/apyrimidinic endonuclease	M81955	99	+
B4	1296	cyclin-dependent protein kinase	M14505	100	+
I4	1275	yeast *COX11* gene	X55731	65	+
J4/J6	1198	—			–
K4	1302	human breakpoint cluster region gene	U07000	70	+
L4	1440	human eIF-4AI	D13748	99	+
M4	1506				–
N4	1060	—			–
P4	1659	human oligodendrocyte myelin glycoprotein	L05367	84	+
A5*	1574	human lysozyme	M19045	82	–
B5	1681	disulfide isomerase-related protein P5	D49489	99	+
D5	982	—			–
E5	1418	—			–

(*Continued on following page.*)

Table 1. (*Continued*)

Clone[a]	Length (bp)	Match	Sequence: full length (acc. no.)	Sequence: similarity (%)	Sequence: end match
G5	749	—			–
H5*	1719	human IgM heavy chain	X57331	73	–
I5	1480	albumin D-box binding protein	U06936	99	+
K5	1449	—			–
L5	1451	—			–
M5	1346	—			–
B6	1396	—			–
E6	1616	—			–
F6/G6*	1875	2-oxoglutarate dehydrogenase	D32056	67	–
H6	1833	—			–
K6	1362	—			–
M6	1685	female ovary mRNA for metalloproteinase	D83646	100	+
N6	1792	—			–
O6	1444	—			–
P6	1418	—			–
E7*	1949	dihydrolipoamide acetyltransferase	Y00978	58	–
G7	1637	—			–
H7	1499	—			–
I7	1529	neuronal olfactomedin-related ER localized protein	U03417	89	+
M7	1487	—			–
N7	1610	—			–
P7	1976	—			–
D8/E8	1567	—			–
H8	1738	—			–

[a](*) Computer search matched by clone but not by end sequences.

significant similarities, with 4 exact matches (≥98% identity) and 23 nonexact matches (50%–97% identity) to protein sequences defined previously. The 130 corresponding end sequences derived from the 65 clones were also examined for protein sequence identification and compared to the search results obtained from full-length clone insert sequences (Table 2). The end sequences demonstrated 45

Table 2. Computer Search Comparison of Clone Sequences and End Sequences

	DNA search		Protein search	
	full length	end sequence	full length	end sequence
No match	37	45	38	45
Exact match	12	12	4	4
Nonexact match	16	8	23	16
Total match	28	20	27	20

clones with no matches and 20 clones with significant matches to proteins defined previously, including 4 exact matches and 16 nonexact matches. Therefore, full-length clone insert sequences provided seven more protein similarity matches than their corresponding ESTs.

DISCUSSION

We have developed an efficient methodology for obtaining multiple and complete cDNA clone sequences rapidly and accurately (Andersson et al. 1996, 1997). This report describes the application of CCS for large-scale sequencing of >100 kb of DNA from 69 human brain cDNA clones. The method yielded accurate sequences with coverage in both directions and a similar efficiency compared to other shotgun libraries constructed from single inserts. The statistics from this study showed that 17.0 reads per kilobase and 0.81 primer per kilobase were required of finished sequence; this compared favorably overall to the range of 12–18 reads per kilobase and 0.5–1.2 primer per kilobase for complete sequencing of 30- to 170-kb genomic clone inserts in our laboratory using shotgun library procedures.

Computer analyses were performed on the cDNA clone sequences and end sequences to examine both the DNA and protein similarities in the public databases and to assess the possible advantage of obtaining full-length insert sequences versus ESTs. Of 65 clones used for DNA similarity searches, 28 matches were displayed with complete clone sequences and 20 with ESTs. Of 65 clones used for protein similarity searches, 27 matches were displayed with complete clone sequences and 20 with ESTs. A summary of the database search results is shown in Table 2 and demonstrates that full-length cDNA insert sequences provide significantly more information compared to ESTs, for both DNA and protein similarity search analysis.

Although consistent results were observed between DNA and protein similarity searches using clones such as A1, I1, and J2, we noted a discrepancy among other clones (B1, H1, and F2) that displayed matches from DNA similarity searching but not from protein similarity searching, and vice versa. The protein similarity search can be complicated by the variation in the size of the expressed 3′ untranslated region (3′ UTR) from different genes and even from the same gene undergoing alternative splicing, as human 3′ UTRs up to 3.5 kb in length have been reported, as well as mRNA species with 3′ UTRs differing by 2 kb in length from the same gene (Qian et al. 1993; Tam et al. 1994; Boyd et al. 1995). For mRNAs with a 3′ UTR extended over 2 kb, neither full-length cDNA insert sequence nor 3′ ESTs can provide a correct match for protein similarity searches.

Other factors that can each contribute to a discrepancy between DNA and protein similarity searches may include clone chimerism, a lack of corresponding amino acid sequences in the database, or a deletion or insertion of a single nucleotide during sequencing and assembly.

The development and use of high-quality cDNA libraries is critical to the success of cDNA sequencing efforts (Sikela and Auffray 1993). The availability of cDNA libraries representing a high percentage of complete mRNA molecules will certainly facilitate the generation of a full map of human expressed sequences, whereas the establishment of efficient and large-scale approaches for full-length sequencing of cDNA clones is an important step toward that goal.

METHODS

cDNA Insert Purification and Concatenation

Complementary DNA clones (plate 074) from the Soares human infant brain 1NIB cDNA libraries (Soares et al. 1994) were used in this study. Each clone was grown by standard microbiological methods, and phagemids were purified with the Automated Nucleic Acid Isolation System (AutoGen 740, Integrated Separation System).

Three micrograms of each clone were digested with 30 units of *Hin*dIII (Pharmacia) and 30 units of *Not*I (New England Biolabs, NEB) in 50 µl of 1× NE buffer 3 (NEB) in the presence of 1 mg/ml of bovine serum albumin (NEB) for 2 hr at 37°C. Following digestion, the inserts were electrophoresed and excised from a low melting agarose gel and purified using the freeze-thaw method (Andersson et al. 1996). Sixty-nine cDNA inserts (0.3–0.4 µg each) were pooled and concatenated in a reaction containing 1× One-Phor-All buffer (Pharmacia) and 2 mM ATP in the presence of 20 units of T4 DNA ligase (Pharmacia) at room temperature for 6 hr. The size of the concatenated DNA was verified by agarose gel electrophoresis.

Construction of the M13 Shotgun Library

The M13 shotgun library from concatenated and sheared DNA fragments was constructed using the double adaptor method as described previously (Andersson et al. 1996).

M13 Template Preparation

The single-stranded M13 templates were prepared using a protocol for glass fiber filter purification of M13 DNA in a 96-well format as described previously (Andersson et al. 1997).

DNA Sequencing and Sequence Assembly

DNA sequencing was performed using the fluorescent auto-

mated approach on Applied Biosystems DNA 373 sequencers. Random sequences were genereated by forward sequencing from the M13 universal primer site using dye-labeled primers. The ABI Prism dye terminator cycle sequencing ready reaction kit (Perkin Elmer) was used for both custom primer-directed sequences and clone end sequence determination by forward and reverse sequencing from the original cDNA cloning vector Lafmid BA M13 universal and reverse priming sites. The sequence reads were edited and assembled into contigs using the computer program SEQUENCHER for Macintosh, version 3.0 (Gene Codes Corp.). Gap closure and full coverage of each clone in both directions were completed by using primer-directed sequencing.

Computer Analysis

Computer analyses were performed on both cDNA clone sequences and corresponding ESTs via the BLASTN program (Altschul et al. 1990) for DNA similarity searching and the BLASTX program (Gish and States 1993) for protein similarity searching. The parameters used in the searches included an expect value of 10 (with a cutoff calculated using the expect values), and a word length of 12 for BLASTN and 3 for BLASTX, with the scores calculated using the BLOSUM62 matrix. These programs were accessed from the World Wide Web page (http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html) and via the BCM Search Launcher Batch Client (Smith et al. 1996).

ACKNOWLEDGMENTS

We thank M. Ali Ansari-Lari for helpful discussions. We also thank Harley Gorrell for computer assistance and Kecia Rowland for sequencing assistance. This study was partially supported by National Institutes of Health (NIH) grant RO1 HG00823, Genome Program Center grant P30 HG00210-05, and the Department of Energy. W.Y. is a recipient of a National Research Service Award postdoctoral fellowship (1 F32 HG00169-01) from NIH.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J.C. Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651–1656.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* 355: 632–634.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Andersson, B., M.A. Wentland, J.Y. Ricafrente, W. Liu, and R.A. Gibbs. 1996. A "double adaptor" method for improved shotgun library construction. *Anal. Biochem.* 236: 107–113.

Andersson, B., J. Lu, Y. Shen, M.A. Wentland, and R.A. Gibbs. 1997. Simultaneous shotgun sequencing of multiple cDNA clones. *DNA Sequence* (in press).

Boyd, C.D., T.J. Mariani, Y.H. Kim, and K. Csiszar. 1995. The size heterogeneity of human lysyl oxidase mRNA is due to alternate polyadenylation site and not alternate exon usage. *Mol. Biol. Rep.* 21: 95–103.

Gish, W. and D.J. States. 1993. Identification of protein coding regions by database similarity search. *Nature Genet.* 3: 266–272.

Hillier, L., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M.B. Soares, F. Tan, J. Thierry-Meg, E. Trevaskis, K. Underwood, P. Wohldman, R. Waterston, R. Wilson, and M. Marra. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6: 807–828.

Khan, A.S., A.S. Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* 2: 180–185.

Qian, J.F., E. Lazar-Wesley, C. Breugnot, and E. May. 1993. Human transforming growth factor alpha: Sequence analysis of the 4.5-kb and 1.6-kb mRNA species. *Gene* 132: 291–296.

Sikela, J.M. and C. Auffray. 1993. Finding new genes faster than ever. *Nature Genet.* 3: 189–190.

Smith, R.F., B.A. Wiese, M.K. Wojzynski, D.B. Davison, and K.C. Worley. 1996. BCM Search Launcher—An integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res.* 6: 454–462.

Soares, M.B., M.F. Bonaldo, P. Jelene, L. Su, L. Lawton, and A. Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* 91: 9228–9232.

Tam, S.W., L.R. Cote-Paulino, D.A. Peak, K. Sheahan, and M.J. Murnane. 1994. Human cathepsin B-encoding cDNAs: sequence variations in the 3′-untranslated region. *Gene* 139: 171–176.

Received November 15, 1996; accepted in revised form February 4, 1997.