Analysis of 5′-End Sequences of Chimpanzee cDNAs

Ryuichi Sakate,[1,5] Naoki Osada,[2] Munetomo Hida,[3] Sumio Sugano,[3] Ikuo Hayasaka,[4] Naoko Shimohira,[1] Shinsuke Yanagi,[1] Yumiko Suto,[1] Katsuyuki Hashimoto,[2] and Momoki Hirai[1]

[1]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8562, Japan; [2]Division of Genetic Resources, National Institute of Infectious Diseases, Tokyo 162-8640, Japan; [3]Department of Genome Structure Analysis, Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan; [4]Kumamoto Primate Research Park, Sanwa Kagaku Kenkyusho Co., Ltd., Kumamoto 869-3201, Japan

We constructed full-length enriched cDNA libraries from chimpanzee brain, skin, and liver tissues by the oligo-capping method to establish a database of sequences of chimpanzee genes. Randomly selected clones from the libraries were subjected to one-pass sequencing from their 5′-ends. As a result, we collected 6813 chimpanzee cDNA sequences longer than 400 bp. Homology search against human mRNA sequences (RefSeq mRNAs) revealed that our collection included sequences of 1652 putative chimpanzee genes. In order to calculate the sequence identity between human and chimpanzee homologs, we constructed 5′-end consensus sequences of 226 chimpanzee genes by aligning at least three sequences for individual genes. Sequence identity was estimated by comparing these consensus sequences and the corresponding sequences of their human homologs. The average sequence identity of the 5′-end cDNAs was 99.30%. Those of the 5′-UTRs and CDSs were 98.79% and 99.42%, respectively. The results confirmed that human and chimpanzee genes are highly conserved at the nucleotide level. As for amino acids, the average sequence identity was 99.44%. The average synonymous (K_S) and nonsynonymous (K_A) divergences were estimated to be 1.33% and 0.28%, respectively.

[Supplemental material is available online at www.genome.org. All of the 1947 sequences used for constructing the consensus sequences of 226 chimpanzee genes have been submitted to DDBJ under accession nos. AU296732–AU298678. Two hundred twenty-six consensus sequences and their detailed annotation descriptions are available at our Web site http://www.prigen.org/.]

Since the draft sequence of the human genome was determined (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), efforts have been under way to construct more comprehensive databases of human genes and their expression patterns. For better understanding of the biological characteristics of humans, a comparative analysis of chimpanzee genes with human genes will yield valuable information. The identity of genomic sequences between humans and chimpanzees was first estimated to be about 98.5% by the DNA–DNA hybridization method (Sibley and Ahlquist 1984, 1987). Recently, a comparative map has been constructed through paired alignment of genome-wide chimpanzee bacterial artificial chromosome end sequences with publicly available human genome sequences (Fujiyama et al. 2002). It revealed that the genomic sequence identity between humans and chimpanzees was 98.77%. Although human and chimpanzee genomes are highly identical at the nucleotide level, morphological traits and cognitive abilities are distinct between these two species. A comparative analysis of mRNA sequences may provide clues to the genetic information that affects the differing phenotypes. In contrast to the great number (about 16,000) of human mRNA sequence entries in public databases such as the RefSeq mRNAs of the National Center for Biotechnology Information (NCBI), only a small number of chimpanzee mRNA sequences and expressed sequence tags (ESTs) have been deposited in public databases. Moreover, the variety of chimpanzee genes in the databases is biased; they contain sequences derived mainly from mtDNAs and genes related to the major histocompatibility complex (MHC), which are known to evolve rapidly and are suitable for the analysis of phylogenetic relationships among closely related species. In this study, we attempted to analyze chimpanzee (*Pan troglodytes verus*) mRNA sequences using a substantial number of 5′-end enriched cDNA clones in order to establish a standard reference between the two species. We constructed cDNA libraries from the brain, skin, and liver tissues of two chimpanzees and sequenced the 5′-ends of 6813 clones. As a result, we were able to annotate the consensus sequences of 226 putative chimpanzee genes by comparing with the sequences of human homologs in public databases.

RESULTS

We collected 7064 5′-end sequences of chimpanzee cDNAs. These 5′-expressed sequence tags (5′-ESTs) were annotated by the BLAST program (Altschul et al. 1990). Consequently, 163 mitochondrial, 71 repetitive, and 17 vector sequences that were included in our raw sequence data were eliminated. The

[5]Corresponding author.
EMAIL sakate@prigen.org; FAX 81-4-7136-3687.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.783103.

13:1022–1026 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

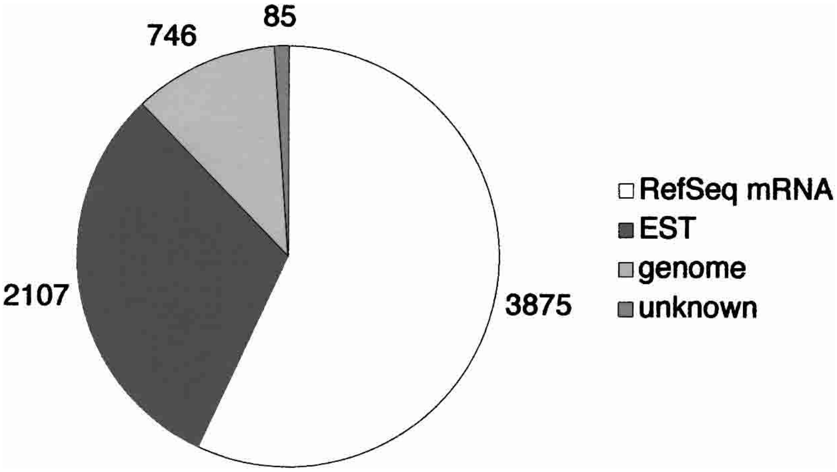

Figure 1 Constitution of 6813 chimpanzee cDNA sequences. Numbers of sequences that matched those of human RefSeq mRNA, EST, and genome sequences are indicated in this figure. "Unknown sequences" denote the sequences that did not match any sequences in public databases.

remaining 6813 ESTs consisted of 3875 sequences that matched human RefSeq mRNAs, 2107 that matched human ESTs, and 746 that matched human genome sequences (Fig. 1). The rest (85 sequences) did not match any sequences in the public databases. Of those that matched human RefSeq mRNAs, 2835 (73.2%) contained the translation start site. These 3875 sequences were clustered into 1652 nonredundant chimpanzee genes. It is worth noting that each of the 1537 sequences (93.0%) were found to correspond to only one human RefSeq mRNA by the BLAST search at a threshold value of $E = 1e^{-120}$, suggesting that these genes are orthologous to the corresponding human genes.

From these 1652 sequences, we constructed a total of 226 consensus sequences of 5′-end cDNAs. As described in the methods, each consensus sequence was established by aligning at least three sequences. The average length of the consensus sequences was 399.9 bp. The molecular functions of the 226 genes were annotated according to the classification system by the Molecular Function of GO categories (http://www.geneontology.org/). The distribution of the functions is shown in the Supplementary Figure 1. These 226 consensus sequences were aligned with 5′-untranslated regions (5′-UTRs) and/or coding sequences (CDSs) of human RefSeq mRNAs. We calculated sequence identities between 226 chimpanzee consensus sequences and the corresponding human RefSeq mRNA sequences (Suppl. Table 1). Among these 226 5′-end cDNA sequences, three consisted solely of 5′-UTR and 29 consisted solely of CDS. One hundred ninety-four sequences contained both 5′-UTRs and CDSs. When we calculated the average sequence identity, we excluded 28 5′-UTR regions and one CDS region of the consensus sequences because they were not sufficiently long enough to calculate reliable values. The distribution of the sequence identities (%) of the 5′-end consensus sequences of cDNAs and their CDS regions are shown separately in Figure 2. Two sequences with an exceptionally low identity were those of the polymorphic MHC-related genes. The remaining genes showed sequence identities greater than 97.0%.

The average sequence identity (%) with standard deviation of the 5′-UTR in 169 genes was 98.79% ± 1.71%. That of the CDS in 222 genes was 99.42% ± 0.62%, and that of the amino acids was 99.44% ± 1.20%. As for all the 5′-end regions of 226 genes, the average sequence identity was 99.30% ± 0.62%. The average K_S and K_A values (%) based on the data of 222 genes were determined to be 1.62% ± 1.75% and 0.26% ± 0.55%, respectively, by the method of Miyata-Yasunaga (1980), and 1.33% ± 1.54% and 0.28% ± 0.59%, respectively, by the method of Li (1993).

When all the bases for 5′-UTRs (197 genes, 15,665 bp) and CDSs (223 genes, 72,708 bp) were combined, their sequence identities were 98.79% and 99.42%, respectively. As for amino acids (24,011 a. a.), the average identity was 99.44%. As for all the 5′-end regions of 226 genes (90,381 bp), the sequence identity was 99.31%. When these combined data were used, the K_S and K_A values were respectively found to be 1.65% and 0.26%, by the method of Miyata and Yasunaga (1980), and 1.35% and 0.27%, by the method of Li (1993). The K_S values obtained by the two methods were different, while the K_A values were the same irrespective of the method used.

DISCUSSION

When analyzing sequence data of cDNAs, limitations in accuracy must be considered. mRNAs are fragile by nature and the final sequencer outputs are apt to include sequence errors generated during the cloning and sequencing processes. Therefore, sequence data derived from only one cDNA clone is not sufficient to obtain reliable information on genes. In this study, we aimed to precisely calculate base-by-base sequence identities. From our collection of 6813 cDNA 5′-end sequences, we were able to construct the consensus sequences of 226 chimpanzee genes based on at least three sequences for each gene. This is probably the first report concerning the comparative sequence analysis between humans and chimpanzees using such a substantial number of genes. In previous studies using 20 GenBank cDNAs (Varki 2000), the sequence identity (%) between human and chimpanzee cDNAs was 99.31% ± 0.38% (mean ± S.D.), and that of amino acid was 99.36% ± 0.66%. These values are not different from those obtained in our study (99.30% ± 0.62% and 99.44% ± 1.20%, respectively). Therefore, the sequence identity in the coding regions between humans and chimpanzees is higher, as expected, than that of the genome sequences (98.77%) reported by Fujiyama et al. (2002). It should be emphasized that we collected the 5′-UTR sequences of mRNAs. Thus far, there is no sufficient information on the 5′-UTR region based on which the substitution rate can be calculated. The average sequence divergence of 5′-UTRs between humans and chimpanzees was found to be 1.21%. This value is the same as those of the genomic sequence differences reported previously (Fujiyama et al. 2002, 1.23%; Chen and Li 2001, 1.22%).

As seen in Figure 2, two genes showed a low sequence identity (96.4%). These were MHC-related genes (PC_061

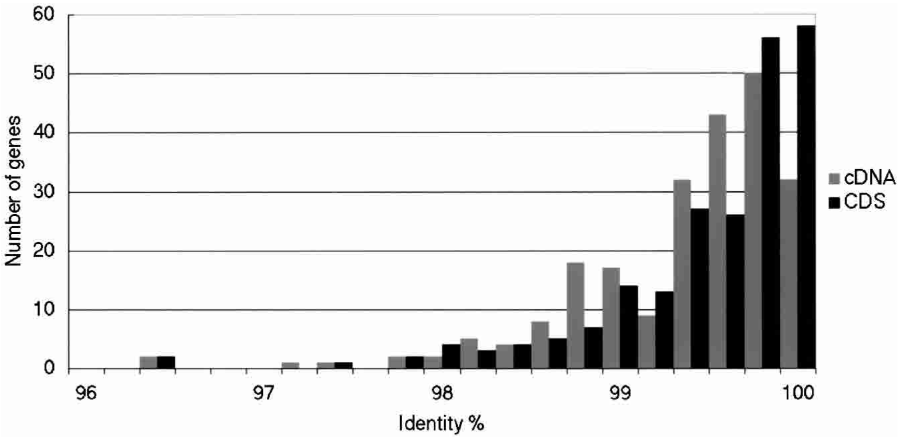

Figure 2 Distribution of sequence identities of 5′-end consensus sequences of chimpanzee cDNAs (226 cDNAs and 222 CDS regions).

[HLA-A homolog] and PC_133 [HLA-B homolog]) as listed in the Supplementary Table 1. The results suggest that these MHC-related genes evolved rapidly. Another MHC-related gene in our collection (PC_134, HLA-E homolog) showed a higher identity (98.2%) with the human homolog than that of the HLA-A and -B genes, and seemed to be relatively conserved. This is consistent with a previous study (Adams and Parham 2001), in which African apes were shown to have orthologs of all human class I MHC-related genes, and HLA-A, B, and C genes were suggested to be highly polymorphic while others (HLA-E, F, and G) conserved.

The synonymous (K_S) and nonsynonymous (K_A) divergences obtained in this study by the method of Miyata and Yasunaga (1980) were 1.62% and 0.26%, respectively. When using the method of Li (1993), the values were 1.33% and 0.28%, respectively. The K_S values calculated by the two methods are different. Li (1993) suggested that the method of Miyata and Yasunaga tends to overestimate the K_S value. Therefore, we considered the values calculated by the method of Li (1993) in this study. In a previous study (Chen et al. 2001), which analyzed 88 GenBank chimpanzee cDNAs, the K_S and K_A values were calculated to be 1.48% and 0.55%, respectively, by the method of Li (1993). These values are larger than our K_S and K_A values obtained by the same method. Distribution of the K_S and K_A values of 88 gene set (Chen et al. 2001) and that of our 226 gene set are shown in the Supplementary Figure 2. In these two gene sets, 15 out of 88 genes (15.9%) and 24 out of 222 genes (10.8%) had a value of $K_A / K_S \geq 1$ (Suppl. Table 1). This implies that the sequence data of 88 GenBank cDNAs include rapidly evolving genes such as immune-related genes and duplicated gene. Our calculation is based on 226 randomly selected genes and may represent a genome-wide average substitution rate of expressed sequences (functional distribution of 226 genes is shown in the Suppl. Fig. 1).

The objective of this study was to precisely calculate sequence diversity between human and chimpanzee homologs. While aligning chimpanzee sequences with those of human RefSeq mRNAs, we noted some problems that need to be clarified. Since studies that address these problems are currently underway, we just briefly report preliminary results.

1. As for the 85 sequences that did not match any human sequences in public databases, they may include those that may match ever-increasing genome and EST sequence databases, those that may be putative chimpanzee-specific transcripts, and those of artifacts. Thus far, we selected ten sequences out of the 85 sequences to confirm the presence of transcripts corresponding to the sequences in human and chimpanzee lymphoblastoid cells by RT-PCR using two sets of primers for each sequence. Primers were designed on the basis of chimpanzee sequences. As a result, three sequences (PorA1155, PstA6283, and PstA7892) were transcribed in both humans and chimpanzees, and one sequence (PccB3689) transcribed only in chimpanzees. The former could be unknown genes and the latter could be a chimpanzee specific transcript, though we did not find any protein domain in the four sequences by using NCBI CD-Search (http://www.ncbi.nlm.nih.gov:80/Structure/cdd/cdd.shtml).
2. We found chimpanzee sequences with ten or more nucleotides at the 5′- or the 3′-end that are completely different from corresponding sequences in human RefSeq. We excluded these inconsistent sequences by computational process and visual inspection because these could affect the results of identity calculation. Thus far, we compared these sequences with the human genome sequences, and found that several sequence inconsistencies could be the product of alternative splicing (Suppl. Table 2). Since a high proportion of human genes have been suggested to undergo alternative splicing (Brett et al. 2002), it is interesting to analyze the possible interspecific alternative splicing affecting gene functions. Recently, Britten has claimed that sequence diversity between human and chimpanzee genomes is as high as 5% when insertions/deletions (indels) are taken into account (Britten 2002). We detected indels in both the 5′-UTR and CDS regions of 226 consensus sequences comparing with human RefSeq mRNAs. In Supplementary Table 3, we listed indels which were confirmed by comparing with corresponding human genome sequences.
3. Recent studies have shown a high frequency of single-nucleotide polymorphisms (SNPs) in the human genome (one SNP per 1.08 kb; The International SNP Map Working Group 2001). Considering threefold to fourfold sequence diversity in chimpanzees compared with that in humans (Kaessmann et al. 1999, 2001), SNPs in chimpanzees are expected to occur frequently. Actually, we found that 16 consensus sequences contained putative SNPs (e.g., PC_121: 210th nucleotide is three As vs. three Gs). Each allele was determined when at least two clones contained the same nucleotide at a polymorphic base position. In total, three SNPs were found in the 5′-UTRs and 14 SNPs in the CDSs (PC_061, HLA-A homolog, containing three SNPs in the CDS region and PC_133, HLA-B homolog, containing 23 SNPs in the CDS region were excluded). The 17 SNPs in the chimpanzee sequences were not found in the human database (dbSNP). Among the 14 SNPs in the CDSs, seven (seven at the third codon) were synonymous and

seven (three at the second codon and four at the first codon) were nonsynonymous (Suppl. Table 4).

For the structural and functional analysis of genes, collection of data on alternative splicing, indels, and SNPs is important. In addition, a comparative analysis of tissue-specific expressions of genes between humans and chimpanzees is expected to shed light on species specificity and evolution of humans. Our collection of full-length cDNA clones and sequence data could be a valuable resource for postgenomic research.

METHODS

Construction of cDNA Libraries and Annotation of cDNA Sequences

Tissue specimens were collected from adult chimpanzees (*Pan troglodytes verus*) kept in the Primate Research Park, Kumamoto, Japan. All the procedures of tissue collection were approved by an institutional board. About 10 g of skin tissues was collected by biopsy from a male chimpanzee. Liver and brain tissues were obtained at autopsy from a female chimpanzee that died of septicemia. The full-length enriched cDNA libraries were constructed by the oligo-capping method (Maruyama and Sugano 1994; Hida et al. 2000). From these libraries, clones were randomly selected and their sequences were determined from the 5′-end by one-pass sequencing using ABI-377 and ABI-3100 sequencers. After eliminating the 5′-end vector sequences and undecided 3′-end sequences using our in-house program, only sequences longer than 400 bp were used for further analysis. Sequence base-calling was performed by the basecaller program attached to the ABI sequencers. The sequence data were subjected to the computer-based homology search against those of vector pME18S-FL3 and those in public databases (human RefSeq mRNAs [including mitochondria] and human ESTs of NCBI [http://www.ncbi.nlm.nih.gov/]; human repetitive sequences of REPBASE [http://www.girinst.org/Repbase_Update.html; Jurka 2000]; and human genome of the University of California Santa Cruz [UCSC, http://genome.ucsc.edu/, April 2001 freeze]) using the BLAST program (Altschul et al. 1990). The threshold values (BLAST expectation values) used to execute the BLAST program for the vector sequences, human RefSeq mRNAs, human mitochondrial sequences, human repetitive sequences, human ESTs, and human genome sequences were 1e,$^{-10}$ 1e,$^{-120}$ 1e,$^{-60}$ 1e,$^{-60}$ 1e,$^{-60}$ and 1e,$^{-30}$ respectively.

Construction of Consensus Sequence and Calculation of Sequence Identity

The chimpanzee sequences that matched the sequences of human RefSeq mRNAs were defined as the sequences of putative chimpanzee genes. Consensus sequences were established from at least three sequences of individual genes. Multiple sequences clustered to each gene were aligned together using the CLUSTAL W program (Thompson et al. 1994) and the output was computationally and visually inspected to remove alignment errors. As a result, we collected 226 consensus sequences from 1947 5′-end cDNA sequences. The average length of the 1947 sequences was 519.6 bp, and the average percentage of undecided bases, denoted as N, was 0.74%. The refined alignment data were processed using our original in-house program that determines individual bases by majority at each nucleotide site. The resulting consensus sequences were aligned with the sequences of the corresponding human RefSeq mRNAs. Then, a base-by-base comparison was conducted to calculate sequence identity of the 5′-end cDNA sequences using our original in-house program. 5′-UTR and CDS included in the 5′-end cDNA sequences were identified and analyzed. 3′-UTRs were omitted from our analysis because of insufficient data set. When the program was executed, gaps and Ns were excluded from the calculation, and alignment mismatches were eliminated by visual inspection. The identity of each amino acid was calculated for sequences spanning from the 5′-end to the 3′-end until an erroneous (not in-frame) gap appeared. The rates of synonymous (K_S) and nonsynonymous (K_A) substitutions were calculated using two previously reported methods (Miyata and Yasunaga 1980; Li 1993).

ACKNOWLEDGMENTS

This study was supported by the health science research grant for the human genome program from the Ministry of Health and Welfare of Japan, and by grant Nos. 13202015 and 13554035 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. We thank Eiichi Sato and Miho Aruga for their assistance.

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, E.J. and Parham, P. 2001. Species-specific evolution of MHC class I genes in the higher primates. *Immunol. Rev.* **20:** 41–64.

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

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. 2002. Alternative splicing and genome complexity. *Nat. Genet.* **30:** 29–30.

Britten, R.J. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl. Acad. Sci.* **99:** 13633–13635.

Chen, F.-C. and Li, W.-H. 2001. Genomic divergence between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68:** 444–456.

Chen, F.-C., Vallender, E.J., Wang, H., Tzeng, C.S., and Li, W.-H. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92:** 481–489.

Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T.D., Itoh, T., Tsai, S.-F., Park, H.-S., Yaspo, M.-L., Lehrach, H., Chen, Z,. et al. 2002. Construction and analysis of a human–chimpanzee comparative clone map. *Science* **295:** 131–134.

Hida, M., Suzuki, Y., Sugano, S., Hashimoto, K., Terao, K., Hayasaka, I., and Hirai, M. 2000. Construction and preliminary characterization of full-length enriched cDNA libraries for nonhuman primates. *Primate Res.* **16:** 95–110.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

The International SNP Map Working Group 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Jurka, J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16:** 418–420.

Kaessmann, H., Wiebe, V., and Paabo, S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286:** 1159–1162.

Kaessmann, H., Wiebe, V., Weiss, G., and Paabo, S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* **27:** 155–156.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.

Maruyama, K. and Sugano, S. 1994. Oligo-capping: A simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138:** 171–174.

Miyata, T. and Yasunaga, T. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16:** 23–36.

Sibley, C.G. and Ahlquist, J.E. 1984. The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. *J. Mol. Evol.* **20:** 2–15.

Sibley, C.G. and Ahlquist, J.E. 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *J. Mol. Evol.* **26:** 99–121.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Varki, A. 2000. A chimpanzee genome project is a biomedical imperative. *Genome Res.* **10:** 1065–1070.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

WEB SITE REFERENCES

http://www.ncbi.nlm.nih.gov/; NCBI.

http://www.girinst.org/Repbase_Update.html; REPBASE UPDATE.

http://genome.ucsc.edu/; UCSC Genome Bioinformatics.

http://www.geneontology.org/; Gene Ontology (GO) Consortium.

http://www.ncbi.nlm.nih.gov:80/Structure/cdd/cdd.shtml; A conserved Domain Database and Search Service, v1.60.

Received September 10, 2002; accepted in revised form March 5, 2003.